

# Unsupervised Quantification of Entity Consistency between Photos and Text in Real-World News

Von der Fakultät für Elektrotechnik und Informatik  
der Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des akademischen Grades  
**Doktor der Ingenieurwissenschaften**

Dr.-Ing.

genehmigte Dissertation von

**M. Eng. Eric Müller-Budack**

2022

Referent: Prof. Dr. Ralph Ewerth  
Korreferent: Prof. Dr.-Ing. Bodo Rosenhahn  
Tag der Promotion: 10.12.2021

# Abstract

In today’s information age, the World Wide Web and social media are important sources for news and information. Different modalities (in the sense of information encoding) such as photos and text are typically used to communicate news more effectively or to attract attention. Communication scientists, linguists, and semioticians have studied the complex interplay between modalities for decades and investigated, e.g., how their combination can carry additional information or add a new level of meaning. The number of shared concepts or entities (e.g., persons, locations, and events) between photos and text is an important aspect to evaluate the overall message and meaning of an article. Computational models for the quantification of image-text relations can enable many applications. For example, they allow for more efficient exploration of news, facilitate semantic search and multimedia retrieval in large (web) archives, or assist human assessors in evaluating news for credibility. To date, only a few approaches have been suggested that quantify relations between photos and text. However, they either do not explicitly consider the cross-modal relations of entities – which are important in the news – or rely on supervised deep learning approaches that can only detect the cross-modal presence of entities covered in the labeled training data. To address this research gap, this thesis proposes an unsupervised approach that can quantify entity consistency between photos and text in multimodal real-world news articles.

The first part of this thesis presents novel approaches based on deep learning for information extraction from photos to recognize events, locations, dates, and persons. These approaches are an important prerequisite to measure the cross-modal presence of entities in text and photos. First, an ontology-driven event classification approach that leverages new loss functions and weighting schemes is presented. It is trained on a novel dataset of 570,540 photos and an ontology with 148 event types. The proposed system outperforms approaches that do not use structured ontology information. Second, a novel deep learning approach for geolocation estimation is proposed that uses additional contextual information on the environmental setting (indoor, urban, natural) and from earth partitions of different granularity. The proposed solution outperforms state-of-the-art approaches, which are trained with significantly more photos. Third, we introduce the first large-scale dataset for date estimation with more than one million photos taken between 1930 and 1999, along with two deep learning approaches that treat date estimation as a classification and regression problem. Both approaches achieve very good results that are superior to human annotations. Finally, a novel approach is presented that identifies public persons and their co-occurrences in news

photos extracted from the *Internet Archive*, which collects time-versioned snapshots of web pages that are rarely enriched with metadata relevant to multimedia retrieval. Experimental results confirm the effectiveness of the deep learning approach for person identification.

The second part of this thesis introduces an unsupervised approach capable of quantifying image-text relations in real-world news. Unlike related work, the proposed solution automatically provides novel measures of cross-modal consistency for different entity types (persons, locations, and events) as well as the overall context. The approach does not rely on any pre-defined datasets to cope with the large amount and diversity of entities and topics covered in the news. State-of-the-art tools for natural language processing are applied to extract named entities from the text. Example photos for these entities are automatically crawled from the Web. The proposed methods for information extraction from photos are applied to both news images and example photos to quantify the cross-modal consistency of entities. Two tasks are introduced to assess the quality of the proposed approach in real-world applications. Experimental results for document verification and retrieval of news with either low (potential misinformation) or high cross-modal similarities demonstrate the feasibility of the approach and its potential to support human assessors to study news.

**Keywords:** Image-text relations, news analytics, multimedia retrieval, image indexing, face recognition, date estimation, geolocation estimation, event classification, deep learning, computer vision, natural language processing

# Zusammenfassung

Das World Wide Web und die sozialen Medien übernehmen im heutigen Informationszeitalter eine wichtige Rolle für die Vermittlung von Nachrichten und Informationen. In der Regel werden verschiedene Modalitäten im Sinne der Informationskodierung wie beispielsweise Fotos und Text verwendet, um Nachrichten effektiver zu vermitteln oder Aufmerksamkeit zu erregen. Kommunikations- und Sprachwissenschaftler erforschen das komplexe Zusammenspiel zwischen Modalitäten seit Jahrzehnten und haben unter anderem untersucht, wie durch die Kombination der Modalitäten zusätzliche Informationen oder eine neue Bedeutungsebene entstehen können. Die Anzahl gemeinsamer Konzepte oder Entitäten (beispielsweise Personen, Orte und Ereignisse) zwischen Fotos und Text stellen einen wichtigen Aspekt für die Bewertung der Gesamtaussage und Bedeutung eines multimodalen Artikels dar. Automatisierte Ansätze zur Quantifizierung von Bild-Text-Beziehungen können für zahlreiche Anwendungen eingesetzt werden. Sie ermöglichen beispielsweise eine effiziente Exploration von Nachrichten, erleichtern die semantische Suche von Multimedia-Inhalten in (Web)-Archiven oder unterstützen menschliche Analysten bei der Evaluierung der Glaubwürdigkeit von Nachrichten. Allerdings gibt es bislang nur wenige Ansätze, die sich mit der Quantifizierung von Beziehungen zwischen Fotos und Text beschäftigen. Diese Ansätze berücksichtigen jedoch nicht explizit die intermodalen Beziehungen von Entitäten, welche eine wichtige Rolle in Nachrichten darstellen, oder basieren auf überwachten multimodalen Deep-Learning-Techniken. Diese überwachten Lernverfahren können ausschließlich die intermodalen Beziehungen von Entitäten detektieren, die in annotierten Trainingsdaten enthalten sind. Um diese Forschungslücke zu schließen, wird in dieser Arbeit ein unüberwachter Ansatz zur Quantifizierung der intermodalen Konsistenz von Entitäten zwischen Fotos und Text in realen multimodalen Nachrichtenartikeln vorgestellt.

Im ersten Teil dieser Arbeit werden neuartige Verfahren auf Basis von Deep Learning zur Extrahierung von Informationen aus Fotos vorgestellt, um Ereignisse (Events), Orte, Zeitangaben und Personen automatisch zu erkennen. Diese Verfahren bilden eine wichtige Voraussetzung, um die Beziehungen von Entitäten zwischen Bild und Text zu bewerten. Zunächst wird ein Ansatz zur Ereignisklassifizierung präsentiert, der neuartige Optimierungsfunktionen und Gewichtungsschemata nutzt um Ontologie-Informationen aus einer Wissensdatenbank in ein Deep-Learning-Verfahren zu integrieren. Das Training erfolgt anhand eines neu vorgestellten Datensatzes, der 570.540 Fotos und eine Ontologie mit 148 Ereignistypen enthält. Der Ansatz übertrifft die Ergebnisse von Referenzsystemen

die keine strukturierten Ontologie-Informationen verwenden. Weiterhin wird ein Deep-Learning-Ansatz zur Schätzung des Aufnahmeortes von Fotos vorgeschlagen, der Kontextinformationen über die Umgebung (Innen-, Stadt-, oder Naturaufnahme) und von Erdpartitionen unterschiedlicher Granularität verwendet. Die vorgeschlagene Lösung übertrifft die bisher besten Ergebnisse von aktuellen Forschungsarbeiten, obwohl diese deutlich mehr Fotos zum Training verwenden. Darüber hinaus stellen wir den ersten Datensatz zur Schätzung des Aufnahmejahres von Fotos vor, der mehr als eine Million Bilder aus den Jahren 1930 bis 1999 umfasst. Dieser Datensatz wird für das Training von zwei Deep-Learning-Ansätzen zur Schätzung des Aufnahmejahres verwendet, welche die Aufgabe als Klassifizierungs- und Regressionsproblem behandeln. Beide Ansätze erzielen sehr gute Ergebnisse und übertreffen Annotationen von menschlichen Probanden. Schließlich wird ein neuartiger Ansatz zur Identifizierung von Personen des öffentlichen Lebens und ihres gemeinsamen Auftretens in Nachrichtenfotos aus der digitalen Bibliothek *Internet Archiv* präsentiert. Der Ansatz ermöglicht es unstrukturierte Webdaten aus dem *Internet Archiv* mit Metadaten, beispielsweise zur semantischen Suche, zu erweitern. Experimentelle Ergebnisse haben die Effektivität des zugrundeliegenden Deep-Learning-Ansatzes zur Personenerkennung bestätigt.

Im zweiten Teil dieser Arbeit wird ein unüberwachtes System zur Quantifizierung von Bild-Text-Beziehungen in realen Nachrichten vorgestellt. Im Gegensatz zu bisherigen Verfahren liefert es automatisch neuartige Maße der intermodalen Konsistenz für verschiedene Entitätstypen (Personen, Orte und Ereignisse) sowie den Gesamtkontext. Das System ist nicht auf vordefinierte Datensätze angewiesen, und kann daher mit der Vielzahl und Diversität von Entitäten und Themen in Nachrichten umgehen. Zur Extrahierung von Entitäten aus dem Text werden geeignete Methoden der natürlichen Sprachverarbeitung eingesetzt. Exemplarbilder für diese Entitäten werden automatisch aus dem Internet beschafft. Die vorgeschlagenen Methoden zur Informationsextraktion aus Fotos werden auf die Nachrichten- und heruntergeladenen Exemplarbilder angewendet, um die intermodale Konsistenz von Entitäten zu quantifizieren. Es werden zwei Aufgaben untersucht um die Qualität des vorgeschlagenen Ansatzes in realen Anwendungen zu bewerten. Experimentelle Ergebnisse für die Dokumentverifikation und die Beschaffung von Nachrichten mit geringer (potenzielle Fehlinformation) oder hoher multimodalen Konsistenz zeigen den Nutzen und das Potenzial des Ansatzes zur Unterstützung menschlicher Analysten bei der Untersuchung von Nachrichten.

**Stichworte:** Bild-Text-Beziehungen, Nachrichtenanalyse, Multimedia Retrieval, Bildindexierung, Personenerkennung, Schätzung des Aufnahmejahres, Schätzung des Aufnahmeortes, Eventklassifikation, Deep Learning, Maschinelles Sehen, Natürliche Sprachverarbeitung

# Acknowledgement

I would like to take this opportunity to thank all the people who have supported me during my time as a PhD student. This thesis was written while I was working in the research group *Visual Analytics* at the *TIB - Leibniz Information Centre for Science and Technology* and the *L3S Research Center* of the *Gottfried Wilhelm Leibniz Universität Hannover*. I thank the German Federal Ministry of Economic Affairs and Energy (BMWi) and the German Research Foundation (DFG) for their financial support in the projects *GoVideo* (BMWi, project number: KF2135608KM3), *FaAM* (BMWi, project number: ZF4210002BZ6), *VIVA* (DFG, project number: 388420599), and *TIB-AV-A* (DFG, project number: 442397862).

My deepest gratitude goes to my PhD supervisor Prof. Dr. Ralph Ewerth for providing me with the opportunity to pursue the research directions of my choice and his advice, support, encouragement, and patience, which contributed greatly to the success of this thesis. I appreciated his open and cordial style, where I always felt welcome to talk to him about my requests and concerns, even if they were not related to the PhD.

I would also like to thank Prof. Dr.-Ing. Bodo Rosenhahn, Prof. Dr. Avishek Anand, and Prof. Dr.-Ing. Markus Fidler for their willingness to be part of my defense committee.

I am grateful to my colleagues and friends at TIB and L3S, external project partners, student assistants, and dedicated bachelor and master students for their support and the friendly working atmosphere that made my time as a PhD student very enjoyable. Special thanks are due to my office mate Matthias Springstein for his endless support, our collaborations on many publications, and the great times. I want to thank Jonas Theiner and Nils Nommensen for helping me with the review of this thesis, as well as Dr. Sherzod Hakimov and Dr. Anett Hoppe for sharing their experiences and advice during my PhD research. Big thanks to Wolfgang Gritz for the creative walks and discussions, as well as to Christian Otto and Kader Pustu-Iren for their help, paper collaborations, and many interesting and fun conversations.

I would like to express my gratitude to the staff in Information Technology and Administration of the TIB who assisted me in technical and administrative requests, as well as the Marketing team that allowed me to present my work to external audiences and at exhibitions.

Finally, a special thanks to my friends and family, particularly my parents and grandparents, as well as my beloved wife Victoria and my son Fynn, who have always supported me and believed in me even in the most difficult times. I could not have done it without you.

*Dedicated to my family*



# Contents

<b>Abstract</b>	<b>III</b>
<b>Zusammenfassung</b>	<b>V</b>
<b>Acknowledgement</b>	<b>VII</b>
<b>List of Tables</b>	<b>XIII</b>
<b>List of Figures</b>	<b>XV</b>
<b>Acronyms</b>	<b>XVII</b>
<b>Notations</b>	<b>XXI</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	3
1.1.1 Importance of Multimodal Information for Communication . . . . .	3
1.1.2 Multimodal Relations in Computer Science . . . . .	3
1.2 Existing Limitations & Challenges . . . . .	6
1.3 Contributions . . . . .	9
1.4 Thesis Structure . . . . .	12
1.5 List of Publications . . . . .	12
1.6 Achievements . . . . .	21
<b>2 Foundations</b>	<b>23</b>
2.1 Basics of Artificial Neural Networks . . . . .	23
2.1.1 Perceptron . . . . .	23
2.1.2 Multilayer Perceptron . . . . .	25
2.1.3 Network Optimization . . . . .	26
2.2 Convolutional Neural Networks for Image Classification . . . . .	27
2.2.1 Convolutional Layer . . . . .	27
2.2.2 Overview of Convolutional Neural Network Architectures . . . . .	29
2.2.3 AlexNet Architecture . . . . .	31
2.2.4 GoogLeNet Architecture . . . . .	32

## Contents

2.2.5	ResNet Architecture . . . . .	35
2.3	Natural Language Processing . . . . .	37
2.3.1	Distributional Semantics . . . . .	37
2.3.1.1	Word2Vec . . . . .	38
2.3.1.2	FastText . . . . .	40
2.3.2	Named Entity Recognition . . . . .	41
2.3.3	Named Entity Disambiguation . . . . .	42
2.4	Semantic Web & Knowledge Graphs . . . . .	45
<b>3</b>	<b>Information Extraction from Photos</b>	<b>49</b>
3.1	Event Classification of Photos . . . . .	50
3.1.1	Related Work . . . . .	51
3.1.2	Dataset and Ontology for Event Type Classification . . . . .	52
3.1.2.1	Notations & Definitions . . . . .	53
3.1.2.2	VisE-O: Visual Event Ontology . . . . .	53
3.1.2.3	Visual Event Classification Dataset . . . . .	57
3.1.3	Ontology-Driven Event Classification . . . . .	58
3.1.3.1	Classification Approach . . . . .	58
3.1.3.2	Integration of Ontology Information . . . . .	59
3.1.3.3	Inference . . . . .	60
3.1.4	Experimental Setup & Results . . . . .	61
3.1.4.1	Network Parameters . . . . .	61
3.1.4.2	Evaluation Metrics . . . . .	61
3.1.4.3	Ablation Study . . . . .	62
3.1.4.4	Experimental Results for Individual Event Types . . . . .	63
3.1.4.5	Comparisons on other Benchmarks . . . . .	65
3.1.5	Summary . . . . .	66
3.2	Geolocation Estimation of Photos . . . . .	68
3.2.1	Related Work . . . . .	70
3.2.2	Partitioning of the Earth Surface for Classification . . . . .	71
3.2.3	Geolocation Estimation using Contextual Information . . . . .	73
3.2.3.1	Environmental Scene Classification . . . . .	73
3.2.3.2	Geolocation Estimation . . . . .	74
3.2.3.3	Predicting Geolocations using Hierarchical Spatial Information . . . . .	76
3.2.4	Experimental Setup & Results . . . . .	77
3.2.4.1	Training Data . . . . .	77
3.2.4.2	Parameters for the Adaptive Partitioning using S2 Cells . . . . .	77
3.2.4.3	Scene Classification Parameters . . . . .	77
3.2.4.4	Network Training . . . . .	78
3.2.4.5	Test Setup . . . . .	79

3.2.4.6	Evaluating the Multi-Partitioning Approach . . . . .	79
3.2.4.7	Evaluating the Individual Scenery Networks . . . . .	80
3.2.4.8	Evaluating the Multi-Task Network . . . . .	82
3.2.4.9	Comparison to the State of the Art . . . . .	82
3.2.5	Demonstrator . . . . .	85
3.2.6	Summary . . . . .	86
3.3	Date Estimation of Historical Photos . . . . .	87
3.3.1	Related Work . . . . .	87
3.3.2	Date Estimation in the Wild Dataset . . . . .	89
3.3.3	Deep Learning Models for Date Estimation . . . . .	90
3.3.3.1	Regression Model . . . . .	90
3.3.3.2	Classification Model . . . . .	91
3.3.4	Experimental Setup & Results . . . . .	91
3.3.5	Summary . . . . .	93
3.4	Person Identification in News Articles of the Internet Archive . . . . .	95
3.4.1	Related Work . . . . .	96
3.4.2	Person Identification in Archived Web News . . . . .	97
3.4.2.1	Learning a Feature Representation for Faces . . . . .	98
3.4.2.2	Creating a Dictionary of Persons for a Domain . . . . .	98
3.4.2.3	Person Identification Pipeline . . . . .	100
3.4.3	Case Study & Qualitative Results . . . . .	100
3.4.3.1	Person Dictionaries & News Dataset . . . . .	101
3.4.3.2	Parameter Selection . . . . .	101
3.4.3.3	Face Recognition in Image Collections of the Internet Archive . . . . .	103
3.4.3.4	Summary . . . . .	105
<b>4</b>	<b>Multimodal Analytics using Measures of Cross-modal Consistency</b>	<b>107</b>
4.1	Related Work . . . . .	109
4.1.1	Quantification of Image-Text Relations . . . . .	109
4.1.2	Image Repurposing Detection . . . . .	111
4.2	Cross-modal Entity Consistency . . . . .	111
4.2.1	Extraction of Entities from the Text . . . . .	112
4.2.2	Extraction of Features from Photos . . . . .	113
4.2.3	Verification of Shared Cross-modal Entities . . . . .	114
4.2.3.1	Verification of Persons . . . . .	114
4.2.3.2	Verification of Locations and Events . . . . .	115
4.3	Cross-modal Context Consistency . . . . .	116
4.3.1	Text Context . . . . .	116
4.3.2	Photo Context . . . . .	117
4.3.3	Cross-modal Context Similarity . . . . .	117

## Contents

4.4	Datasets . . . . .	118
4.4.1	Manipulation Techniques . . . . .	118
4.4.2	TamperedNews Dataset . . . . .	119
4.4.3	News400 Dataset . . . . .	119
4.5	Experimental Setup & Results . . . . .	120
4.5.1	Evaluation Tasks and Metrics . . . . .	120
4.5.2	Parameter Selection . . . . .	122
4.5.3	Experimental Results . . . . .	123
4.5.3.1	Results on TamperedNews . . . . .	124
4.5.3.2	Results on News400 . . . . .	127
4.5.4	Comparison of Event Feature Descriptors . . . . .	129
4.5.5	Limitations and Dependencies . . . . .	129
4.6	Demonstrator . . . . .	131
4.7	Summary . . . . .	132
<b>5</b>	<b>Conclusions</b>	<b>133</b>
5.1	Answers to the Research Questions . . . . .	134
5.2	Limitations & Future Work . . . . .	137
	<b>References</b>	<b>141</b>
<b>A</b>	<b>Appendix</b>	<b>179</b>
A.1	Event Classification of Photos . . . . .	179
A.1.1	Detailed Dataset Statistics . . . . .	179
A.1.2	Results using other Inference Strategies . . . . .	179
A.2	Results for Geolocation Estimation . . . . .	185
A.3	Results on other Subsets of TamperedNews and News400 . . . . .	194
	<b>Curriculum Vitae</b>	<b>199</b>

# List of Tables

2.1	Details of the <i>GoogLeNet</i> architecture . . . . .	34
2.2	Details of <i>ResNet</i> architectures with varying depth . . . . .	37
3.1	Statistics for the <i>Visual Event Ontology</i> and the <i>Visual Event Classification Dataset</i> . . . . .	56
3.2	Event classification results on the <i>VisE-Bing</i> test dataset . . . . .	62
3.3	Event classification results on several benchmark datasets . . . . .	66
3.4	Number of geographical cells for different Earth partitionings . . . . .	78
3.5	Scene classification results on the <i>Places365</i> validation dataset . . . . .	78
3.6	Notations of the geolocalization approaches . . . . .	79
3.7	Number of images for two geolocalization benchmarks for different scenes . . . . .	81
3.8	Scene classification results on <i>Places365</i> for different Multi-Task Networks . . . . .	81
3.9	Parameters used by approaches for geolocation estimation . . . . .	83
3.10	Geolocation estimation results on the <i>Im2GPS</i> and <i>Im2GPS3k</i> test datasets . . . . .	84
3.11	Date estimation results on the <i>Date Estimation in the Wild</i> test dataset . . . . .	93
3.12	Number of photos and faces found in archived news articles . . . . .	102
3.13	Person identification results for different data filtering methods . . . . .	102
4.1	Dataset statistics for the <i>TamperedNews</i> and <i>News400</i> dataset . . . . .	119
4.2	Results using different operator functions on <i>TamperedNews</i> (Top-50%) . . . . .	122
4.3	Results using different image sources on <i>TamperedNews</i> (Top-50%) . . . . .	123
4.4	Results for cross-modal consistency on <i>TamperedNews</i> (Top-50%) . . . . .	124
4.5	Results for cross-modal consistency for a selection of locations on <i>Tampered-News</i> (Top-50%) . . . . .	126
4.6	Results for cross-modal consistency for a selection of events on <i>Tampered-News</i> (Top-50%) . . . . .	127
4.7	Results for cross-modal consistency on all verified <i>News400</i> documents . . . . .	128
4.8	Comparison of a place and event classification approach for the quantification of cross-modal event consistency . . . . .	130
A.1	Results on the <i>VisE-Bing</i> test dataset using the probabilities $\hat{\mathbf{y}}_L^o$ . . . . .	183
A.2	Results on the <i>VisE-Bing</i> test dataset using the probabilities $\hat{\mathbf{y}}_L^{cos}$ . . . . .	184
A.3	Results on the <i>Im2GPS</i> test dataset of all images . . . . .	186

List of Tables

A.4 Results on the *Im2GPS* test dataset of all images classified as *indoor* . . . . . 187

A.5 Results on the *Im2GPS* test dataset of all images classified as *natural* . . . . . 188

A.6 Results on the *Im2GPS* test dataset of all images classified as *urban* . . . . . 189

A.7 Results on the *Im2GPS3k* test dataset of all images . . . . . 190

A.8 Results on the *Im2GPS3k* test dataset of all images classified as *indoor* . . . . . 191

A.9 Results on the *Im2GPS3k* test dataset of all images classified as *natural* . . . . . 192

A.10 Results on the *Im2GPS3k* test dataset of all images classified as *urban* . . . . . 193

A.11 Results for cross-modal consistency on *TamperedNews* (Top-25%) . . . . . 195

A.12 Results for cross-modal consistency on all *TamperedNews* documents . . . . . 196

A.13 Results for cross-modal consistency on the Top-50% verified *News400* documents . . . . . 197

# List of Figures

1.1	Exemplary multimodal news articles with cross-modal entity similarities computed by the proposed system . . . . .	2
1.2	Taxonomy and examples for eight image-text relationships . . . . .	4
1.3	Examples of image-text relations in advertisements . . . . .	5
1.4	Reference and test images of the <i>Multimodal Entity Image Repurposing</i> dataset . . . . .	6
2.1	Mathematical model of the perceptron . . . . .	24
2.2	Popular activation functions used in neural networks . . . . .	25
2.3	Illustration of a multilayer perceptron . . . . .	25
2.4	Illustration of a two-dimensional convolution and a convolutional layer . . . . .	28
2.5	Strided convolution and pooling . . . . .	29
2.6	Illustration of the <i>AlexNet</i> architecture . . . . .	31
2.7	Illustration of the <i>Inception</i> module . . . . .	33
2.8	Illustration of two residual block variants . . . . .	36
2.9	Illustration of two <i>Word2Vec</i> models . . . . .	39
2.10	Exemplary output of <i>spaCy</i> for <i>Named Entity Recognition</i> . . . . .	41
2.11	Exemplary output of <i>Ambiverse</i> for <i>Named Entity Disambiguation</i> . . . . .	43
2.12	<i>Named Entity Disambiguation</i> with <i>Wikifier</i> using a mention-entity graph . . . . .	45
2.13	Exemplary <i>Resource Description Framework</i> graph . . . . .	46
3.1	Exemplary subset of the <i>Ontology</i> and photos of the proposed <i>Visual Event Classification Dataset</i> . . . . .	51
3.2	Exemplary subset of the initial and final <i>Ontology</i> . . . . .	54
3.3	Event classification results for a selection of <i>Event Nodes</i> on <i>VisE-Bing</i> . . . . .	64
3.4	Qualitative event classification results on <i>VisE-Wiki</i> . . . . .	64
3.5	Workflow and sample images for geolocation estimation . . . . .	69
3.6	Partitioning of the Earth into geographical cells . . . . .	72
3.7	Pipeline of the geolocation estimation frameworks . . . . .	73
3.8	Comparison of the geolocation approaches trained with and without multiple partitionings . . . . .	80
3.9	Qualitative results for different partitionings as well as hierarchical result . . . . .	80
3.10	Comparison of the <i>Individual Scenery Networks</i> to the baseline approaches . . . . .	81
3.11	Comparison of the <i>Multi-Task Network</i> to a baseline approach . . . . .	82

List of Figures

3.12	Screenshot of the demonstrator for geolocation estimation . . . . .	85
3.13	Example images from the <i>Date Estimation in the Wild</i> dataset . . . . .	88
3.14	Number of crawled images and the accuracy of the provided timestamps for each year in the <i>Date Estimation in the Wild</i> dataset . . . . .	90
3.15	Workflow of the proposed person identification approach for news articles in the <i>Internet Archive</i> . . . . .	96
3.16	Exemplary results for person identification in the <i>Internet Archive</i> . . . . .	104
4.1	Test and reference images of the <i>Multimodal Entity Image Repurposing</i> dataset and two news from <i>BreakingNews</i> with outputs of the proposed system . . . .	108
4.2	Workflow for the quantification of <i>Cross-modal Entity Similarities</i> . . . . .	112
4.3	Workflow for the quantification of the <i>Cross-modal Context Similarity</i> . . . .	117
4.4	Cross-modal similarity values for person, location, and event entities . . . .	121
4.5	Qualitative results for cross-modal document verification . . . . .	125
4.6	Screenshot of the demonstrator for multimodal news analytics . . . . .	131
A.1	Number of training images for all <i>Leaf Event Nodes</i> in the <i>Visual Event Classification Dataset</i> . . . . .	180
A.2	Number of images for all <i>Leaf Event Nodes</i> in the <i>VisE-Bing</i> test dataset. . .	181
A.3	Number of images for all <i>Leaf Event Nodes</i> in the <i>VisE-Wiki</i> test dataset. . .	182



# Acronyms

- AP** *Average Precision*. 121, 124, 128, 195–197
- API** *Application Programming Interface*. 89, 112, 119
- AUC** *Area Under Receiver Operating Curve*. 121–124, 126–128, 130, 195–197
- BERT** *Bidirectional Encoder Representations from Transformers*. 38
- BoW** *Bag of Words*. 42
- CBOW** *Continuous Bag-of-Words*. 38, 39
- CMCS** *Cross-modal Context Similarity*. XVI, 116, 117
- CMES** *Cross-modal Event Similarity*. 116
- CMI** *Cross-modal Mutual Information*. 4, 5, 9, 11, 49, 107–110, 132, 133
- CMLS** *Cross-modal Location Similarity*. 116, 126, 130
- CMPS** *Cross-modal Person Similarity*. 112, 115, 121
- CMS** *Cross-modal Similarity*. 114, 116, 125, 132
- CNN** *Convolutional Neural Network*. 23, 26–32, 35, 52, 58, 59, 68, 69, 71, 74, 75, 78, 83, 86–88, 90, 91, 93, 96–101, 105, 113, 115, 117, 130, 136, 139
- CORE** *Computing Research & Education*. 21
- CRF** *Conditional Random Field*. 41, 42
- CV** *Computer Vision*. 7, 23
- ELMo** *Embeddings from Language Models*. 38, 42
- FNV** *Fowler-Noll-Vo*. 41
- GAN** *Generative Adversarial Network*. 97
- GCD** *Great Circle Distance*. 79, 82, 84, 85, 118, 122, 124–126, 128, 186–193, 195–197
- GCNN** *Graph Convolutional Neural Network*. 52, 67, 138
- GloVe** *Global Vectors for Word Representation*. 37
- GPS** *Global Positioning System*. 69–73, 76, 79, 82, 83, 86

## Acronyms

- GPT** *Generative Pre-training*. 38
- HMM** *Hidden Markov Model*. 41
- HOG** *Histogram of Oriented Gradients*. 99
- HTML** *Hypertext Markup Language*. 46
- HTTP** *Hypertext Transfer Protocol*. 47
- ILSVRC** *ImageNet Large Scale Visual Recognition Challenge*. 29–32, 34, 36, 58, 77, 78, 83, 89, 91, 102, 118
- ISN** *Individual Scenery Network*. 75–85, 186–193
- LFW** *Labeled Faces in the Wild*. 96, 102, 105, 113, 122
- LSTM** *Long short-term memory*. 26, 38, 42
- MEIR** *Multimodal Entity Image Repurposing*. XV, XVI, 6, 108, 118
- MLP** *Multilayer Perceptron*. 25, 27, 32
- MP16** *MediaEval Placing Task 2016*. 77
- MS-Celeb-1M** *Microsoft-Celebrity-1M*. 96, 98, 101
- MTN** *Multi-Task Network*. XV, 76, 79, 81, 82, 186–193
- MvMF** *Mixture of von-Mises Fisher*. 71, 83
- NAS** *Neural Architecture Search*. 30
- NED** *Named Entity Disambiguation*. XV, 37, 42–45, 95, 112, 113, 125, 129
- NER** *Named Entity Recognition*. XV, 37, 41, 42, 44, 112, 113, 138
- NER & NED** *Named Entity Recognition and Disambiguation*. 7, 12, 23, 44, 98, 108, 111, 112, 114, 131–134, 137
- NLP** *Natural Language Processing*. 7, 12, 23, 26, 31, 37, 41, 42
- OWL** *Web Ontology Language*. 46, 47
- RDF** *Resource Description Framework*. XV, 46, 47
- RED** *Rare Event Dataset*. 50, 65, 66
- ReLU** *Rectified Linear Unit*. 24, 25, 27, 31, 32, 34, 36, 37
- SARE** *Stochastic Attraction and Repulsion*. 71
- SC** *Semantic Correlation*. 4, 5, 109, 110

- SE** *Squeeze-and-Excitation*. 30
- SGD** *Stochastic Gradient Descent*. 61, 78, 91, 102
- SIFT** *Scale-invariant Feature Transform*. 52
- SocEID** *Social Event Image Dataset*. 50, 65, 66
- SPARQL** *SPARQL Protocol and RDF Query Language*. 47
- SVM** *Support Vector Machine*. 41, 66, 99, 111
- TF-IDF** *Term Frequency–Inverse Document Frequency*. 42
- TIB** *Leibniz Information Centre for Science and Technology*. 85, 131
- ULMFiT** *Universal Language Model Fine-tuning*. 38
- URI** *Uniform Resource Identifier*. 42, 43, 46, 47
- URL** *Uniform Resource Locator*. 119
- VA** *Verification Accuracy*. 120, 124, 128, 130, 195–197
- VGG** *Visual Geometry Group*. 29, 32, 83, 111
- VisE-D** *Visual Event Classification Dataset*. XIII, XV, XVI, 11, 50–52, 66, 130, 135, 179, 180
- VisE-O** *Visual Event Ontology*. XIII, 11, 50, 52, 66, 113, 135, 138
- WIDER** *Web Image Dataset for Event Recognition*. 50, 52, 65, 66
- YFCC100M** *Yahoo Flickr Creative Commons 100 Million dataset*. 77, 78



# Notations

In this chapter the description of notations used in this thesis are presented. In general, the notation from Goodfellow et al. [79] are used.

## Scalars, Arrays, and Sets

$a$	A scalar (integer or real)
$\mathbf{a}$	A one-dimensional vector
$\mathbf{A}$	A two-dimensional matrix
$\mathbf{A}$	A three-dimensional tensor
$\mathbb{A}$	A set
$ \mathbb{A} $	The number of items in set $\mathbb{A}$
$\mathbb{R}$	The set of real numbers
$\{0, 1\}$	A set containing 0 and 1
$\{0, \dots, n\}$	A set containing all integers from 0 to $n$
$\mathcal{G}$	A graph
$\ \mathbf{a}\ _1$	The <i>Manhattan (or l1) norm</i> of a vector $\mathbf{a}$
$\ \mathbf{a}\ _2$	The <i>Euclidean (or l2) norm</i> of a vector $\mathbf{a}$

## Indexing

$a_i$	Element with index $i$ of vector $\mathbf{a}$
$A_{i,j}$	Element in row $i$ and column $j$ of matrix $\mathbf{A}$

## Datasets

$\mathbb{X}$	A set of training examples
$\mathbf{X}^{(i)}$ or $\mathbf{x}^{(i)}$	Input matrix $\mathbf{X}^{(i)}$ or vector $\mathbf{x}^{(i)}$ of the $i$ -th element in the training set $\mathbb{X}$
$\mathbf{y}^{(i)}$ or $y^{(i)}$	Target vector $\mathbf{y}^{(i)}$ or scalar $y^{(i)}$ of the $i$ -th element in the training set $\mathbb{X}$



# 1 Introduction

With the widespread availability and use of digital environments, the World Wide Web plays an essential role in the dissemination of information and news. In particular, social media platforms like *Twitter* (<https://twitter.com/>) allow users to follow worldwide events and are a popular source of information [44, 214, 256]. Typically, multimedia articles and news published on the World Wide Web include different *modalities*, such as photos, text, video, or sound. According to Guo et al., a "*modality refers to a particular way or mechanism of encoding information*" [85]. The various modalities act as mechanisms that can describe different aspects of the same object to convey information about objects in the world [85].

An essential aspect of understanding multimodal messages is the complex interplay between different modalities [31], e.g., the semantic correlation and the number of co-occurring concepts or entities. Due to the rapidly growing amount of multimodal articles and news on the Web, automated approaches for multimodal content analysis are becoming increasingly important. The quantification of cross-modal relations in news articles is particularly challenging as they are typically centered around real-world entities such as persons, locations, and events. Moreover, new entities and topics can emerge every day. A fully-automatic system capable of quantifying cross-modal relations of entities (e.g., persons, locations, and events) between a photo and its associated text (illustrated in Figure 1.1) can enable many tasks and applications in news analytics, semiotics, and multimedia retrieval.

Approaches that automatically quantify cross-modal relations allow users to efficiently explore news articles and other multimodal documents, e.g., to reveal related parts of the text to the accompanying photo. Named entities can be linked to knowledge bases such as *Wikipedia* or *Wikidata* [268] to provide additional information, which can be helpful for readers that might not be familiar with a news topic and the mentioned entities. The quantification of entity relations can also help semioticians, linguists, and communication scientists investigate the inter-dependencies of photo-text pairs, for example, to evaluate the amount of shared and contrasting entities between both modalities.

Another application of approaches that quantify cross-modal relations is the evaluation of inconsistent entity relations between photos and text, which can be helpful to detect misinformation in news. Media and individual users may copy, paraphrase, or manipulate news stories and use social media platforms to disseminate an intended narrative or opinion. *Fake news*, i.e., articles that deliberately spread rumors or misleading information, have

## 1 Introduction

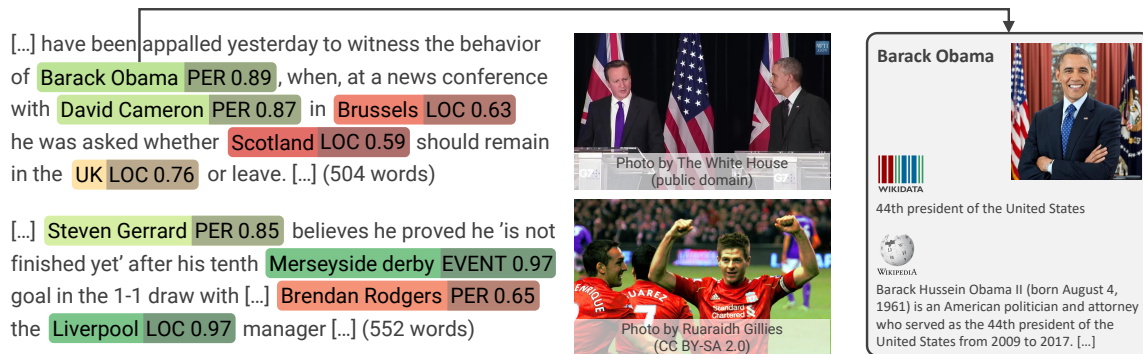


Figure 1.1: **Left:** Exemplary multimodal news articles with cross-modal entity similarities for locations (LOC), persons (PER), and events (EVENT) computed by the proposed approach presented in Chapter 4 of this thesis. **Right:** Entity information extracted from *Wikipedia* and *Wikidata* [268]. A web application is available at: <https://labs.tib.eu/newsanalytics>. Photos are replaced with similar ones depicting the same entity relations due to image copyright restrictions. Links to the original documents can be found on: [https://github.com/TIBHannover/cross-modal\\_entity\\_consistency/tree/master/supplemental\\_material](https://github.com/TIBHannover/cross-modal_entity_consistency/tree/master/supplemental_material)

become a critical problem in recent years and have even been used repeatedly, e.g., during the *2016 United States elections* [13, 38]. In some cases, measures of cross-modal consistency can be an important first step towards supporting human assessors and expert-oriented fact-checking efforts such as *PolitiFact*<sup>1</sup> and *Snopes*<sup>2</sup> to identify *fake news*. Examples for news articles that report about certain events at a claimed location but use photos of another location have been reported in the media:

- Example 1: "COVID-19: Old Video from Azerbaijan Shared as Lockdown in Spain" (archived web link from 12th April 2020 <https://bit.ly/3wSU5kZ>)
- Example 2: "CBS admits crowded New York hospital was actually in Italy" (archived web link from 1st February 2021: <https://bit.ly/3wS5TE6>)

While the aforementioned applications focus on the evaluation of individual multimodal documents, the quantification of cross-modal relations also allows news retrieval from large multimedia collections or news archives. For example, news with low cross-modal consistency can be retrieved that are potentially *check-worthy* for fact-checking efforts. On the other hand, retrieval of multimodal articles with high cross-modal consistency more likely provides users with credible news articles. Moreover, entity relations, e.g., between (public) persons and events, can be indexed for information retrieval, allowing users to retrieve documents that are likely to represent one or multiple specified entities in both photo and text.

<sup>1</sup><https://www.politifact.com/>

<sup>2</sup><https://www.snopes.com/>



## 1.1 Background

The interplay of different modalities or information channels has been studied in communication and computer science. This section briefly summarizes the importance of multimodal relations for communication as well as related work in computer science.

### 1.1.1 Importance of Multimodal Information for Communication

Different modalities such as photo and text, diagrams and text, or video and speech (audio) can help convey information more efficiently or attract attention [31, 153]. Therefore, multimodal information is essential for different media types such as television, books, or social media across various domains, e.g., education, entertainment, or news. Each information channel carries specific information. Their combination enables the communication of a multimodal message that can yield additional information and sometimes a new level of meaning referred to as *meaning multiplication* [31, 134]. For example, in static multimodal articles, the role of the photo can range from decorative (with little or no information compared to the text) over depicting rich information enhancements (important or additional details) to even misleading (contradictory) visual information. Examples of different relations between photos and text according to Otto et al. [185] are shown in Figure 1.2.

In the past decades, linguists, semioticians, and communication scientists have been investigating the visual/verbal divide and attempted to define types of interrelations between visual (e.g., photos, diagrams, or video) and verbal information (e.g., text or speech) using suitable taxonomies to describe the complex interplay between different modalities [30, 88, 153, 155, 162, 262]. According to Bateman [31], the consideration of relationships between the modalities, such as the semantic correlation and mutual concepts, is crucial to understand and evaluate the overall message and meaning of multimodal documents.

### 1.1.2 Multimodal Relations in Computer Science

Computational models require rich features from both modalities to determine relations between photo and text. Smeulders et al. [237] identified the *semantic gap* as one of the biggest challenges for image retrieval applications and defined it as "*the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation*" [237]. Twenty years ago, this problem was mainly related to the fact that computer vision approaches were only able to describe photo content based on low-level features (e.g., color, texture, shape). Substantial progress has been made in recent years due to the introduction of deep learning approaches [91, 126, 285] for many computer vision tasks, such as object and scene recognition, that are capable of extracting visual concepts as semantic, high-level features [303]. In recent years,

# 1 Introduction

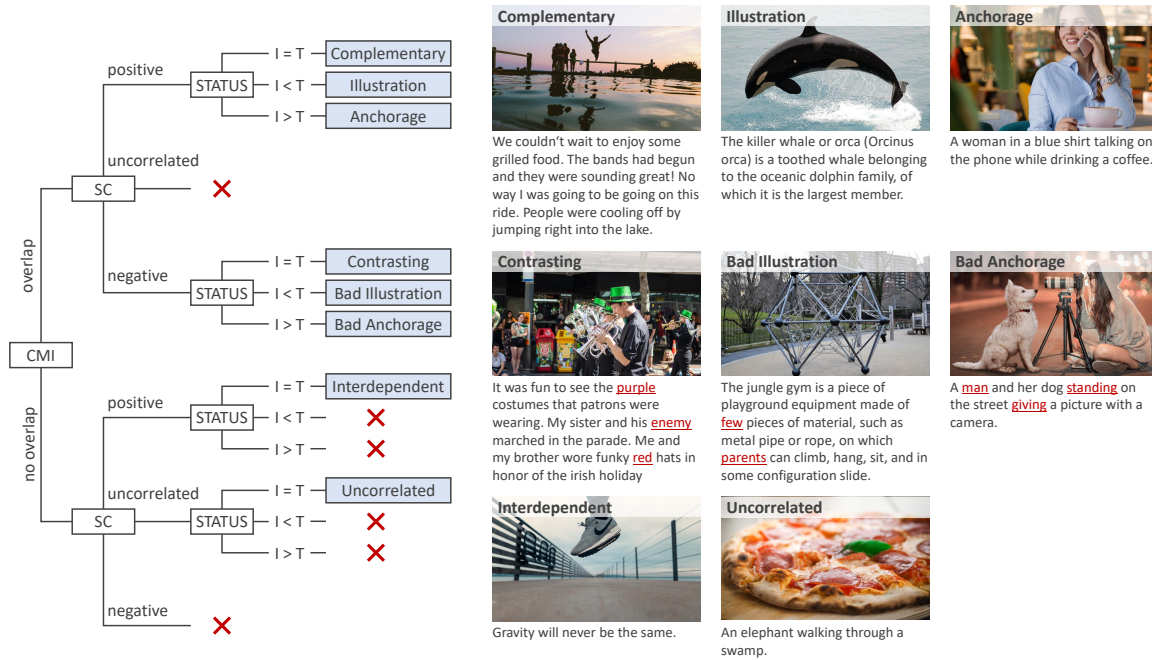


Figure 1.2: **Left:** Categorization of eight image-text relations based on three computable dimensions, namely *Cross-modal Mutual Information* (CMI), *Semantic Correlation* (SC), and *Status* according to Otto et al. [185]. For *Status*, both modalities can have the same relative importance to the multimodal message ( $I = T$ ), the image can be superordinate to the text ( $I > T$ ), or vice versa ( $I < T$ ). Note that there are no hierarchical relations implied and that discarded subtrees are marked with a red cross. **Right:** Examples for each type of image-text relations [68, 185]. Underlined red text is contradictory to the photo content.

researchers have tackled challenging cross-modal tasks such as image captioning and visual question answering [16, 19, 118, 138]. However, the proposed solutions focus on answers and precise descriptions of the visual content based on (rather generic) concepts such as objects, actions, or persons. Thus, they do not aim to describe more complex image-text relations that are relevant in practice and can include complementary information [96, 306]. Moreover, these approaches lack the capabilities of (human) scene understanding, including the interpretation of gestures, symbols, and other contextual information, and are unable to capture the deeper semantic information and meaning of images. Thus, even with the availability of high-level features from recent deep learning approaches, the *semantic gap* is still a critical challenge, especially for automated solutions that aim to understand complex relationships of multimodal information.

**Quantification of Image-Text Relations:** So far, only a few approaches [96, 127, 185, 294, 306] have been presented that aim to bridge the *semantic gap* with respect to multimodal relationships. Henning and Ewerth [96, 97] proposed the first approach that quantifies image-text relations using two computable dimensions: *Cross-modal Mutual Informa-*



Figure 1.3: Examples of image-text relations similar to Zhang et al. [306]. Due to image copyrights restrictions, the original advertisement images (can be found in Zhang et al. [306]) were replaced with similar ones. The text and images in examples A and B share the same meaning (parallel relationship), although the modality-specific information is not equivalent. Non-parallel relationships are shown in examples C and D. While in C, the text does not mention the concept of smoking and the image does not establish a connection to potential health-related issues, the information in D seems contradictory since the fuzzy duckling is supposed to be "not soft enough".

*tion* (CMI) and *Semantic Correlation* (SC). While CMI focuses on the mutual cross-modal presence of concepts, SC describes the shared meaning of image and text. SC is comparable to the different levels of semantic relations (low, medium, high) in the taxonomy of Marsh and White [153]. Henning and Ewerth [96, 97] train an autoencoder that reconstructs the multimodal input to learn a low-dimensional representation of the image-text pairs. Subsequently, the encoder-network is used with labeled training data to train a classifier that determines scores for CMI and SC. Otto et al. [185] suggested a third dimension called *Status*, initially introduced by Barthes [30], which determines the relative importance of each information channel to the multimodal message. They suggest training a multimodal deep learning approach that quantifies scores for CMI, SC, and *Status* that can be used to categorize image-text relations into eight distinct classes, as shown in Figure 1.2. Zhang et al. [306] investigate the relation of images and embedded slogans in advertisements to predict parallel or non-parallel relationships based on a variety of visual and textual features, as well as methods that analyze the semantics within and across channels. As illustrated in Figure 1.3, text and image in parallel relationships are considered either redundant or complementary and convey the same message, while in non-parallel relations, one modality can be contradictory, ambiguous, or decorative compared to the other. Ye et al. [294] further extended this approach by interpreting the rhetoric of visual advertisements using cross-modal embeddings and image embeddings for symbol regions. Kruk et al. [127] identified that *Instagram* posts often contain complex image-text relations and propose a deep multimodal classifier to determine the author's intent as well as the semiotic and contextual relationships between image and caption in *Instagram* posts.

## 1 Introduction

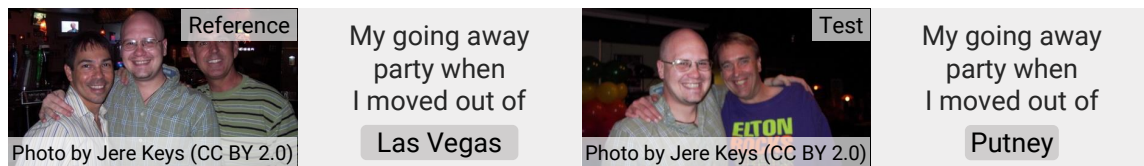


Figure 1.4: Reference and test images of the *Multimodal Entity Image Repurposing* (MEIR) dataset [219] and corresponding texts with original and manipulated entities. Please note that these examples have significantly shorter text and fewer entities compared to real-world news illustrated in Figure 1.1. Furthermore, the reference dataset is very closely related to the images used for testing.

However, the aforementioned methods [96, 127, 185, 294, 306] do not explicitly consider cross-modal relations of named entities such as public figures, locations, and events as shown in Figure 1.1. As a result, the occurrence of, for example, an arbitrary person in text and image can already result in a valid cross-modal relation since it matches the concept *person* even if the identity of this person differs between the modalities. Many real-world applications require more differentiated measures that are able to evaluate the cross-modal consistency of every individual named entity.

**Image Repurposing Detection:** In recent years, solutions for image repurposing detection [114, 115, 219] addressed a similar problem and have evaluated the consistency between an image and its associated entities (e.g., persons, locations, or organizations) claimed in the text or metadata. These approaches rely on multimodal deep learning techniques that require appropriate datasets containing non-manipulated pairs of image and text. Such datasets are hard to collect automatically since they need to be verified for valid cross-modal relations. In addition, these methods are limited to the verification of the cross-modal occurrence of entities that appear in the training and reference data. Thus, these supervised approaches cannot cope well with the dynamic nature of news and other multimodal documents that can cover new topics and entities every day. Moreover, experimental evaluation has been performed on images with relatively short image captions [114, 219] or existing metadata [115], which do not reflect real-world characteristics, as illustrated in Figure 1.4.

## 1.2 Existing Limitations & Challenges

Multimedia articles and news are typically centered around entities such as persons, locations, or events, and new topics emerge every day. As discussed in the previous section, current solutions that quantify image-text relation [96, 127, 185, 294, 306] do not focus on evaluating the cross-modal occurrence of specific named entities and supervised approaches for image repurposing detection [114, 115, 219] cannot handle the vast and ever-growing

entity diversity. Whether an unsupervised approach can address these existing limitations is the main research question of this thesis:

**Research Question 1:** “*Can we develop an unsupervised approach for the quantification of cross-modal entity consistency in news articles? What are the advantages, limitations, and challenges in comparison to supervised approaches?*”

However, such an unsupervised approach requires combining solutions from both *Natural Language Processing* (NLP) and *Computer Vision* (CV) to extract meaningful information from image-text pairs and leads to further research questions.

Tremendous progress has been made in NLP, and solutions for *Named Entity Recognition and Disambiguation* (NER & NED) [40, 98, 125, 281] have achieved promising results in identifying named entities in a text. Also, recent advancements in deep learning have led to significant progress in many computer vision areas such as object classification [92, 126, 255, 322], place classification [315], and face recognition [59, 228, 245, 282]. These approaches can help verify the cross-modal presence of object classes (e.g., types of cars, animals, and food), place categories (e.g., beach, church, and plaza), and persons (e.g., public figures). However, it still remains challenging to identify the vast amount and diversity of persons mentioned in the news and other multimodal documents every day. Moreover, evaluating the cross-modal occurrence of events, locations (latitude and longitude), and dates requires extracting rich geospatial, temporal, and spatio-temporal information from photos, but comparatively few approaches with several limitations were presented in the related areas of event classification [4, 5, 8, 86, 273, 287], geolocation estimation [89, 90, 229, 267, 279], and date estimation [72, 77, 189, 221]. The individual limitations and challenges for the related computer vision areas are discussed in the following.

Only a few datasets [8, 66, 287] have been introduced for the **classification of events or event types** in photos. These datasets are relatively small and disregard many event types relevant to the news, like *epidemics* or *natural disasters*. Due to the lack of a large-scale image dataset, recent approaches on event classification focus on ensemble models [5, 6, 273] and the integration of descriptors from local image regions [4, 78, 86, 287]. However, the models used in these ensemble approaches are trained for related tasks such as object and place (or scene) recognition and lack important features for event classification. For example, some event types such as *parades* and *protests* can be similar in terms of place (e.g., street or plaza) and object information (e.g., humans) but can significantly differ in the sentiment of the depicted persons.

**Geolocation estimation** of photos is very challenging due to the considerable amount of intra-class (e.g., different daytimes, objects, or camera settings) and extra-class variations (e.g., architecture, flora and fauna, or style of interior furnishings). Therefore, most approaches simplify photo geolocalization by restricting the problem to urban photos of, for example, well-known landmarks and cities [20, 142, 226, 280, 302, 313] or natural areas like

## 1 Introduction

deserts or mountains [24, 225, 261]. Only a few proposals [89, 90, 229, 267, 279] treat the task at global scale without any prior assumptions. However, according to Vo et al. [267], a single deep neural network, even with tens of millions of parameters, can struggle to memorize the visual appearance of locations in the entire world. Moreover, the photos taken all over the world are very unevenly distributed [279], making it difficult to train a deep learning approach using regression-based loss functions. To prevent bias, previous solutions [229, 267, 279] have divided the Earth into partitions with a similar number of images to treat geolocation as a classification problem. However, the choice of granularity for this partitioning entails a trade-off problem [229]. While fewer but larger (in terms of geographic area) cells decrease the geospatial resolution of the model outputs, more but smaller cells are more challenging to distinguish and also make the model susceptible to overfitting due to the lower number of available training images per cell. Moreover, geographic information at different spatial resolutions is important to identify locations of varying granularity (e.g., buildings, cities, or countries) relevant in news.

**Date estimation** approaches that aim to predict the acquisition year of (historical) photographs have not attracted much attention in recent years. Current solutions and datasets on date estimation are restricted to historical color photographs [72, 154, 189] or specific concepts such as cities [227], cars [133], persons [77, 221], or historical documents [94, 139]. No large-scale datasets and approaches are available for unconstrained date estimation of black-and-white and color photographs depicting arbitrary motifs.

Unlike the previously mentioned tasks, **face recognition** is a very well-studied computer vision area. Previous solutions [59, 228, 245, 282] have proposed deep learning models for representation learning to verify persons based on reference (example) images that depict them. However, acquiring these reference images automatically from the Web, e.g., using image search engines such as *Google Images*, poses additional challenges such as possible selection biases or the acquisition of irrelevant photos that portray other entities instead.

Overall, the lack of datasets and the limitations of previous solutions in these computer vision areas lead to:

**Research Question 2:** “How suitable are deep learning approaches in recognizing events, locations, dates, and persons in photos specifically with respect to information extraction from news articles?”

As mentioned in Section 1.1, one of the biggest challenges for information extraction in images is the *semantic gap* [237]. Many approaches in computer vision areas such as object classification [92, 126, 255, 322] and face recognition [59, 228, 245, 282] work on a descriptive level. However, tasks such date estimation, geolocation estimation, or event classification require a more profound scene understanding. For example, estimating a photo’s geolocation based on its visual content requires the consideration of various high-level features, e.g., architecture, street signs, flora and fauna, or style of interior furnishings depending on the

environmental context. Similarly, approaches to date estimation can benefit from additional features that describe the fashion style or types of cars. The classification of (news) events can benefit from event relations to learn the fundamental differences of event types in different domains such as politics, health, and sports. However, as mentioned above, solutions in related computer vision areas may lack scene interpretation capabilities, world knowledge, and other contextual information that are important for these tasks, resulting in:

**Research Question 3:** “*Can contextual information, derived from knowledge bases or related tasks like scene classification, improve image recognition and interpretation and provide better performance for computer vision tasks?*”

### 1.3 Contributions

The goal of this thesis is to answer the research questions mentioned above by presenting an **unsupervised approach** that is **applicable to real-world news articles** and other multimodal documents and provides **differentiated cross-modal relations for specific named entities** such as public figures, locations, and events. Such an approach is a crucial step towards the quantification of fine-grained *Cross-modal Mutual Information* (CMI) of multimodal documents. As illustrated in Figure 1.2, more reliable measures for CMI can improve the categorization of image-text relations. As pointed out in the previous section, the extraction of information from photos is an important prerequisite for this task. Thus, another goal of this thesis is to **improve information extraction from photos**. The contributions can be grouped into the following two categories.

**Information Extraction from Photos:** As discussed in Section 1.2, current solutions for date estimation [72, 77, 189, 221], geolocation estimation [89, 90, 229, 267, 279], and event classification [4, 5, 8, 86, 273, 287] have several limitations, such as the lack of appropriate training datasets and insufficient scene interpretation capabilities. This motivates the approaches presented in this thesis.

Existing datasets for **event classification** in photos [8, 66, 287] are relatively small and contain only a few event types relevant in news. We introduce a large-scale dataset for event classification that comprises 570,540 images along with an ontology of 148 newsworthy event types extracted from *Wikidata* [268]. To date, the dataset covers the most diverse and complete set of event classes. Unlike previous work that either uses ensemble models [5, 6, 273] trained for similar tasks or integrates descriptors from local image regions [4, 78, 86, 287], the dataset allows for the training of deep learning models from scratch. Besides, ontology-driven deep learning models based on novel weighting schemes and loss functions are presented that leverage event relations extracted from structured knowledge graph information. In this way, the network is provided with additional contextual information to understand and learn from

## 1 Introduction

the fundamental similarities and differences of various event types, including sports, social and cultural events, natural disasters, and health crises. Experimental results on several benchmark datasets, including two novel test sets, have demonstrated the superiority of this approach, outperforming baselines trained without structured ontology information.

Although related work [89, 90, 146, 229, 267, 279] has presented powerful deep learning models for **geolocation estimation**, it remains a challenging task due to geographic ambiguities, the vast amount of intra- and extra-class variations, and the trade-off problem introduced by dividing the Earth into geographic cells. In this thesis, novel deep learning approaches for geolocation estimation are suggested that combine the outputs of hierarchical cell partitions of different granularity and consider the environmental context (e.g., indoor, urban, rural) of the photo. Hierarchical cell partitions alleviate the entailed trade-off problem since the network learns features at multiple geographical scales and allows for hierarchical predictions using the outputs of each partitioning. Besides, it can learn geographic features at different spatial resolutions, which is important because news articles mention locations of varying granularity, e.g., buildings, cities, or countries. Finally, a complementary deep learning approach for scene classification is incorporated to train individual expert networks for different environmental settings (e.g., indoor, urban, rural) on photos with fewer intra- and extra-class variations, which allows them to learn more discriminative features for the particular setting. Experimental results have shown that the proposed approaches outperform strong baselines from the literature [229, 267, 279] on popular benchmark datasets while using a significantly smaller number of training images.

A new dataset for **date estimation** is presented that comprises more than one million images from *Flickr* captured between 1930 and 1999. Unlike previous datasets, the *Date Estimation in the Wild* dataset is neither restricted to specific concepts [77, 94, 133, 139, 221, 227] nor to historical color photographs [72, 154, 189]. Two deep learning methods are proposed that use different loss functions to treat date estimation as a regression and classification problem, respectively. Experimental results on a novel test dataset have shown the superiority of both approaches compared to human annotators.

While tremendous progress has been made in **face recognition** [59, 228, 245, 282], many real-world applications can pose additional challenges. For example, persons mentioned in the media are usually not known in advance. We present a multimedia retrieval approach to identify (public) persons and their joint co-occurrences with other individuals in photos extracted from news articles in the *Internet Archive* (<https://archive.org/>), which is a digital library that has been capturing (multimedia) web pages since the mid-1990s. We show how to automatically create a dictionary containing the most relevant persons for a given time period and domain (e.g., politics or entertainment). Furthermore, we propose an unsupervised approach that can identify persons without manual effort from the user. To achieve this goal, example images for the relevant persons are crawled automatically from the Web. An additional filtering step is since the image search results can contain irrelevant



photos that can portray multiple or different persons. A case study has demonstrated the feasibility of the solution for person identification in news photos.

**Measures of Cross-modal Entity Consistency:** This thesis presents an automatic system for the quantification of cross-modal entity consistency. We go beyond existing approaches that quantify image-text relations [96, 127, 185, 294, 306] by providing more differentiated measures that allow for an evaluation of *Cross-modal Mutual Information* (CMI) based on individual and more specific entities. Unlike related work on image repurposing detection [114, 115, 219], the system is entirely unsupervised and does not rely on any pre-defined reference or training data. To the best of our knowledge, we present the first system that is *applicable to real-world news articles* by tackling several news-specific challenges such as the excessive length of news documents, entity diversity, and unrelated reference (example) images. Based on visual features extracted by appropriate deep learning approaches, novel measures for different entity types (persons, locations, and events) as well as for a more general news context are introduced to quantify cross-modal relationships between photo and text. The feasibility of the proposed approach is demonstrated on two novel datasets, namely *TamperedNews* and *News400*, covering different languages and domains.

**The main contributions of this thesis can be summarized as follows:**

- *Visual Event Ontology* (VisE-O) and *Visual Event Classification Dataset* (VisE-D) for event type classification comprising 570,540 images for 148 event types
- Ontology-driven event classification approach including novel loss functions and weighting schemes that outperforms baseline systems that were not trained with structured ontology information
- Geolocation approach that leverages contextual geographical and environmental information with state-of-the-art performance on two benchmark datasets
- Large-scale *Date Estimation in the Wild* dataset comprising more than one million images for unrestricted date estimation for the period from 1930 - 1999
- Regression and classification-based deep learning models for date estimation that surpass human performance
- Unsupervised person identification approach applicable to news articles from the *Internet Archive* that reveals individual and joint occurrences of public figures
- Novel benchmark datasets covering different languages and domains for multimodal document verification and retrieval as well as in-depth results of the proposed system
- Unsupervised news analytics system that provides measures of cross-modal context and entity consistency

## 1.4 Thesis Structure

The remainder of this thesis is structured as follows. The mathematical and theoretical foundations to understand the proposed approaches of this thesis are explained in **Chapter 2**. The chapter includes the foundations of the deep learning techniques applied for information extraction from photos. In addition, NLP methods to generate word embeddings as well as for *Named Entity Recognition and Disambiguation* (NER & NED) are explained. Finally, definitions and notations for knowledge graphs are introduced. In **Chapter 3**, several computer vision approaches are suggested for event classification, geolocation estimation, date estimation, and person identification. These approaches allow us to obtain rich image features that are used in conjunction with suitable techniques for NER & NED to quantify the relation between image and text in news articles, as explained in **Chapter 4**. The thesis concludes with a summary and outlines potential areas of future work in **Chapter 5**.

## 1.5 List of Publications

In this section, the publications that have been published in the context of this thesis are listed. Parts of these publications are reused in this thesis.

### Event Classification

- [174] Eric Müller-Budack, Matthias Springstein, Sherzod Hakimov, Kevin Mrutzek, and Ralph Ewerth. “Ontology-driven Event Type Classification in Images”. In: *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*. IEEE, 2021, pp. 2927–2937. DOI: 10.1109/WACV48630.2021.00297

**Abstract:** Event classification can add valuable information for semantic search and the increasingly important topic of fact validation in news. To date only few approaches address image classification for newsworthy event types such as natural disasters, sports events, or elections. Previous work distinguishes only between a limited number of event types and relies on rather small datasets for training. In this paper, we present a novel ontology-driven approach for the classification of event types in images. We leverage a large number of real-world news events to pursue two objectives: First, we create an ontology based on *Wikidata* comprising the majority of event types. Second, we introduce a novel large-scale dataset of images that was obtained by crawling the Web. Several baselines are proposed including an ontology-driven learning approach that aims to exploit structured information of a knowledge graph to learn relevant event relations using deep neural networks. Experimental results on

novel and existing benchmark datasets demonstrate the superiority of the proposed ontology-driven approach.

**Source Code & Dataset:** <https://github.com/TIBHannover/VisE>

### Geolocation Estimation

- [173] Eric Müller-Budack, Kader Pustu-Iren, and Ralph Ewerth. “Geolocation Estimation of Photos Using a Hierarchical Model and Scene Classification”. In: *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*. ed. by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss. Vol. 11216. Lecture Notes in Computer Science. Springer, 2018, pp. 575–592. DOI: 10.1007/978-3-030-01258-8\_35. URL: [https://doi.org/10.1007/978-3-030-01258-8\\_35](https://doi.org/10.1007/978-3-030-01258-8_35)

**Abstract:** While the successful estimation of a photo’s geolocation enables a number of interesting applications, it is also a very challenging task. Due to the complexity of the problem, most existing approaches are restricted to specific areas, imagery, or worldwide landmarks. Only a few proposals predict GPS coordinates without any limitations. In this paper, we introduce several deep learning methods, which pursue the latter approach and treat geolocalization as a classification problem where the Earth is subdivided into geographical cells. We propose to exploit hierarchical knowledge of multiple partitionings and additionally extract and take the photo’s scene content into account, i.e., indoor, natural, or urban setting etc. As a result, contextual information at different spatial resolutions as well as more specific features for various environmental settings are incorporated in the learning process of the convolutional neural network. Experimental results on two benchmarks demonstrate the effectiveness of our approach outperforming the state of the art while using a significant lower number of training images and without relying on retrieval methods that require an appropriate reference dataset.

**Source Code:** <https://github.com/TIBHannover/GeoEstimation>

**Web Demo:** <https://labs.tib.eu/geoestimation>

- [254] Golsa Tahmasebzadeh, Endri Kacupaj, Eric Müller-Budack, Sherzod Hakimov, Jens Lehmann, and Ralph Ewerth. “GeoWINE: Geolocation based Wiki, Image, News and Event Retrieval”. In: *SIGIR ’21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. Ed. by Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai. ACM, 2021, pp. 2565–2569. DOI: 10.1145/3404835.3462786

## 1 Introduction

**Abstract:** In the context of social media, geolocation inference on news or events has become a very important task. In this paper, we present the GeoWINE (Geolocation-based Wiki-Image-News-Event retrieval) demonstrator, an effective modular system for multimodal retrieval which expects only a single image as input. The GeoWINE system consists of five modules in order to retrieve related information from various sources. The first module is a state-of-the-art model for geolocation estimation of images. The second module performs a geospatial-based query for entity retrieval using the Wikidata knowledge graph. The third module exploits four different image embedding representations, which are used to retrieve most similar entities compared to the input image. The last two modules perform news and event retrieval from EventRegistry and the Open Event Knowledge Graph (OEKG). GeoWINE provides an intuitive interface for end-users and is insightful for experts for reconfiguration to individual setups. The GeoWINE achieves promising results in entity label prediction for images on Google Landmarks dataset.

**Web Demo:** <http://cleopatra.ijs.si/geowine/>

- [257] Jonas Theiner, Eric Müller-Budack, and Ralph Ewerth. “Interpretable Semantic Photo to Geolocation”. In: *IEEE Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 4-8, 2022*. IEEE, 2022, pp. 750–760

**Abstract:** Planet-scale photo geolocalization is the complex task of estimating the location depicted in an image solely based on its visual content. Due to the success of convolutional neural networks (CNNs), current approaches achieve super-human performance. However, previous work has exclusively focused on optimizing geolocalization accuracy. Due to the black-box property of deep learning systems, their predictions are difficult to validate for humans. State-of-the-art methods treat the task as a classification problem, where the choice of the classes, that is the partitioning of the world map, is crucial for the performance. In this paper, we present two contributions to improve the interpretability of a geolocalization model: (1) We propose a novel semantic partitioning method which intuitively leads to an improved understanding of the predictions, while achieving state-of-the-art results for geolocalization accuracy on benchmark test sets; (2) We introduce a metric to assess the importance of semantic visual concepts for a certain prediction to provide additional interpretable information, which allows for a large-scale analysis of already trained models. Source code and dataset are publicly available.

**Source Code:** [https://github.com/jtheiner/semantic\\_geo\\_partitioning](https://github.com/jtheiner/semantic_geo_partitioning)

### Date Estimation

- [179] Eric Müller, Matthias Springstein, and Ralph Ewerth. “When Was This Picture Taken?” - Image Date Estimation in the Wild”. In: *Advances in Information Retrieval*

- *39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings*. Ed. by Joemon M. Jose, Claudia Hauff, Ismail Sengör Altingövde, Dawei Song, Dyaa Albakour, Stuart N. K. Watt, and John Tait. Vol. 10193. Lecture Notes in Computer Science. 2017, pp. 619–625. DOI: 10.1007/978-3-319-56608-5\_57. URL: [https://doi.org/10.1007/978-3-319-56608-5\\_57](https://doi.org/10.1007/978-3-319-56608-5_57)

**Abstract:** The problem of automatically estimating the creation date of photos has been addressed rarely in the past. In this paper, we introduce a novel dataset *Date Estimation in the Wild* for the task of predicting the acquisition year of images captured in the period from 1930 to 1999. In contrast to previous work, the dataset is neither restricted to color photography nor to specific visual concepts. The dataset consists of more than one million images crawled from Flickr and contains a large number of different motives. In addition, we propose two baseline approaches for regression and classification, respectively, relying on state-of-the-art deep convolutional neural networks. Experimental results demonstrate that these baselines are already superior to annotations of untrained humans.

**Source Code:** <https://github.com/TIB-Visual-Analytics/DEW-Model>

**Dataset:** <https://doi.org/10.22000/0001abcde>

### Person Identification

- [178] Eric Müller, Christian Otto, and Ralph Ewerth. “Semi-supervised Identification of Rarely Appearing Persons in Video by Correcting Weak Labels”. In: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ICMR 2016, New York, New York, USA, June 6-9, 2016*. Ed. by John R. Kender, John R. Smith, Jiebo Luo, Susanne Boll, and Winston H. Hsu. ACM, 2016, pp. 381–384. DOI: 10.1145/2911996.2912073

**Abstract:** Some recent approaches for character identification in movies and TV broadcasts are realized in a semi-supervised manner by assigning transcripts and/or subtitles to the speakers. However, the labels obtained in this way achieve only an accuracy of 80% – 90% and the number of training examples for the different actors is unevenly distributed. In this paper, we propose a novel approach for person identification in video by correcting and extending the training data with reliable predictions to reduce the number of annotation errors. Furthermore, the intra-class diversity of rarely speaking characters is enhanced. To address the imbalance of training data per person, we suggest two complementary prediction scores. These scores are also used to recognize whether or not a face track belongs to a (supporting) character whose identity does not appear in the transcript etc. Experimental results demonstrate the feasibility of the proposed approach, outperforming the current state of the art.

## 1 Introduction

- [169] Markus Mühling, Nikolaus Korfhage, Eric Müller, Christian Otto, Matthias Springstein, Thomas Langelage, Uli Veith, Ralph Ewerth, and Bernd Freisleben. “Deep learning for content-based video retrieval in film and television production”. In: *Multim. Tools Appl.* 76.21 (2017), pp. 22169–22194. DOI: 10.1007/s11042-017-4962-9

**Abstract:** While digitization has changed the workflow of professional media production, the content-based labeling of image sequences and video footage, necessary for all subsequent stages of film and television production, archival or marketing is typically still performed manually and thus quite time-consuming. In this paper, we present deep learning approaches to support professional media production. In particular, novel algorithms for visual concept detection, similarity search, face detection, face recognition and face clustering are combined in a multimedia tool for effective video inspection and retrieval. The analysis algorithms for concept detection and similarity search are combined in a multi-task learning approach to share network weights, saving almost half of the computation time. Furthermore, a new visual concept lexicon tailored to fast video retrieval for media production and novel visualization components are introduced. Experimental results show the quality of the proposed approaches. For example, concept detection achieves a mean average precision of approximately 90% on the top-100 video shots, and face recognition clearly outperforms the baseline on the public *Movie Trailers Face* Dataset.

- [171] Eric Müller-Budack, Kader Pustu-Iren, Sebastian Diering, and Ralph Ewerth. “Finding Person Relations in Image Data of News Collections in the Internet Archive”. In: *Digital Libraries for Open Knowledge, 22nd International Conference on Theory and Practice of Digital Libraries, TPDL 2018, Porto, Portugal, September 10-13, 2018, Proceedings*. Ed. by Eva Méndez, Fabio Crestani, Cristina Ribeiro, Gabriel David, and João Correia Lopes. Vol. 11057. Lecture Notes in Computer Science. Springer, 2018, pp. 229–240. DOI: 10.1007/978-3-030-00066-0\_20. URL: [https://doi.org/10.1007/978-3-030-00066-0\\_20](https://doi.org/10.1007/978-3-030-00066-0_20)

**Abstract:** The amount of multimedia content in the World Wide Web is rapidly growing and contains valuable information for many applications in different domains. The Internet Archive initiative has gathered billions of time-versioned web pages since the mid-nineties. However, the huge amount of data is rarely labeled with appropriate metadata and automatic approaches are required to enable semantic search. Normally, the textual content of the Internet Archive is used to extract entities and their possible relations across domains such as politics and entertainment, whereas image and video content is usually disregarded. In this paper, we introduce a system for person recognition in image content of web news stored in the Internet Archive. Thus, the system complements entity recognition in text and allows researchers and analysts to track media coverage and relations of persons more precisely. Based on a deep learning face recognition approach, we suggest a system that detects persons of interest and

gathers sample material, which is subsequently used to identify them in the image data of the Internet Archive. We evaluate the performance of the face recognition system on an appropriate standard benchmark dataset and demonstrate the feasibility of the approach with two use cases.

**Source Code:** <https://github.com/TIB-Visual-Analytics/PIIA>

- [172] Eric Müller-Budack, Kader Pustu-Iren, Sebastian Diering, Matthias Springstein, and Ralph Ewerth. “Image Analytics in Web Archives”. In: *The Past Web: Exploring Web Archives*. Ed. by Daniel Gomes, Elena Demidova, Jane Winters, and Thomas Risse. Cham: Springer International Publishing, 2021, pp. 141–151. ISBN: 978-3-030-63291-5. DOI: 10.1007/978-3-030-63291-5\_11. URL: [https://doi.org/10.1007/978-3-030-63291-5\\_11](https://doi.org/10.1007/978-3-030-63291-5_11)

**Abstract:** The multimedia content published on the World Wide Web is constantly growing and contains valuable information in various domains. The Internet Archive initiative has gathered billions of time-versioned web pages since the mid-nineties, but unfortunately, they are rarely provided with appropriate metadata. This lack of structured data limits the exploration of the archives, and automated solutions are required to enable semantic search. While many approaches exploit the textual content of news in the Internet Archive to detect named entities and their relations, visual information is generally disregarded. In this chapter, we present an approach that leverages deep learning techniques for the identification of public personalities in the images of news articles stored in the Internet Archive. In addition, we elaborate on how this approach can be extended to enable detection of other entity types such as locations or events. The approach complements named entity recognition and linking tools for text and allows researchers and analysts to track the media coverage and relations of persons more precisely. We have analysed more than one million images from news articles in the Internet Archive and demonstrated the feasibility of the approach with two use cases in different domains: politics and entertainment.

### Cross-modal Entity Consistency

- [175] Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. “Multimodal Analytics for Real-world News using Measures of Cross-modal Entity Consistency”. In: *Proceedings of the 2020 on International Conference on Multimedia Retrieval, ICMR 2020, Dublin, Ireland, June 8-11, 2020*. Ed. by Cathal Gurrin, Björn Þór Jónsson, Noriko Kando, Klaus Schöffmann, Yi-Ping Phoebe Chen, and Noel E. O’Connor. ACM, 2020, pp. 16–25. DOI: 10.1145/3372278.3390670
- [176] Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, Sherzod Hakimov, and Ralph Ewerth. “Multimodal news analytics using measures of cross-

## 1 Introduction

modal entity and context consistency”. In: *Int. J. Multim. Inf. Retr.* 10.2 (2021), pp. 111–125. DOI: 10.1007/s13735-021-00207-4

**Abstract:** The World Wide Web has become a popular source for gathering information and news. Multimodal information, e.g., enriching text with photos, is typically used to convey the news more effectively or to attract attention. The photos can be decorative, depict additional details, or even contain misleading information. Quantifying the cross-modal consistency of entity representations can assist human assessors in evaluating the overall multimodal message. In some cases such measures might give hints to detect fake news, which is an increasingly important topic in today’s society. In this paper, we present a multimodal approach to quantify the entity relations between image and text in *real-world news*. Named entity linking is applied to extract persons, locations, and events from news texts. Several measures are suggested to calculate the cross-modal similarity of these entities with the news photo, using state-of-the-art computer vision approaches. In contrast to previous work, our system automatically gathers example data from the Web and is applicable to real-world news. The feasibility is demonstrated on two novel datasets that cover different languages, topics, and domains.

### Source Code & Dataset:

[https://github.com/TIBHannover/cross-modal\\_entity\\_consistency](https://github.com/TIBHannover/cross-modal_entity_consistency)

- [68] Ralph Ewerth, Christian Otto, and Eric Müller-Budack. “Computational Approaches for the Interpretation of Image-text Relations”. In: *Empirical Multimodality Research: Methods, Evaluations, Implications*. Ed. by Jana Pflaeging, Janina Wildfeuer, and John A. Bateman. De Gruyter, 2021, pp. 109–138. DOI: 10.1515/9783110725001-005. URL: <https://doi.org/10.1515/9783110725001-005>

**Abstract:** In this paper, we present approaches that automatically estimate semantic relations between textual and (pictorial) visual information. We consider the interpretation of these relations as one of the key elements for empirical research on multimodal information. From a computational perspective, it is difficult to automatically “comprehend” the meaning of multimodal information and to interpret cross-modal semantic relations. One reason is that already the automatic understanding and interpretation of a single source of information (e.g., text, image, or audio) is difficult – and it is even more difficult to model and understand the interplay of two different modalities. While the complex interplay of visual and textual information has been investigated in communication sciences and linguistics for years, they have been rarely considered from a computer science perspective. To this end, we review the few currently existing approaches to automatically recognize semantic cross-modal relations. In previous work, we have suggested to model image-text relations along three main dimensions: cross-modal mutual information, semantic correlation, and the sta-



tus relation. Using these dimensions, we characterized a set of eight image-text classes and showed their relations to existing taxonomies. Moreover, we have shown how the cross-modal mutual information can be further differentiated in order to measure image-text consistency in news at the entity level of persons, locations, and scene context. Experimental results demonstrate the feasibility of the approaches.

- [238] Matthias Springstein, Eric Müller-Budack, and Ralph Ewerth. “QuTII! Quantifying Text-Image Consistency in Multimodal Documents”. In: *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. Ed. by Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai. ACM, 2021, pp. 2575–2579. DOI: 10.1145/3404835.3462796

**Abstract:** The World Wide Web and social media platforms have become popular sources for news and information. Typically, multimodal information, e.g., image and text is used to convey information more effectively and to attract attention. While in most cases image content is decorative or depicts additional information, it has also been leveraged to spread misinformation and rumors in recent years. In this paper, we present a web-based demo application that automatically quantifies the cross-modal relations of entities (persons, locations, and events) in image and text. The applications are manifold. For example, the system can help users to explore multimodal articles more efficiently, or can assist human assessors and fact-checking efforts in the verification of the credibility of news stories, tweets, or other multimodal documents.

**Web Demo:** <https://labs.tib.eu/newsanalytics>

#### Further Publications (Abstracts are omitted)

- [42] Andreas Breitbarth, Eric Müller, Peter Kühmstedt, Gunther Notni, and Joachim Denzler. “Phase unwrapping of fringe images for dynamic 3D measurements without additional pattern projection”. In: *Dimensional Optical Metrology and Inspection for Practical Applications IV*. ed. by Kevin G. Harding and Toru Yoshizawa. Vol. 9489. International Society for Optics and Photonics. SPIE, 2015, pp. 8–17. URL: <https://doi.org/10.1117/12.2176822>
- [69] Ralph Ewerth, Matthias Springstein, Eric Müller, Alexander Balz, Jan Gehlhaar, Tolga Naziyok, Krzysztof Dembczynski, and Eyke Hüllermeier. “Estimating relative depth in single images via rankboost”. In: *2017 IEEE International Conference on Multimedia and Expo, ICME 2017, Hong Kong, China, July 10-14, 2017*. IEEE Computer Society, 2017, pp. 919–924. DOI: 10.1109/ICME.2017.8019434
- [177] Eric Müller-Budack, Jonas Theiner, Robert Rein, and Ralph Ewerth. “Does 4-4-2 exist?” - An Analytics Approach to Understand and Classify Football Team Formations

## 1 Introduction

- in Single Match Situations”. In: *Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports, MMSports@MM 2019, Nice, France, October 25, 2019*. Ed. by Rainer Lienhart, Thomas B. Moeslund, and Hideo Saito. ACM, 2019, pp. 25–33. DOI: 10.1145/3347318.3355527
- [168] David Morris, Eric Müller-Budack, and Ralph Ewerth. “SlideImages: A Dataset for Educational Image Classification”. In: *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II*. ed. by Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins. Vol. 12036. Lecture Notes in Computer Science. Springer, 2020, pp. 289–296. DOI: 10.1007/978-3-030-45442-5\_36. URL: [https://doi.org/10.1007/978-3-030-45442-5\\_36](https://doi.org/10.1007/978-3-030-45442-5_36)
- [253] Golsa Tahmasebzadeh, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth. “A Feature Analysis for Multimodal News Retrieval”. In: *Proceedings of the 1st International Workshop on Cross-lingual Event-centric Open Analytics co-located with the 17th Extended Semantic Web Conference (ESWC 2020), Heraklion, Crete, Greece, June 3, 2020 (online event due to COVID-19 outbreak)*. Ed. by Elena Demidova, Sherzod Hakimov, Jane Winters, and Marko Tadic. Vol. 2611. CEUR Workshop Proceedings. CEUR-WS.org, 2020, pp. 43–56. URL: <http://ceur-ws.org/Vol-2611/paper4.pdf>
- [50] Gullal S. Cheema, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth. “On the Role of Images for Analyzing Claims in Social Media”. In: *Proceedings of the 2nd International Workshop on Cross-lingual Event-centric Open Analytics co-located with the 30th The Web Conference (WWW 2021), Ljubljana, Slovenia, April 12, 2021 (online event due to COVID-19 outbreak)*. Ed. by Elena Demidova, Sherzod Hakimov, Jane Winters, and Marko Tadic. Vol. 2829. CEUR Workshop Proceedings. CEUR-WS.org, 2021, pp. 32–46. URL: <http://ceur-ws.org/Vol-2829/paper3.pdf>
- [49] Gullal S. Cheema, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth. “A Fair and Comprehensive Comparison of Multimodal Tweet Sentiment Analysis Methods”. In: *MMPT@ICMR2021: Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding, Taipei, Taiwan, August 21, 2021*. Ed. by Bei Liu, Jianlong Fu, Shizhe Chen, Qin Jin, Alexander G. Hauptmann, and Yong Rui. ACM, 2021, pp. 37–45. DOI: 10.1145/3463945.3469058
- [195] Kader Pustu-Iren, Eric Müller-Budack, Sherzod Hakimov, and Ralph Ewerth. “Visualizing Copyright-Protected Video Archive Content Through Similarity Search”. In: *Linking Theory and Practice of Digital Libraries - 25th International Conference on Theory and Practice of Digital Libraries, TPDL 2021, Virtual Event, September 13-17, 2021, Proceedings*. Ed. by Gerd Berget, Mark Michael Hall, Daniel Brenn, and Sanna Kumpulainen. Vol. 12866. Lecture Notes in Computer Science. Springer, 2021,

pp. 123–127. DOI: 10.1007/978-3-030-86324-1\_15. URL: [https://doi.org/10.1007/978-3-030-86324-1\\_15](https://doi.org/10.1007/978-3-030-86324-1_15)

- [239] Matthias Springstein, Eric Müller-Budack, and Ralph Ewerth. “Unsupervised Training Data Generation of Handwritten Formulas using Generative Adversarial Networks with Self-Attention”. In: *MMPT@ICMR2021: Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding, Taipei, Taiwan, August 21, 2021*. Ed. by Bei Liu, Jianlong Fu, Shizhe Chen, Qin Jin, Alexander G. Hauptmann, and Yong Rui. ACM, 2021, pp. 46–54. DOI: 10.1145/3463945.3469059

## 1.6 Achievements

In the scope of this thesis, 22 papers (thereof nine as first author) have been published at peer-reviewed conferences [42, 69, 168, 171, 173–175, 178, 179, 257], workshops [49, 50, 177, 239, 253], demo tracks [195, 238, 254], and journals [169, 176], as well as invited chapters [68, 172] in two books.

Five papers were published at conferences that are ranked A\* [173] or A [168, 174, 179, 257] within their domain according to the Australian *Computing Research & Education (CORE) Conference Portal*<sup>3</sup> (source: CORE2021) and two papers [174, 178] were published at the *ACM International Conference on Multimedia Retrieval (ICMR)*. The *ICMR* is one of the top-tier conferences on multimedia as reported by *Google Scholar’s* h5-index metric<sup>4</sup> and “the premier forum of knowledge exchange for researchers and practitioners of multimedia retrieval algorithms, tools, and systems”, according to Candan et al. [45]. Moreover, two web applications were accepted as demos [238, 254] at the *ACM SIGIR Conference on Research and Development in Information Retrieval*, which is an A\* ranked conference for information retrieval according to the Australian *CORE Conference Portal* (source: CORE2021).

**Best Paper Award:** Our paper “Multimodal Analytics for Real-world News using Measures of Cross-modal Entity Consistency” [175] has received the *Best Paper Award* at the *International Conference on Multimedia Retrieval (ICMR) 2020*. Thus, an extended version of the paper was invited for publication in the *International Journal of Multimedia Retrieval (IJMIR)* [176].

**Honorable Mention Award:** Our paper “Finding Person Relations in Image Data of News Collections in the Internet Archive” [171] has received the *Honorable Mention Award* at the *International Conference on Theory and Practice of Digital Libraries (TPDL) 2018*.

<sup>3</sup><http://portal.core.edu.au/conf-ranks/>

<sup>4</sup>[https://scholar.google.com/citations?view\\_op=top\\_venues&hl=de&vq=eng\\_multimedia](https://scholar.google.com/citations?view_op=top_venues&hl=de&vq=eng_multimedia)

## 1 Introduction

**Exhibitions and Media Attention:** The geolocalization approach presented in "Geolocation Estimation of Photos Using a Hierarchical Model and Scene Classification" [173] has attracted attention in the media and was presented in a *c't* article<sup>5</sup> and was also mentioned in a *Computer Bild* article<sup>6</sup>, which are both popular *German* computer magazines. In addition, the demonstrator (details are provided in Section 3.2.5) was or will be presented as an exhibit at the following events and exhibitions:

- *MS Wissenschaft 2019*, 16th May 2019 to 24th October 2019, Exponat 26: "Woher stammt das Bild?"<sup>7</sup>
- *Science Station Tour 2019*<sup>8</sup>, 24th April 2019 to 26th September 2019,
- *Deutsches Museum Bonn* - Exhibition on the topic artificial intelligence 2021<sup>9,10</sup>
- *Zukunftsmobil*, scheduled for 2021

---

<sup>5</sup>Archived link from 19th September 2020 to the article: <https://web.archive.org/web/20200919051036/https://www.heise.de/select/ct/2019/5/1551091142351937>

<sup>6</sup>Archived link from 9th August 2020 to the article: <https://web.archive.org/web/20200809150022/https://www.computerbild.de/artikel/cb-News-Internet-Google-erkennt-uebermenschlich-genau-wo-ein-Foto-aufgenommen-wurde.-15152421.html>

<sup>7</sup>Archived link from 8th December 2020 to the article: <https://web.archive.org/web/20201208142719/https://archiv.ms-wissenschaft.de/2019/ausstellung/rundgang/index.html#accordion-shipplan-collapse-26>

<sup>8</sup>Web link: <https://www.wissenschaft-im-dialog.de/projekte/sciencestation/archiv/>

<sup>9</sup>Web link: <https://www.tib.eu/de/die-tib/neuigkeiten-und-termine/termine/detail/tib-exponat-im-deutschen-museum-bonn-ausgestellt>

<sup>10</sup>Archived link from 29th July 2021: <http://web.archive.org/web/20210729105948/https://www.deutsches-museum.de/bonn/ausstellung/mission-ki>

## 2 Foundations

In this chapter, the theoretical and mathematical foundations to understand the approaches proposed in this thesis are introduced. First, the building blocks of neural networks are presented in Section 2.1. In Section 2.2, convolutional layers and popular *Convolutional Neural Network* (CNN) architectures for image classification are explained that are used in Chapter 3 to train models for information extraction from photos. This information is used to quantify the cross-modal consistency of named entities according to Section 4.2. Methods for *Named Entity Recognition and Disambiguation* (NER & NED) are applied to detect named entities in the text automatically. Furthermore, word embeddings, i.e., vector representations of linguistic items (e.g., a word or sentence), are extracted to verify the contextual consistency between photos and text, as explained in Section 4.3. The related work and foundations of *Natural Language Processing* (NLP) approaches to generate word embeddings as well as for *Named Entity Recognition and Disambiguation* (NER & NED) are presented in Section 2.3. Finally, Section 2.4 introduces relevant concepts for *Semantic Web* and knowledge graphs. These concepts are an important prerequisite for the event classification approach presented in Section 3.1 and the quantification of cross-modal entity consistency in Section 4.2.

### 2.1 Basics of Artificial Neural Networks

Deep learning approaches have achieved impressive performances and are widely applied in many *Computer Vision* (CV) and *Natural Language Processing* (NLP) tasks. In this section, the foundations to understand the perceptron (Section 2.1.1), neural networks (Section 2.1.2) as well as network optimization (Section 2.1.3) are presented.

#### 2.1.1 Perceptron

The perceptron [215], also referred to as neuron, is the smallest element of neural networks and its functionality is inspired by biological processes in the brain of mammals. As illustrated in Figure 2.1, it takes a stimulation  $\mathbf{x} = \langle x_1, x_2, \dots, x_n \rangle \in \mathbb{R}^n$  with dimension  $n$  as input and outputs a value  $\hat{y}$  based on the weights  $\mathbf{w} = \langle w_1, w_2, \dots, w_n \rangle \in \mathbb{R}^n$  that represent

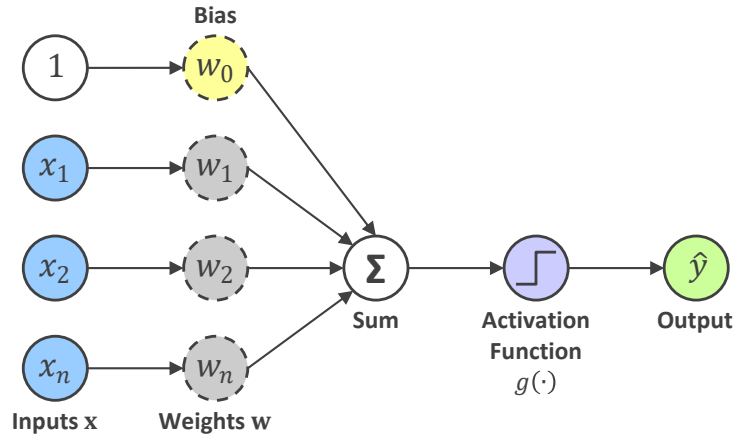


Figure 2.1: Mathematical model of the perceptron. The input stimuli  $\mathbf{x}$  are multiplied with the weights  $\mathbf{w}$ . An activation function  $g(\cdot)$  is applied on the weighted sum, including a bias  $w_0$  to obtain the output  $\hat{y}$ .

the synapses connecting the perceptrons according to the following equation:

$$\hat{y} = g\left(w_0 + \sum_{i=1}^n x_i w_i\right) = g\left(w_0 + \mathbf{x}^\top \mathbf{w}\right) \quad (2.1)$$

The bias  $w_0$  enables to horizontally shift the weighted sum or dot product  $\mathbf{x}^\top \mathbf{w}$  of the inputs  $\mathbf{x}$  and their corresponding weights  $\mathbf{w}$ . Finally, a non-linear activation function  $g(\cdot)$  is applied, allowing neural networks to solve complex non-linear tasks. Perceptrons are typically either activated  $\hat{y} = 1$ , which simulates that a perceptron "fires", or not activated  $\hat{y} = 0$ . This behavior can be reproduced by a step activation function where the bias  $w_0$  defines the activation threshold:

$$\hat{y}_{step} = \begin{cases} 0, & \mathbf{x}^\top \mathbf{w} > w_0, \\ 1, & \text{otherwise.} \end{cases} \quad (2.2)$$

However, the optimization of neural networks, e.g., with the *gradient descent algorithm* [37, 120, 213] and *backpropagation* [217] (explained in Section 2.1.3), requires that each operation in a neural network is differentiable. Since the step function does not fulfill this criterion, it is approximated. Figure 2.2 illustrates three activation functions commonly used in neural networks: *sigmoid*, *hyperbolic tangent*, and *Rectified Linear Unit* (ReLU) [180, 246]. In particular, the ReLU and its variants such as *parametric ReLU* [91] or *leaky ReLU* [288] are widely applied because they are fast to compute due to their mathematical simplicity. Since a single neuron cannot solve complex problems, multiple neurons are combined to form a neural network, as explained in the next section.

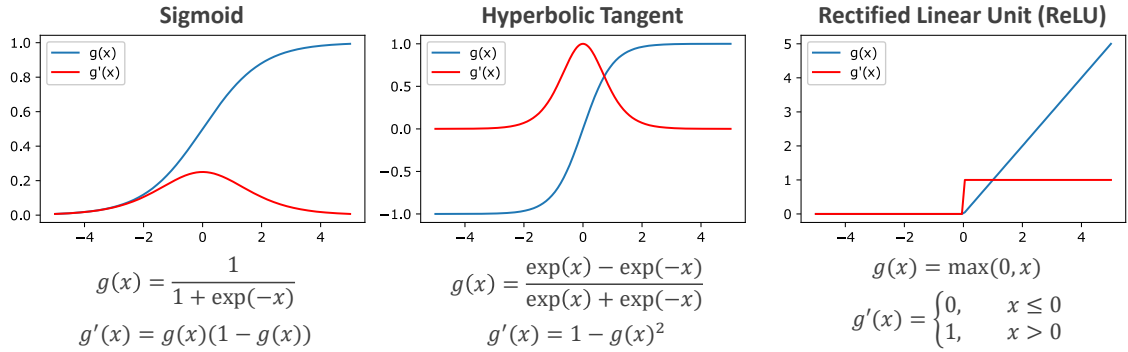


Figure 2.2: Popular activation functions and their first-order derivatives used in neural networks. From left to right: sigmoid, hyperbolic tangent, ReLU.

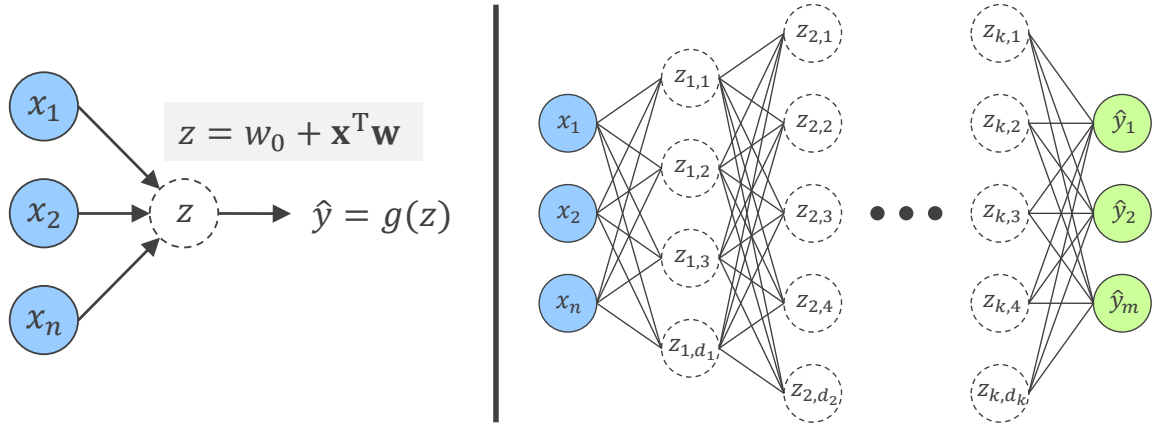


Figure 2.3: **Left:** Simplified illustration of the perceptron. The bias, weight labels, and non-linear activations are omitted in the illustration. It is assumed that each connection is associated with a weight and that an activation function is applied to an output  $z$ . **Right:** *Multilayer Perceptron* (MLP) based on the simplified perceptron illustration with  $n$  input neurons,  $m$  output neurons, and  $k$  hidden layers. The number of neurons in a hidden layer with index  $l \in \{1, 2, \dots, k\}$  corresponds to  $d_l$ .

### 2.1.2 Multilayer Perceptron

The *Multilayer Perceptron* (MLP) is the basis of any type of neural network. As illustrated in Figure 2.3, an MLP consists of a composition of neurons arranged in different layers to model more complex functions and to obtain multidimensional outputs  $\hat{\mathbf{y}} = \langle \hat{y}_1, \hat{y}_2, \dots, \hat{y}_m \rangle \in \mathbb{R}^m$ . The layer(s) between the input and output layer are referred to as hidden layer(s). Unlike the input and output layer, the states of neurons in hidden layers are typically unobserved. The network weights connecting the neurons of the different layers are learned during network optimization. An MLP is a particular type of feedforward neural network in which the information flows solely in one direction from the input to the output layer. Thus, these networks can be considered as a chain  $f(\mathbf{x}) = f_3(f_2(f_1(\mathbf{x})))$  of individual network functions,

## 2 Foundations

where  $f_1(\mathbf{x})$  is the function of the first network layer,  $f_2(f_1(\mathbf{x}))$  the function of the second, etc. [79]. In general, feedforward neural networks can be defined as a function  $\hat{\mathbf{y}} = f(\mathbf{x}; \mathbb{W})$  that outputs a vector  $\hat{\mathbf{y}}$  based on the input vector  $\mathbf{x}$  parametrized by the set of weights  $\mathbb{W} = \{\mathbf{w}^{(0)}, \mathbf{w}^{(1)}, \dots, \mathbf{w}^{(k)}\}$  (including the bias) of the input (layer index  $l = 0$ ) and all  $k$  hidden layers. As explained in Section 2.2, CNNs are another type of feedforward neural network tailored towards processing spatial data such as images. Unlike feedforward neural networks, recurrent neural networks (e.g., *Long short-term memory* (LSTM)), can process sequential data, which are particularly useful for NLP applications (Section 2.3).

### 2.1.3 Network Optimization

In order to optimize a neural network, it is necessary to define a loss or cost function for the given problem, e.g., a classification problem  $y = f^*(\mathbf{x})$  that maps an input vector  $\mathbf{x}$  to a numeric class  $y$ . Given a dataset  $\mathbb{X} = \{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), (\mathbf{x}^{(2)}, \mathbf{y}^{(2)}), \dots, (\mathbf{x}^{(D)}, \mathbf{y}^{(D)})\}$  with  $D$  data points, this loss function  $\mathcal{L}(f(\mathbf{x}; \mathbb{W}), \mathbf{y})$  returns a loss value for the prediction of the network  $f(\mathbf{x}; \mathbb{W})$  for a given input vector  $\mathbf{x}$  and its target vector  $\mathbf{y}$  (or alternatively target value  $y$ ). The goal is to learn a set of network weights  $\mathbb{W}^*$  that results in the best function approximation and consequently achieves the lowest loss:

$$\mathbb{W}^* = \arg \min_{\mathbb{W}} \frac{1}{D} \sum_{i=1}^D \mathcal{L} \left( f \left( \mathbf{x}^{(i)}; \mathbb{W} \right), \mathbf{y}^{(i)} \right) = \arg \min_{\mathbb{W}} J(\mathbb{W}) \quad (2.3)$$

According to Equation (2.3) the optimization function  $J(\mathbb{W})$  can be treated as function of network weights [79]. These weights are first initialized, either at random or with pre-trained weights, i.e., weights learned by a previous network training. At each training step, the loss  $J(\mathbb{W})$  is calculated based on the current weights  $\mathbb{W}$ . A popular approach to optimize the loss applied in many neural networks is the *gradient descent algorithm* [37, 120, 213]. This algorithm updates a weight  $w_i \in \mathbb{W}$  with a specified learning rate  $\eta$  according to the following equation:

$$w_i \leftarrow w_i - \eta \frac{\partial J(\mathbb{W})}{\partial w_i} \quad (2.4)$$

The partial derivative of the weight  $\frac{\partial J(\mathbb{W})}{\partial w_i}$  is calculated with respect to the loss using *back-propagation* [217]. However, neural networks can contain millions of weights [92, 235, 251], and real-world optimization problems can be very complex. Consequently, many local minima exist, and it is not guaranteed that greedy optimization algorithms such as *gradient descent* find a global minimum. Adaptive learning rate algorithms such as *Momentum* [198] or *Adam* [124] have been introduced to mitigate this problem. In addition, the loss is typically calculated on mini-batches with  $B < D$  data points [37] because it is much faster and reliable than using all  $D$  data points (computationally expensive) or a single or very few data points (prone to noise, i.e., data points with inaccurate labels).



## 2.2 Convolutional Neural Networks for Image Classification

Computer vision approaches process images or videos that contain important spatial information. However, the previously MLPs (introduced in Section 2.1.2) process one-dimensional input vectors and flatten (linearize) the image to generate a one-dimensional representation. This linearization has several drawbacks. (1) By using a one-dimensional input vector, the spatial properties of the visual information are neglected. (2) These types of networks are also called fully-connected neural networks because they consist of dense layers in which each neuron is connected to all neurons in the preceding and the following layer, as illustrated in Figure 2.3. As a result, fully-connected neural networks contain many parameters as each connection represents one weight that needs to be learned.

CNNs, which are a particular type of feedforward neural network, provide a solution to counteract these issues. In the remainder of this section, the convolutional layer (Section 2.2.1), related work on CNN architectures for image classification (Section 2.2.2), and widely applied network architectures, namely *AlexNet* (Section 2.2.3), *GoogLeNet* (Section 2.2.4) and *ResNet* (Section 2.2.5), are explained in more detail.

### 2.2.1 Convolutional Layer

In the 1950s to 1960s, Hubel and Wiesel [109] have investigated the visual cortex of mammals (cats, monkeys) and found that neurons respond to the direct environment. CNNs imitate the visual cortex and convolve an  $n \times n$  filter or kernel matrix  $\mathbf{W}$  with an input matrix  $\mathbf{X}$ , as shown in Figure 2.4 (left). Note that padding (e.g., zero-padding) is applied to maintain the spatial resolution of the input. As a result, a single output neuron is connected to an input patch (if  $n > 1$ ), which allows the integration of spatial information. The respective values in the kernel represent the weights  $\mathbf{W}$  that are learned during network training. Moreover, the kernel is spatially shared and outputs a two-dimensional feature map  $\hat{\mathbf{Y}}$  by sliding the filter pixel by pixel over the whole input. This parameter sharing drastically decreases the memory compared to dense layers used in fully-connected neural networks, and it is also highly parallelizable and computationally efficient. As for any neural network, a non-linear activation function such as ReLU is applied on the output feature map to introduce non-linearities (Section 2.1.1).

In reality, the convolution is performed on three-dimensional input tensors  $\mathbf{X}$ , e.g., an RGB image with three channels or the output tensor  $\mathbf{X}_l$  of another (preceding) convolutional layer  $l$  with  $d_l$  channels. Thus, the filter is also a three-dimensional tensor  $\mathbf{W}^{n \times n \times d_l}$  defined by its kernel size  $n$  and the number of channels  $d_l$  of the input tensor  $\mathbf{X}_l$ . Typically, a number of  $d$  filters are learned in a convolutional layer, and *each* filter  $\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_d$  produces a two-dimensional feature map  $\hat{\mathbf{Y}}$ . The result of a convolutional layer is a three-dimensional output tensor  $\hat{\mathbf{Y}}$ . Its output dimension is defined by the spatial resolution (width and height)

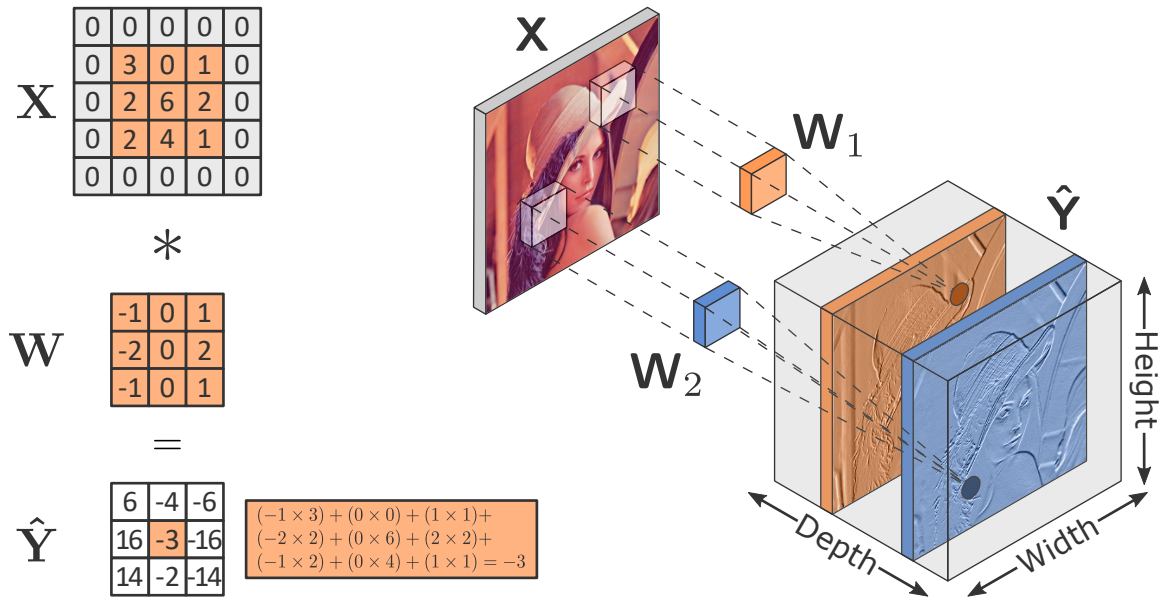


Figure 2.4: **Left:** Example of a two-dimensional convolution of an input matrix  $\mathbf{X}$  with an exemplary convolutional filter  $\mathbf{W}$  (in this case, the weights represent a *Sobel* filter in  $x$ -direction). The output  $\hat{\mathbf{Y}}$  is produced by moving the filter pixel by pixel over the input. The convolution of the orange area in  $\mathbf{X}$  with  $\mathbf{W}$  results in a single output value in  $\hat{\mathbf{Y}}$ . **Right:** Illustration of a convolutional layer. The three-dimensional convolution of each filter  $\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_d$  with the input tensor  $\mathbf{X}$  produces an individual two-dimensional feature map in the output tensor  $\hat{\mathbf{Y}}$ . The dimension of the output tensor is defined by its spatial resolution (width, height) after the convolution and the number of kernels (depth)  $d$  to be learned.

of the feature maps obtained by the convolution and the specified number of filters (depth)  $d$  to be learned, as illustrated in Figure 2.4 (right).

In most state-of-the-art CNN architectures [92, 251, 255], strided convolution or pooling is applied to decrease the spatial resolution of the feature maps, as shown in Figure 2.5. These operations simultaneously increase the receptive field, i.e., the size of the region in the input that produces the feature in the subsequent layers. Consequently, more global and complex filters in later layers of a CNN can be learned [303]. The stride determines the number of pixels by which the convolution filter moves (shifts) each time. Usually, the convolutional filter is applied pixel by pixel (stride = 1) to maintain the input resolution. However, by increasing the stride (stride > 1), the spatial resolution can be decreased. Similarly, a pooling operation decreases the spatial resolution by calculating the minimum (min), maximum (max), or average (avg) of an  $n \times n$  patch (typically  $n = 2$ ) in a feature map. Max pooling (using the maximum value within a patch) is widely applied in many CNN architectures. In general, pooling can help CNNs gain invariance against small translations in the inputs [79].

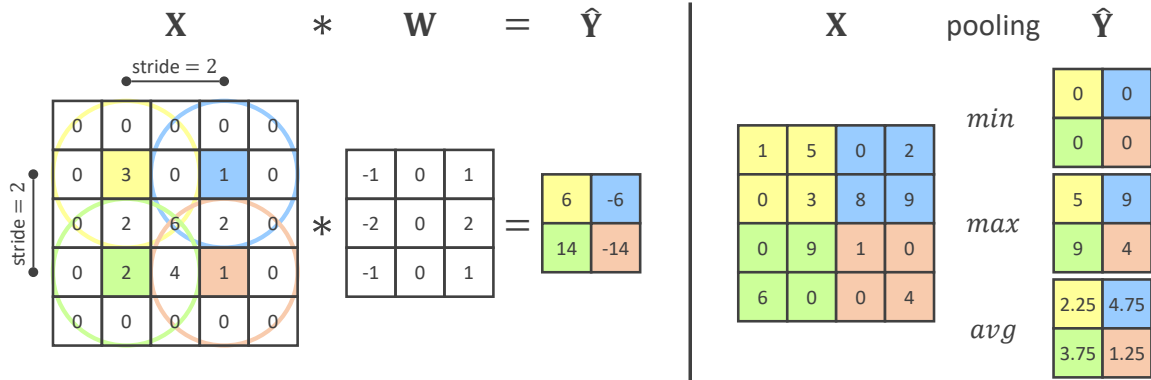


Figure 2.5: **Left:** Example of a convolution with stride = 2, where the filter  $W$  is applied on every 2<sup>nd</sup> pixel (colored) of the input  $X$ . **Right:** Example of min, max, and avg pooling using a kernel with size  $2 \times 2$  and stride = 2. The respective output values  $\hat{Y}$  of a certain input area are highlighted with the same color.

Any neural network that contains at least one convolutional layer is considered a CNN. The architectural design of CNNs, including the type of layers (e.g., convolutional layers or fully-connected layers), the number of layers (network depth), and the number of filters (network width) for solving complex computer vision problems has been extensively researched in the last decade, as discussed in the next section.

### 2.2.2 Overview of Convolutional Neural Network Architectures

The *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) [58] has been widely applied to evaluate computer vision approaches for object classification. In the scope of the ILSVRC 2012, Krizhevsky et al. [126] were the first to apply a CNN, called *AlexNet* (Section 2.2.3), for object classification in images and significantly outperformed previous approaches based on hand-crafted features. In recent years, many improvements regarding the network architectures have been introduced.

Simonyan and Zisserman [235] from the *Visual Geometry Group* (VGG) of the *University of Oxford* suggested using deeper networks and increased the number of convolutional layers from eight (*AlexNet*) to 16 (VGG-16) and 19 (VGG-19) to learn more complex filters and consequently improve the performance for object classification. Szegedy et al. [251, 252] also proposed to leverage deeper (more layers) and wider (more filters) network architectures and introduced the *GoogLeNet* architecture (Section 2.2.4). It uses an *Inception* module that combines outputs from multiple branches of convolutional layers with different filter sizes to extract features at different spatial resolutions. However, very deep neural network architectures suffer from the *Vanishing Gradient Problem* (explained in Section 2.2.5) during optimization. To alleviate this issue, He et al. introduced the *ResNet* architecture [92, 93] that uses residual layers with skip connections to maintain the gradient in the first network

## 2 Foundations

layers. This enables to drastically increase the network depth, e.g., to 152 layers (*ResNet-152*). Szegedy et al. [250] (*Inception-ResNetv2*) and also Xie et al. [286] (*ResNeXt*) exploit the benefits of both residual architectures and inception modules and add the skip connection around an inception module that consists of multiple branches of convolutional filters with different kernel sizes. This approach was extended by Chollet [54] in the *Xception* architecture by applying depthwise separable convolutions, which can be considered another form of the *Inception* module. Unlike standard convolutions that perform both filter operations in one step to combine inputs into a new set of outputs, depthwise separable convolutions split this operation into two layers. Spatial correlations are extracted using a spatial convolution performed over each input channel independently, followed by a point-wise  $1 \times 1$  convolution to extract cross-channel correlations. This factorization significantly reduces the computational time and model size. Based on this observation, Howard et al. [100] and Sandler et al. [222] proposed the *MobileNet* architectures that aim to reduce the network complexity while maintaining competitive results compared to the state of the art.

Based on the idea of improving the information flow of CNNs with skip connections, Huang et al. [106] proposed a *Densely Connected Convolutional Network (DenseNet)*. Instead of a single skip connection between a layer and its subsequent layer in residual architectures [92, 93], the *DenseNet* uses skip connections to all subsequent layers within a network block in the architecture to reduce the *Vanishing Gradient Problem* further and increase information propagation throughout the network. Unlike previous solutions that primarily aim to improve the spatial encodings throughout the network, Hu et al. [103] proposed the *Squeeze-and-Excitation (SE)* block to model the interdependencies between feature channels within a neural network and consequently use this global information to emphasize informative features across all channels and simultaneously suppress less useful ones.

CNN architectures have become very complex, and the definition of the optimal hyperparameters, layer types as well as their arrangement is a tedious task for computer scientists. For this reason, Zoph and Le [321] and Zoph et al. [322] suggest a *Neural Architecture Search (NAS)* to automatically find the optimal network architecture using a reinforcement learning approach. Howard et al. [101] applied a NAS to find *MobileNetv3* that, like its predecessors [100, 222], aims to solve computer vision tasks with less complex network architectures. Tan and Le [255] applied a NAS and further studied network scaling to balance the network depth (number of layers), width (number of filters in a layer), and resolution (spatial resolution of input images) in order to automatically find suitable architectures with varying complexity ranging from less complex models for mobile use to very complex but powerful architectures. Their *EfficientNet* architectures have achieved state-of-the-art results among networks with varying complexity on the ILSVRC 2012 dataset with 1,000 classes in 2019. Xie et al. [285] have presented a semi-supervised deep learning approach and first trained an *EfficientNet* as a *teacher network* on labeled data that subsequently generates noisy pseudo labels for 300 million unlabeled images. Then, they have trained a

## 2.2 Convolutional Neural Networks for Image Classification

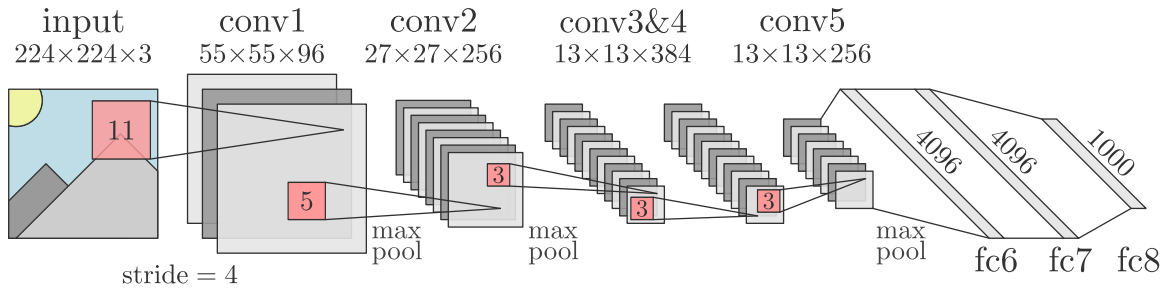


Figure 2.6: Illustration of the *AlexNet* architecture [126] with five convolutional layers and three fully-connected (fc) or also called dense layers. The number of channels in a convolution layer (conv) is defined by the number of specified filters from the previous layer (Figure 2.4). Kernel sizes (quadratic) are displayed in the red squares. A strided convolution with stride = 4 as well as max pooling with a  $3 \times 3$  kernel and stride = 2 is applied for dimension reduction. The number of neurons in the last fully-connected layer (fc8) corresponds to the number of 1,000 classes in the ILSVRC 2012 dataset [58, 218].

*student network* using a more complex *EfficientNet* architecture on both labeled and pseudo labeled images. They have iterated the process by setting students as new teachers and demonstrated that this strategy can improve generalization and performance for image classification. Inspired by the success of transformer models in NLP [61, 203, 204] (Section 2.3), Dosovitskiy et al. [64] recently proposed a visual transformer model. They split the image into a sequence of patches and use a trainable linear projection (implemented via a single convolutional layer) to create embeddings for each of them. Furthermore, they add positional embeddings to incorporate spatial information and feed the resulting sequence into a transformer [264]. Experimental results have shown that this alternative approach matches or outperforms *ResNet*-like CNN architectures [92, 101, 255, 285, 322] with comparable complexity on many image classification datasets including ILSVRC 2012 dataset [58, 218] and poses an interesting research direction.

### 2.2.3 AlexNet Architecture

The *AlexNet* architecture introduced by Krizhevsky et al. [126] was the first CNN applied on the ILSVRC 2012 dataset [58, 218] for object classification and won the challenge by outperforming traditional approaches based on hand-crafted features. An overview of the architecture is illustrated in Figure 2.6. It can be divided into two parts: feature extraction and classification.

The first part of the network comprises five convolutional layers with varying kernel sizes and number of features. It aims to learn a feature representation of the RGB input image with dimension  $224 \times 224 \times 3$ . After each convolutional layer, a ReLU is used as a non-linear activation function. Strided convolution (stride = 4) and max pooling are applied to decrease

## 2 Foundations

the spatial dimension of the feature maps, according to Figure 2.6. The classification part of the network is an MLP (similar to Figure 2.3) with three fully-connected layers. They require a one-dimensional input vector produced by linearizing the feature representations obtained by the first part of the network. The number of neurons in the last fully-connected layer matches the number of  $m$  classes ( $m = 1,000$  for the ILSVRC 2012 dataset [58, 218]) that are to be distinguished. In contrast to the first two fully-connected layers, which use ReLU as activation function, *softmax* is applied on the outputs of the last fully-connected layer to produce the final class probabilities  $\hat{\mathbf{y}} = \langle \hat{y}_1, \hat{y}_2, \dots, \hat{y}_m \rangle$ . The *softmax* maps an input vector  $\mathbf{x} \in \mathbb{R}^m$  to another vector  $\hat{\mathbf{y}}$  with the same dimension according to:

$$\hat{y}_i = \text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^m \exp(x_j)} \quad \forall i \in \{1, 2, \dots, m\} \quad (2.5)$$

The *softmax* represents a discrete probability distribution over  $m$  values, as each class probability ranges from  $0 \leq \hat{y}_i \leq 1$  and the sum of all probabilities equals  $\sum_{i=1}^m \hat{y}_i = 1$ ,

As discussed in Section 2.2.2, first attempts such as the VGG networks [235] mainly increased the number of network layers and filters compared to the *AlexNet* to improve the network capabilities. However, this has several drawbacks. (1) Larger networks typically contain more parameters, making them more prone to overfitting, particularly if the amount of training data is limited. (2) The network size in terms of memory requirements drastically increases. (3) Very deep CNN with many convolutional layers suffer from the *Vanishing Gradient Problem*. In the following, the *GoogLeNet* [251, 252] and *ResNet* [92, 93] architectures are presented that aim to address these problems.

### 2.2.4 GoogLeNet Architecture

**Inception Module:** Szegedy et al. [251] introduced the *Inception* module, which learns features at different kernel sizes to improve the network capabilities while drastically reducing the number of parameters and consequently memory requirements of the neural network. Figure 2.7 shows a *naïve Inception* module and their proposed implementation of the *Inception* module. The kernel sizes in the network layers can significantly impact the performance, and finding the optimal values is challenging. The *Inception* module allows for applying filters with different kernel sizes within one layer by stacking their individual outputs in the output tensor. Each module learns a number of  $1 \times 1$ ,  $3 \times 3$ , and  $5 \times 5$  filters. Besides, it also applies a parallel *max pooling* path, which does not require any parameters but can increase translation invariance [79] (Section 2.2.1). While *max pooling* is usually used for dimension reduction (stride = 2) in CNNs, stride = 1 is applied in the *Inception* module to maintain the spatial resolution.

However, a naïve implementation of this approach is computationally-intensive, as the example in Figure 2.7 reveals. Let the input of the network layer  $l$  be a tensor  $\mathbf{X}_l$  with

## 2.2 Convolutional Neural Networks for Image Classification

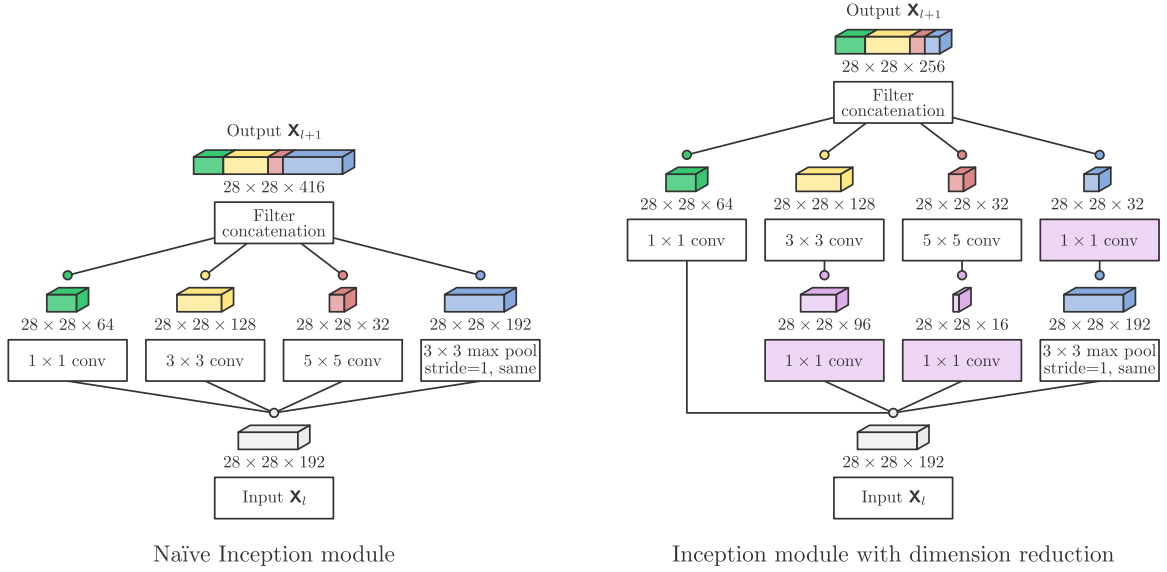


Figure 2.7: Exemplary *naïve Inception* module (left) and *Inception* module (in this case *Inception (3a)* from Table 2.1) with dimension reduction (right) proposed by Szegedy et al. [251]. Convolutional filters with kernel size  $1 \times 1$  (purple) are used to decrease the number of features before the "heavier" convolutional operations with larger kernel sizes are computed. Same padding and stride = 1 is used for the max pooling operation to maintain the spatial resolution.

dimension  $28 \times 28 \times 192$ . The computational cost without dimension reduction to calculate 32 feature maps (with the same spatial resolution) using a  $5 \times 5$  kernel can be approximated as follows. For each of the  $28 \times 28 \times 32$  values in the output  $\mathbf{X}_{l+1}$ ,  $5 \times 5 \times 192$  multiplications are necessary. This equals to a large number of approximately 120 million multiplications.

As a solution, Szegedy et al. [251] propose to apply  $1 \times 1$  convolution filters in order to decrease the number of channels of the input tensor  $\mathbf{X}_l$  before the "heavier" convolutional operations are computed. First, the number of channels is decreased by learning a smaller number of  $1 \times 1$  convolutional filters, in this case 16. These filters only require to perform  $1 \times 1 \times 192$  multiplications for each of the  $28 \times 28 \times 16$  output values, which totals in approximately 2.4 million multiplications. Following the previous example, the computational cost of  $5 \times 5$  convolutional filters is drastically decreased because only  $5 \times 5 \times 16$  multiplications are required to produce each of the  $28 \times 28 \times 32$  output values. As a result, the number of multiplications is decreased by a factor of around 10, from 120 million to 12.4 million (2.4 million for  $1 \times 1$  and 10 million for  $5 \times 5$  convolutions). The number of features from the parallel *max pooling* branch is equal to the number of input channels and is comparatively high. To address this issue, another  $1 \times 1$  convolution is applied in order to reduce the number of *max pooling* features in the output tensor. Overall, the *Inception* module allows for building more complex models as the memory requirement can be drastically decreased.

## 2 Foundations

Table 2.1: Details of the *GoogLeNet* architecture from Szegedy et al. [251]. All convolutions (also inside the *Inception* modules) use ReLU as non-linear activation function. The input is an RGB image tensor with dimension  $224 \times 224 \times 3$ . The number of  $1 \times 1$  filters used for reducing the dimension (illustrated in Figure 2.7) before the  $3 \times 3$  and  $5 \times 5$  convolution is denoted as " $3 \times 3$  reduce" and " $5 \times 5$  reduce". The number of  $1 \times 1$  filters to reduce the feature dimension of the max-pooling output can be found in the column "pool proj."

type	kernel size (stride)	output size	$1 \times 1$	$3 \times 3$ reduce	$3 \times 3$	$5 \times 5$ reduce	$5 \times 5$	pool proj.
convolution	$7 \times 7$ (2)	$112 \times 112 \times 64$						
max pool	$3 \times 3$ (2)	$56 \times 56 \times 64$						
convolution	$3 \times 3$ (1)	$56 \times 56 \times 192$		64	192			
max pool	$3 \times 3$ (2)	$28 \times 28 \times 192$						
inception (3a)		$28 \times 28 \times 256$	64	96	128	16	32	32
inception (3b)		$28 \times 28 \times 480$	128	128	192	32	96	64
max pool	$3 \times 3$ (2)	$14 \times 14 \times 480$						
inception (4a)		$14 \times 14 \times 512$	192	96	208	16	48	64
inception (4b)		$14 \times 14 \times 512$	160	112	224	24	64	64
inception (4c)		$14 \times 14 \times 512$	128	128	256	24	64	64
inception (4d)		$14 \times 14 \times 528$	112	144	288	32	64	64
inception (4e)		$14 \times 14 \times 832$	256	160	320	32	128	128
max pool	$3 \times 3$ (2)	$7 \times 7 \times 832$						
inception (5a)		$7 \times 7 \times 832$	256	160	320	32	128	128
inception (5b)		$7 \times 7 \times 1024$	384	192	384	48	128	128
avg pool	$7 \times 7$ (1)	$1 \times 1 \times 1024$						
dropout (40%)		$1 \times 1 \times 1024$						
fully-connected		$1 \times 1 \times 1000$						
softmax		$1 \times 1 \times 1000$						

**GoogLeNet:** Table 2.1 contains the details of the *GoogLeNet* architecture suggested by Szegedy et al. [251]. First, a strided convolution (stride = 2) and *max pooling* are applied to decrease the spatial resolution of the input and, consequently, memory requirements. The remainder of the architecture uses the proposed *Inception* modules. *Max pooling* is used between blocks of these modules to further decrease the spatial resolutions in the later network stages. Finally, *average pooling* is applied to generate a one-dimensional feature representation that is subsequently used for classification. In order to prevent overfitting, the activations of 40% of the neurons (randomly selected) are set to zero. This technique is referred to as *Dropout* [240]. The number of neurons in the final dense layer matches the number of  $m$  classes ( $m = 1,000$  for the ILSVRC 2012 dataset [58, 218]), and the *softmax* (Equation (2.5)) is calculated to produce the final class probabilities as explained in Section 2.2.3.



### 2.2.5 ResNet Architecture

**Vanishing Gradient Problem:** As mentioned in Section 2.2.2, especially very deep CNNs with a large number of hidden layers suffer from the *Vanishing Gradient Problem*. The origin of this problem lies in the backpropagation algorithm [217], which is used to calculate the weight gradients for optimization (Section 2.1.3). Referring to Figure 2.3, the outputs of a layer  $l$  are used as input of layer  $l + 1$ . Therefore, any neural network with an input vector  $\mathbf{x}$  (or tensor  $\mathbf{X}$ ) can be considered as a chain  $f(\mathbf{x}) = f_3(f_2(f_1(\mathbf{x})))$  of individual functions, where  $f_1(\mathbf{x})$  is the function of the first layer,  $f_2(f_1(\mathbf{x}))$  the function of the second, etc. [79]. Therefore, the chain rule needs to be applied to backpropagate the loss from the last to the initial network layer. These gradients can have small values, particularly if activation functions such as the *sigmoid* or *tanh* are applied (Figure 2.2), that are multiplied together. As a result, the gradient exponentially decreases while backpropagating to the initial layers causing gradients that are almost zero. This problem is referred to as *Vanishing Gradient*. It complicates the optimization of the weights in early network layers leading to an overall inaccuracy of the whole network. It is worth noting that conversely, the possibility of an *Exploding Gradient* (high gradients are multiplied together) exists.

**Residual Layers:** He et al. [92] introduced residual blocks to overcome this problem. They suggest an identity mapping using a shortcut between the input tensor  $\mathbf{X}_l$  of a residual block and its output  $\mathbf{X}_{l+1}$ . As illustrated in Figure 2.8, the input is directly added to a residual mapping  $f(\mathbf{X}_l)$  that is to be learned:

$$\mathbf{X}_{l+1} = f(\mathbf{X}_l) + \mathbf{X}_l \quad (2.6)$$

The residual mapping makes it easier to propagate information through the network. In addition, the derivative  $\frac{\partial \mathbf{X}_{l+1}}{\partial f(\mathbf{X}_l)} + \frac{\partial \mathbf{X}_{l+1}}{\partial \mathbf{X}_l}$  of the block is much larger since the partial derivation of the shortcut corresponds to  $\frac{\partial \mathbf{X}_{l+1}}{\partial \mathbf{X}_l} = 1$ , which alleviates the *Vanishing Gradient Problem*.

Figure 2.8 shows different variants of the residual block proposed by He et al. [92, 93]. They investigated the arrangement of the operations in a residual block to identify the optimal gradient flow and found that pre-activations with batch normalization (*Residual block v2* in Figure 2.8) generally provide the best results. Batch normalization [110] produces normalized outputs by subtracting the mean and by dividing the standard deviation of the outputs within a batch. It adds two trainable parameters to each layer to scale and shift the normalized outputs for optimization. Batch normalization increases the stability of the network training and further alleviates the problem of *Vanishing Gradients*. For deeper CNNs, a residual block with bottleneck is introduced. It follows the general idea of the *Inception* module [251] (Section 2.2.4) and uses a  $1 \times 1$  convolutional layer to decrease the feature dimension of the input and the number of computations.

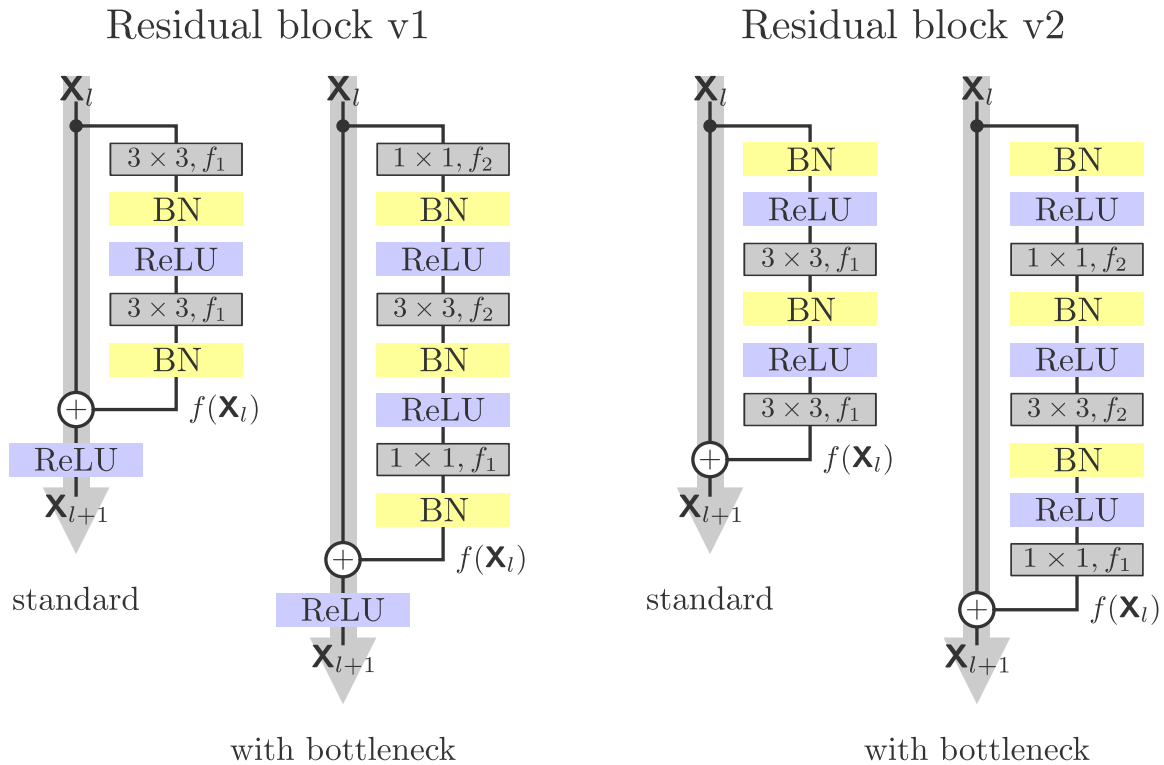


Figure 2.8: Structure of the first [92] and second variant [93] of the residual block. The second (improved) variant applies batch normalization (BN) and ReLU before the convolution and generally achieves better results. A residual block with bottleneck applies a  $1 \times 1$  convolution to decrease the number of features  $f_1$  (tensor depth) of  $\mathbf{X}_l$  to  $f_2 < f_1$  similar to the *Inception* module [251] (Section 2.2.4). This reduces the number of computations in the convolutional layer with a  $3 \times 3$  kernel. Finally, another  $1 \times 1$  convolution with  $f_1$  filters is applied to match the original feature dimension of the input tensor  $\mathbf{X}_l$ .

**ResNet:** The residual blocks allow for building deeper networks for image classification. He et al. [92, 93] investigated *ResNet* architectures with varying network depth (number of layers). The architecture details are provided in Table 2.2. All proposed variants first apply a strided convolution (stride = 2) and *max pooling* to decrease the spatial resolution. Afterward, the variants use a different amount of residual blocks within each network block. Residual blocks with bottlenecks are used in deeper network architectures with 50 or more layers to increase computational efficiency. The first convolutional layer of each network block performs a strided convolution (stride = 2) to decrease the spatial resolution of the features. This dimension reduction enables the network to learn more filters (less memory required due to the smaller spatial resolution) at a larger receptive field. *Average pooling* is applied to generate a feature vector. Finally, a fully-connected layer with *softmax* activation outputs the probabilities of  $m$  classes ( $m = 1,000$  for the ILSVRC 2012 dataset [58, 218]).

Table 2.2: Details of *ResNet* architectures [92] with varying depth, i.e., different number of residual blocks. The input size is  $224 \times 224 \times 3$  pixels for RGB images. Residual blocks with bottleneck (Figure 2.8) are used and denoted as follows: [kernel size  $\times$  kernel size, number of kernels]  $\times$  number of blocks stacked. A convolution with stride = 2 is performed by conv3\_1, conv4\_1, and conv5\_1 to reduce the spatial resolution. All convolutions use ReLU as non-linear activation function.

layer name	output size	ResNet-50	ResNet-101	ResNet-152
conv1	$112 \times 112 \times 64$	7 $\times$ 7 conv, 64 filters, stride 2		
max pool	$56 \times 56 \times 64$	7 $\times$ 7 max pooling, stride 2		
conv2_x	$56 \times 56 \times 256$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
conv3_x	$28 \times 28 \times 512$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 8$
conv4_x	$14 \times 14 \times 1024$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 23$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 36$
conv5_x	$7 \times 7 \times 2048$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
avg pool	2048	7 $\times$ 7 avg pool, stride=1		
fc	1000	fully-connected layer with 1000 neurons		
softmax	1000	softmax		

## 2.3 Natural Language Processing

The extraction of textual features in the form of word embeddings and named entities such as persons, organizations, and locations is an important preprocessing step to quantify relations between photos and text. This section introduces related work and the basic concepts of NLP methods for distributional semantics (Section 2.3.1), *Named Entity Recognition* (Section 2.3.2), and *Named Entity Disambiguation* (Section 2.3.3).

### 2.3.1 Distributional Semantics

Distributional semantics aim to map a linguistic item (e.g., a word or sentence) to a vector representation, i.e., word embeddings, using their semantic similarities and distributional properties in a large corpus of language data. Word embeddings are an essential prerequisite for many NLP tasks, e.g., *Named Entity Recognition* (NER) and question answering, and have been an active research topic for decades. *Word2Vec* [164], *Global Vectors for Word*

## 2 Foundations

*Representation* (GloVe) [191], and *fastText* [35] rely on the *distributional hypothesis* that words used in the same context tend to share a similar meaning. These approaches are trained in an unsupervised manner on large-scale corpora to learn from this *distributional hypothesis*. However, each word is associated with a single global word embedding without considering the textual context. For this reason, research has focused on contextual models that aim to consider an entire input sequence to determine a representation of a word or token. Thus, the same word or token can have different representations based on the context. First approaches such as *Embeddings from Language Models* (ELMo) [192, 193] and *Universal Language Model Fine-tuning* (ULMFiT) [102] have used LSTM models to process input sequences, while more recent approaches like *Bidirectional Encoder Representations from Transformers* (BERT) [61], *Generative Pre-training* (GPT) [203, 204], *XLNet* [293], and other BERT variants [130, 149, 223, 248] instead rely on transformer models. Unlike LSTM models [102, 192, 193] that process the inputs sequentially, transformers [61, 130, 149, 203, 204, 223, 248, 293] compute contextual embeddings for the input sequence in parallel and model connections between words using an attention mechanism. This parallel processing drastically reduces the computational time, and the use of skip connections [92, 93] (Section 2.2.5) in transformer models can alleviate the *Vanishing Gradient Problem*.

*Word2Vec* from Mikolov et al. [164, 165] is one of the most fundamental approaches in distributional semantics. The details are presented in Section 2.3.1.1. Furthermore, the *fastText* [35] algorithm, which can be considered as an extension of *Word2Vec*, is described in Section 2.3.1.2 as it is used to extract word embeddings from the text in Section 4.3.

### 2.3.1.1 Word2Vec

*Word2Vec* [164, 165] is a neural network that learns unique vector representations for each word in a large text corpus given for training. Based on the *distributional hypothesis*, the goal is to learn similar word embeddings for words used in the same context as they tend to share a similar meaning. Mikolov et al. [165] propose two different model architectures, namely a *Continuous Bag-of-Words* (CBOW) *Model* and a *Continuous Skip-gram Model* for learning distributed representations of words, as shown in Figure 2.9.

Unlike deep neural networks (Section 2.1.2), the proposed *Word2Vec* architectures are relatively simple and comprise only a single hidden layer. While the CBOW model predicts the current word (target) using the surrounding words as context, the *Continuous Skip-gram Model* conversely estimates the surrounding words within a certain range given the current word. The algorithmic details of both models are very similar. However, word embeddings from the *Continuous Skip-gram Model* tend to work better because each context-target pair is statistically a new observation. Therefore, the remainder of this section focuses on the *Continuous Skip-gram Model*.



## 2 Foundations

Morin and Bengio [166] to decrease the number of output nodes to evaluate from  $n_w$  to  $\log_2 n_w$ . During inference, the word embedding of a word is generated by multiplying the respective one-hot encoded word vector  $\mathbf{x}_i$  with the input weight matrix  $\mathbf{W}_0^{n_w \times d}$ .

Mikolov et al. [164] identified that most of the complexity is caused by the non-linear hidden layer in the model. They proposed a much simpler log-linear model that omits the hidden layer to decrease the computational complexity. However, some extensions of the *Word2Vec* approach, such as *fastText* [35], rely on the neural network variant with a hidden layer, as explained in the next section.

### 2.3.1.2 FastText

*Word2Vec* assigns a distinct word embedding to each word in the vocabulary. For this reason, it cannot handle so-called *out-of-vocabulary words* that it has not encountered during the training process. For example, there might be no valid embedding of the word *fasttext*, even if the individual parts *fast* and *text* exist in the vocabulary. Furthermore, parameters for words with the same radicals, such as *write* and *writing*, are not shared.

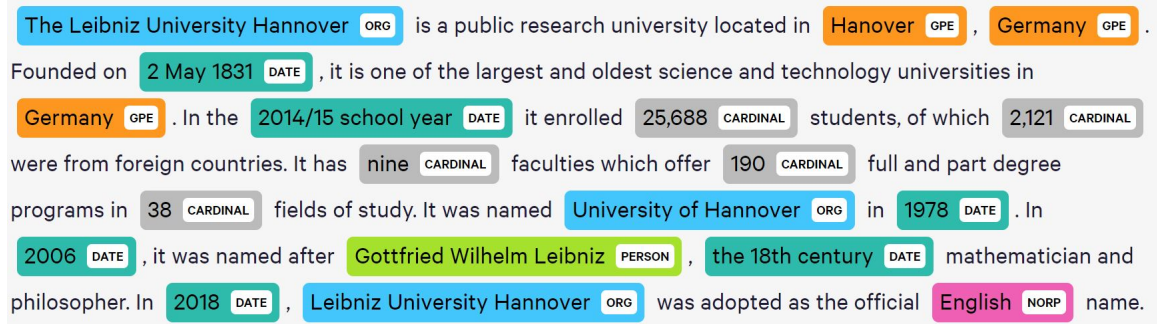
As part of the *fastText* framework, Bojanowski et al. [35] derive a skip-gram model from *Word2Vec* [165] in which each word is represented as a bag of character  $n$ -grams. More specifically, they add boundary symbols  $\langle$  and  $\rangle$  to indicate the beginning and end of a word, which allows distinguishing prefixes and suffixes from other character sequences. Furthermore, they create a set of  $n$ -grams for a given word, containing the word itself and all character  $n$ -grams. In practice, the *fastText* [35] algorithm uses all  $n$ -grams for  $3 \leq n \leq 6$ . Given the word *fasttext*, the bag of  $n$ -grams contains:

$n = 3$ :  $\langle \text{fa, fas, ast, stt, tte, tex, ext xt} \rangle$   
 $n = 4$ :  $\langle \text{fas, fast, astt, stte, ttex, text, ext} \rangle$   
 $n = 5$ :  $\langle \text{fast, fastt, astte, sttex, ttext, text} \rangle$   
 $n = 6$ :  $\langle \text{fastt, fastte, asttex, sttext, ttext} \rangle$

and the word itself:

$\langle \text{fasttext} \rangle$

The vocabulary size is defined by the number of different  $n$ -grams extracted from words in a large text corpus. A multi-hot encoded vector  $\mathbf{x}$  is created for each word, indicating the  $n$ -grams appearing in the word. In contrast to *Word2Vec* [164, 165], which uses a one-hot encoded vector for each word, this allows the integration of subword information and drastically decreases the number of out-of-vocabulary words. The word embedding for the whole word is defined as the sum of individual vector representations of its  $n$ -gram. The feature dimension of the vector representation used in *fastText* is set to  $d = 300$ . However, the number of  $n$ -grams and consequently the vocabulary can be very large. To decrease the memory



The Leibniz University Hannover ORG is a public research university located in Hanover GPE, Germany GPE. Founded on 2 May 1831 DATE, it is one of the largest and oldest science and technology universities in Germany GPE. In the 2014/15 school year DATE it enrolled 25,688 CARDINAL students, of which 2,121 CARDINAL were from foreign countries. It has nine CARDINAL faculties which offer 190 CARDINAL full and part degree programs in 38 CARDINAL fields of study. It was named University of Hannover ORG in 1978 DATE. In 2006 DATE, it was named after Gottfried Wilhelm Leibniz PERSON, the 18th century DATE mathematician and philosopher. In 2018 DATE, Leibniz University Hannover ORG was adopted as the official English NORP name.

Figure 2.10: Exemplary output of *spaCy* [99] for *Named Entity Recognition* (NER). Named entities are detected in an unstructured text and classified into pre-defined categories such as organizations/institutions (ORG), countries/cities/states (GPE), nationalities or religious or political groups (NORPS), dates (DATE), cardinals (CARDINAL), persons (PERSON), etc. The screenshot is taken from: <https://explosion.ai/demos/displacy-ent>

requirements Bojanowski et al. [35] also propose to use a hashing function (*Fowler-Noll-Vo* (*FNV*)-1a hashing<sup>11</sup>) that maps the  $n$ -grams to a number of integers (bucket size) that is smaller than the vocabulary size. The remaining model and training details correspond to the *Word2Vec* approach in Section 2.3.1.1.

### 2.3.2 Named Entity Recognition

Given an input document  $D$  with unstructured textual information, approaches on NER aim to detect a set of named entities  $\mathbb{E}$  in it and to subsequently classify them into pre-defined categories such as persons, locations, events, or organizations. The start and end characters (span) of each mention in a document are extracted and stored with a corresponding type label. An example is shown in Figure 2.10. In information retrieval, named entities are defined as physical or abstract real-world objects designated by a proper name consisting of a continuous span of tokens without nesting. For example, "Leibniz University Hannover" is considered as a single named entity, although "Hannover" is itself a name within this span. Moreover, approaches for NER usually also consider temporal (e.g., dates, weekdays, etc.) and numerical expressions (e.g., percentages, amounts of money, etc.) as named entities.

In recent years, several tools for NER such as *Stanford Core NLP* [73, 152], *Illinois NLP* [209, 298], *Dandelion* [39], *spaCy* [99], *FLAIR* [9–11], and *Stanza* [196] have been introduced. Traditional NER systems [73, 209] have used hand-crafted features (e.g., lower and upper case, word orders, etc.) extracted from text. Based on these features, machine learning algorithms such as *Hidden Markov Models* (HMMs) [65], *Decision Trees* [201], *Support Vector Machines* (SVMs) [95], and *Conditional Random Fields* (CRFs) [128] have been trained for NER. Lample et al. [129] were among the first to replace hand-crafted

<sup>11</sup><http://www.isthe.com/chongo/tech/comp/fnv/>

features with automatically generated features from a deep learning approach. More specifically, they applied a bidirectional LSTM [84] to generate text representations, which are subsequently used to train a CRF [128]. Similar to solutions in the field of distributional semantics (Section 2.3.1), state-of-the-art systems such as *FLAIR* [9–11] use pre-trained language models (e.g., ELMo [192, 193]) to include contextualized string embeddings for a more robust NER. Qi et al. [196] introduced a complete neural pipeline system called *Stanza* as an extension for *Stanford Core NLP* [73, 152]. Unlike most previous approaches, *Stanza* [196] can process raw input texts as solves all related tasks (e.g., tokenization, lemmatization, part-of-speech tagging) and NER within the same toolkit. A more comprehensive survey that focuses on deep learning approaches for NER is provided by Li et al. [135].

Tools from industry, such as *spaCy* [99], are typically updated more frequently and are used for NER in this thesis. The latest version 3 of *spaCy* uses a deep learning approach based on word representations from *RoBERTa* [149], a recent transformer model for distributional semantics<sup>12</sup>. Furthermore, they provide off-the-shelf solutions for NER including all related tasks (e.g., tokenization, lemmatization, part-of-speech tagging, etc.), making them able to process raw input texts.

### 2.3.3 Named Entity Disambiguation

For many applications, including the quantification of cross-modal entity relations, the text spans of named entities extracted from the text are not sufficient for the following reasons. Due to name variations, a specific entity extracted from the text can be denoted by several mentions (e.g., *U.S. President*, *President Obama*, *Barack Obama*). Conversely, a single mention can represent a candidate for multiple distinct named entities in a knowledge base because of name ambiguities. For example, "*Hanover*" can refer to the capital city of the German federated state of Lower Saxony but also many other cities, e.g., in the U.S. states of New Hampshire, Pennsylvania, Indiana, Kansas, and Minnesota or the short form "*Tesla*" can refer to the organization "*Tesla, Inc.*" or the person "*Nicola Tesla*". In order to extract valuable and complete information from the text, it is necessary to link each mention of a named entity recognized in the text unambiguously to an actual entity in a knowledge base such as *Wikipedia*, *Wikidata* [268], *DBpedia* [22], or *YAGO* [241] by assigning a distinct *Uniform Resource Identifier* (URI) to it. An example is shown in Figure 2.11.

As for NER, traditional approaches for *Named Entity Disambiguation* (NED) use hand-crafted features (e.g., *Bag of Words* (BoW) or *Term Frequency–Inverse Document Frequency* (TF-IDF)) to calculate the similarity between a given string (also called mention) of a named entity to its candidate linked entities in a knowledge base [12, 233]. Some of these approaches, e.g., *DBpedia Spotlight* [56, 163] or Fang et al. [70], solely use their lexical similarity or empirical co-occurrence of these candidates to disambiguate each mention.

---

<sup>12</sup>Documentation of spaCy version 3: <https://spacy.io/usage/v3>



The [Leibniz University Hannover](#) is a [public research university](#) located in [Hanover, Germany](#). Founded on 2 May 1831, it is one of the largest and oldest [science and technology universities](#) in [Germany](#). In the 2014/15 [school year](#) it enrolled 25,688 students, of which 2,121 were from foreign [countries](#). It has nine faculties which offer 190 full and [part degree programs](#) in 38 fields of study. It was named [University of Hannover](#) in 1978. In 2006, it was named after [Gottfried Wilhelm Leibniz](#), the 18th century mathematician and [philosopher](#). In 2018, [Leibniz University Hannover](#) was adopted as the official [English name](#).

Entity Type	Entity Name	Description	Salience
Organization	University of Hanover	public university located in Hanover, Germany	0.94
Location	Hanover	capital city of the German federated state of Lower Saxony	0.57
Location	Germany	country in central Europe	0.61
Person	Gottfried Wilhelm Leibniz	German mathematician and philosopher	0.16
Concept	Public university	university that is predominantly funded by public means	0.00

Figure 2.11: Exemplary output of *Ambiverse* [98] for *Named Entity Disambiguation* (NED). Named entities are linked to entries in *Wikipedia* to assign distinct URIs. The screenshot is taken from: <https://ambiversenlu.mpi-inf.mpg.de/>

In this respect, the text context (e.g., surrounding words) and textual entity descriptions from external sources such as *Wikipedia* are used to rank the candidates. Other methods [40, 41, 55, 98, 167, 194, 263] additionally consider that entities mentioned in the same text tend to share similar topics and consequently aim to maximize the topic consistency within a document. Most approaches such as *AIDA* [98], *Agdistis* [263], *Babelfy* [167, 181], *TagME* [194], and *Wikifier* [40, 41] form a probabilistic graph that models the similarity between a mention (text span) and entity as well as the relationship between entities. The graph connectedness [167, 194, 263], *Pagerank* [40, 41], or dense sub-graphs containing exactly one connected entity per mention [98] are used to find the optimal entity set with the highest topic consistency.

Recent approaches for NED extensively apply deep neural networks [205, 231] and achieve competitive results [12]. Unlike traditional methods, they rely on word embeddings to represent the words in a continuous vector space and use either approaches from distributional semantics (Section 2.3.1), e.g., *Word2Vec* [164, 165] or deep neural networks to learn these features automatically. While first neural approaches [74, 242] considered all surrounding words (context) of a named entity as equally important, more recent proposals [67, 76] apply attention mechanisms to assign graded importance to words to improve NED. However, the

## 2 Foundations

majority of these solutions cannot link mentions to unseen entities, which is a huge drawback for many applications. For this reason, Logeswaran et al. [150] and *BLINK* [281] propose neural approaches to zero-shot entity linking that aim to link mentions in the text to unseen entities without in-domain labeled data.

Typically, NED requires text spans containing the raw text of the named entities found by NER as input. Some frameworks [98, 125, 132, 167, 183] also propose joint solutions for both *Named Entity Recognition and Disambiguation* (NER & NED). More comprehensive overviews can be found in recent surveys from Al-Moslmi et al. [12], Sevgili et al. [231], and Martínez-Rodríguez et al. [156]. We have used *Wikifier* for NED in Chapter 4 of this thesis. The details are presented in the following.

**Pagerank-based Wikification:** As mentioned above, there are several approaches for *Named Entity Disambiguation* (NED). Approaches that are based on the hypothesis that named entities in a given document tend to share similar topics [40, 41, 55, 98, 167, 194, 263] are one possible solution to disambiguate mentions of named entities. These approaches aim to maximize the topic consistency within a document. As shown in Figure 2.12, Brank et al. [40, 41] propose a global disambiguation method called *Wikifier* that constructs a mention-entity or mention-concept graph and computes the *PageRank* over it to disambiguate a set of entities with corresponding concepts from *Wikipedia*.

The mention-entity graph is a bipartite graph where the left set of vertices corresponds to the mentions of named entities extracted from the text document  $D$ , and the right set corresponds to distinct entities in *Wikipedia*.

Given a mention (text span)  $M$  of a named entity, an edge is assigned to a target entity  $E$  (if available), i.e., a link to the *Wikipedia* page of entity  $E$ . For example, a text document might mention  $M = \textit{Hanover}$ . This string is used as a link to different entities (pages) in *Wikipedia*, like the capital city of the German federated state of Lower Saxony (<https://en.wikipedia.org/wiki/Hannover>) or a town in the U.S state Indiana ([https://en.wikipedia.org/wiki/Hanover,\\_Indiana](https://en.wikipedia.org/wiki/Hanover,_Indiana)). A transition probability  $p(M \rightarrow E)$  is assigned to each of these edges according to the ratio:

$$p(M \rightarrow E) = \frac{n_{M \rightarrow E}}{n_M}, \quad (2.8)$$

where  $n_M$  is the total number of times the mention  $M$  is used as anchor text in *Wikipedia* and  $n_{M \rightarrow E}$  is the number of times the mention  $M$  links to the specific *Wikipedia* page of entity  $E$ .

The graph is subsequently augmented by relations between entities  $E \rightarrow E'$  to capture the semantic relationships between concepts. The intuition behind this step is that semantically related entities often occur together within the same document. The internal link structure

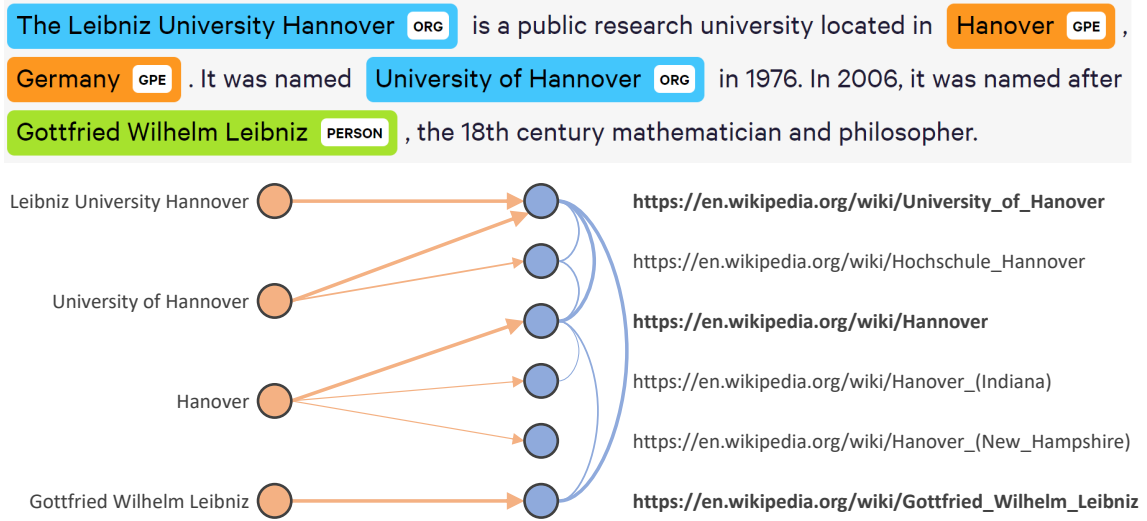


Figure 2.12: *Named Entity Disambiguation* (NED) with *Wikifier* [40, 41] using a mention-entity graph (bottom) based on mentions extracted from the text (top, the screenshot is taken from: <https://explosion.ai/demos/displacy-ent>). The number of times an anchor text of a mention  $M$  (left, orange vertices) links to a specific entity  $E$  (right, blue vertices) defines the transition probability  $p(M \rightarrow E)$  (orange lines). The semantic relatedness (blue lines) measures the proportion of pages that link from one entity to another and vice versa. The line widths indicate the size of the (fictional) values.

of the *Wikipedia* is used to calculate the *semantic relatedness*  $p(E \rightarrow E')$ . Let  $\mathbb{L}_E$  and  $\mathbb{L}'_E$  be the set of *Wikipedia* pages that contain links to the pages of entity  $E$  and entity  $E'$ , respectively, and  $|\mathbb{E}|$  the total number of entities in *Wikipedia*, then the semantic relatedness is defined as follows:

$$p(E \rightarrow E') = 1 - \frac{\log \max(|\mathbb{L}_E|, |\mathbb{L}'_E|) - \log |\mathbb{L}_E \cap \mathbb{L}'_E|}{\log \mathbb{E} - \log \min(|\mathbb{L}_E|, |\mathbb{L}'_E|)} \quad (2.9)$$

According to this equation, two entities are considered semantically related if a large proportion of pages linking to one of these entities also links to the other and vice versa. For each vertex in the mention-entity graph, a vector of *PageRank* scores according to Page et al. [188] is calculated. Finally, the entity with the highest *PageRank* score is used to disambiguate the corresponding mention.

## 2.4 Semantic Web & Knowledge Graphs

A knowledge graph, or knowledge base, is a data model and format that stores information in a structured way for easy processing and interpretation by machines. Its origins are based on the vision of the *Semantic Web* proposed by Berners-Lee et al. [33] in 2001, which can be conceived as an extension of the World Wide Web. The *Semantic Web* supports the in-

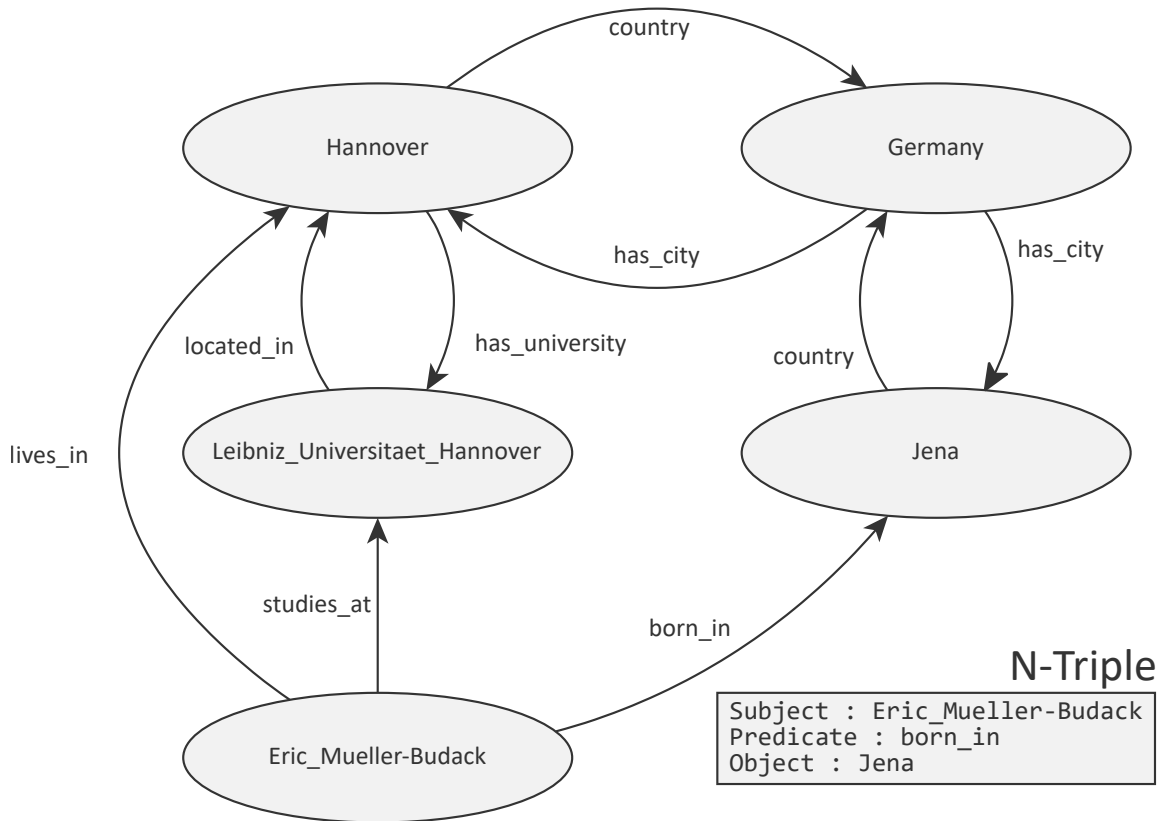


Figure 2.13: Exemplary *Resource Description Framework* (RDF) graph that visualizes the relations (edges) between different resources (vertices) as well as an RDF statement (bottom right) in N-Triple format.

clusion of semantic content and meta information into unstructured *Hypertext Markup Language* (HTML) documents using standards such as *Resource Description Framework* (RDF) and *Web Ontology Language* (OWL).

RDF is a flexible, structured data model, which uses data triples to formulate a statement about (web) resources. These data triples are defined as  $\text{subject} \rightarrow \text{predicate} \rightarrow \text{object}$ . The subject represents the resource, and the predicate defines the relation between the subject and the object, which can be another resource or a literal. Resources can represent anything from a person, location, or event to more general or abstract objects and are described by a distinct *Uniform Resource Identifier* (URI). Figure 2.13 shows a collection of RDF statements intrinsically represents a labeled, directed multi-graph  $\mathcal{G}(\mathbb{V}, \mathbb{E})$ , where the resources form a set of vertices  $\mathbb{V}$  and the predicates a set of edges  $\mathbb{E}$ .

*Linked Open Data* allows everyone to contribute to the *Semantic Web* by publishing structured RDF data. The main vision is to link resources to other datasets on the Web using the same standards in order to build a globally linked database. In this way, persons and machines can explore the Web of data and find other related data. The links between the same

resources across different databases can be established by the OWL attribute "owl:sameAs". The rules were formalized by Berners-Lee [32] in 2006:

- Use URIs as names for things.
- Use *Hypertext Transfer Protocol* (HTTP) URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF, *SPARQL Protocol and RDF Query Language* (SPARQL))
- Include links to other URIs, so that they can discover more things.

The ability to publish structured RDF data as *Linked Open Data* enabled the creation of many domain-specific and cross-domain knowledge bases over the last decades. In the scope of this thesis, knowledge bases such as *Wikidata* [268], *DBpedia* [22], or *YAGO* [241] are of particular interest because they contain factual information and relations for named entities and concepts. They provide a great source of world knowledge that can be used to quantify cross-modal consistency in multimodal news articles. For example, *Wikidata* provides a free and open knowledge base containing nearly 100 million items<sup>13</sup> that can be edited by both humans and machines and acts as central storage for other sister projects, including *Wikipedia*.

---

<sup>13</sup>Archived statistics from 8th August 2021: <https://web.archive.org/web/20210811234343/https://www.wikidata.org/wiki/Special:Statistics>



## 3 Information Extraction from Photos

News articles report on worldwide events and typically cover a variety of entities such as locations, dates, persons, and the events themselves. As mentioned in Section 1.3, the goal of this thesis is to present an unsupervised approach that evaluates cross-modal relations for these types of named entities to provide more differentiated measures of *Cross-modal Mutual Information* (CMI) compared to previous work [96, 127, 185, 294, 306]. This chapter presents novel solutions that extract rich information from photos and evaluates them to estimate their capabilities in recognizing these types of entities (research question 2). This information is an important prerequisite for quantifying the cross-modal presence of entities in articles containing photos and text, according to Chapter 4. As discussed in Section 1.2, there are relatively few approaches with some limitations that focus on identifying events, locations, and dates from photos. For this reason, novel approaches for event classification (Section 3.1), geolocation estimation (Section 3.2), and date estimation (Section 3.3) are proposed in the remainder of this chapter. We evaluate the impact of integrating contextual information, e.g., from related tasks or knowledge bases, into deep learning models to improve image recognition and interpretation (research question 3). Unlike the aforementioned computer vision areas, person identification is a very well-studied computer vision problem that has already attracted attention for decades. Section 3.4 presents an unsupervised approach that addresses application-specific challenges for person recognition, such as the web-based retrieval of example images, to automatically identify relevant persons (e.g., politicians or actors) in news articles extracted from the *Internet Archive*.

### 3.1 Event Classification of Photos

News articles cover events of various domains, such as society, culture, politics, or sports, that are of significant importance to a target audience. As a consequence, event classification in photos is an essential task for various applications. It enables semantic search or semantic retrieval in archives and news collections. In addition, it provides valuable features for multimedia approaches [114, 175, 219] to quantify image-text relations that can help to understand the overall multimodal message and sentiment or might even indicate misinformation, i.e., *Fake News*.

Despite its clear potential, so far, only a few approaches [8, 36, 113, 136, 287] were proposed that aim to recognize types of real-world events in contrast to other computer vision tasks. Event classification is a challenging task in many regards, such as data collection, visual similarities of related event types like *elections* and *political campaigns*, and class imbalance due to the large number of expected (scheduled or regular) compared to unexpected or rare events. Datasets for event classification mostly cover only specific event categories, e.g., social [3, 8, 211], sports [136], or cultural events [66]. To the best of our knowledge, the *Web Image Dataset for Event Recognition* (WIDER) [287] is the largest corpus with 50,574 photos that considers a variety of event types (61). Nonetheless, many event types that are important for news, like *epidemics* or *natural disasters*, are missing. Due to the lack of large-scale datasets, related work has focused on ensemble approaches [5, 6, 273] typically based on pre-trained models for object and place classification and the integration of descriptors from local image regions [4, 78, 86, 287] to learn features for event classification. One of the main challenges is to define a complete lexicon of important event categories. For this purpose, Ahsan et al. [8] suggest to mine *Wikipedia* and gathered 150 generic social events. However, the experiments were only conducted on WIDER and two other datasets, namely *Social Event Image Dataset* (SocEID) and *Rare Event Dataset* (RED), which cover eight social event types and a selection of 21 real-world events. Progress in the field of Semantic Web has shown that it is possible to define a knowledge graph for newsworthy events [81, 82] but has not been leveraged by computer vision approaches yet. Particularly the relations between events extracted from a knowledge base such as *Wikidata* [268] provide valuable information that can be used to train powerful models for event classification.

In this section, we introduce an ontology along with a dataset that enables us to train a novel ontology-driven deep learning approach for event classification. Our **primary contributions** can be summarized as follows: (1) Based on a set of real-world events from *EventKG* [81, 82], we propose the *Visual Event Ontology* (VisE-O) containing 409 nodes that describe 148 unique event types such as different kinds of sports, disasters, and social events with high news potential. The ontology can be created with little supervision and covers the largest number of event types for image classification to date. (2) In order to train deep learning models, we have gathered a large-scale dataset, called *Visual Event*



### 3.1 Event Classification of Photos

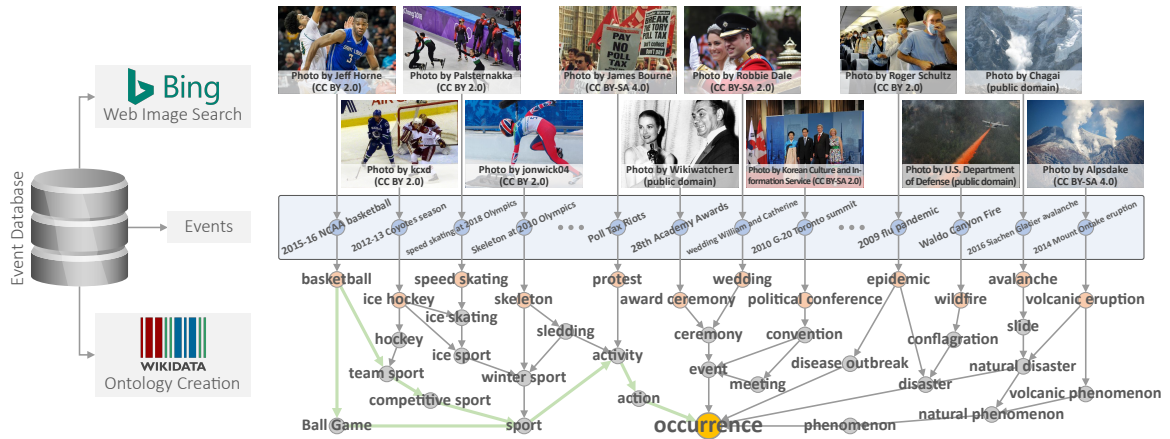


Figure 3.1: Exemplary subset of the *Ontology* (complete version is provided on our *GitHub* page<sup>14</sup>) and photos of the *Visual Event Classification Dataset* (VisE-D). *Leaf Event Nodes* (orange) and *Branch Event Nodes* (gray) are extracted based on relations (e.g., "subclass of") to a set of *Events* (blue) using the *Wikidata* knowledge base. The nodes connected by the green path define the *Subgraph of basketball* to the *Root Node* (yellow). The combination (union) of all *Subgraphs* defines the *Ontology*. Definitions are according to Section 3.1.2.1.

*Classification Dataset* (VisE-D), of 570,540 photos crawled automatically from the Web. It contains 531,080 training and 28,543 validation photos, as well as two test sets with 2,779 manual annotated and 8,138 *Wikimedia* photos. Figure 3.1 depicts some example images. (3) We provide several baselines, including ontology-driven deep learning approaches that integrate the relations of event types extracted from structured information in the ontology using novel loss functions and weighting schemes. As a consequence, it can understand the fundamental differences of event types in different domains such as *sports*, *crimes*, or *natural disasters*. Experimental results on several benchmark datasets demonstrate the feasibility of the proposed approaches. Dataset and source code are publicly available<sup>14</sup>.

The remainder of this section is organized as follows. Section 3.1.1 reviews related work on event classification of photos. The ontology and dataset for newsworthy event types are presented in Section 3.1.2. In Section 3.1.3, we propose ontology-driven deep learning approaches for event classification. Experimental results for several benchmarks are presented in Section 3.1.4. Section 3.1.5 summarizes this work and outlines areas of future work.

#### 3.1.1 Related Work

Since there are different definitions of an event, approaches for event classification are very diverse and range from specific actions in videos [259, 289] over the classification of more personal events in photo collections [25, 36, 284] to the classification of social, cultural, and

<sup>14</sup><https://github.com/TIBHannover/VisE>

### 3 Information Extraction from Photos

sport event types in photos [86, 136, 273, 287]. In the sequel, datasets and proposals for the recognition of events and event types in images with potential news character are reviewed.

Early approaches for event classification have used hand-crafted features such as *Scale-invariant Feature Transform* (SIFT) to classify events in particular domains like sports [113, 136]. As one of the first deep learning approaches, Xiong et al. [287] suggested a multi-layer framework that leverages two CNNs to incorporate the visual appearance of the whole image as well as interactions among humans and objects. Similarly, several approaches aim to integrate local information from image patches or regions extracted by object detection frameworks [4, 78, 86] to learn rich features for event classification. Guo et al. [86] proposed *Graph Convolutional Neural Networks* (GCNNs) to leverage relations between objects. Another line of studies applies ensemble models and feature combinations [5, 6, 273] to exploit the capabilities of deep learning models trained for different computer vision tasks, most typically for object and place (or scene) classification. In the absence of a large-scale dataset for many event types, Ahsan et al. [8] train classifiers based on images crawled for a set of 150 social event concepts mined from *Wikipedia*, while Wang et al. [273] apply transfer learning to object and place representations to learn compact representations for event recognition with few training images. A more detailed review of deep learning techniques for event classification can be found in the survey from Ahmad and Conci [2].

There are many datasets and also challenges such as the *MediaEval Social Event Detection Task* [211] and *ChaLearn Looking at People* [66] for event classification. However, they mostly cover specific domains such as social events [3, 211], cultural events [66], or sports [136]. Besides, the datasets are either too small [136] to train deep learning models or contain very few and incomplete event classes [3]. Other proposals have introduced datasets and approaches to detect concrete real-world news events [8, 66, 78] but only distinguish between a small pre-defined selection of events. To the best of our knowledge, WIDER [287] is the most complete dataset in terms of the number of event categories that can be leveraged by deep learning approaches. It contains 50,574 images for 61 event types, but many important event types for news, such as *epidemics* or *natural disasters*, are missing.

#### 3.1.2 Dataset and Ontology for Event Type Classification

In contrast to prior solutions, this section presents an ontology and dataset for event classification that covers a larger number of event types important for news across all domains such as *sports*, *crimes*, and *natural disasters*. Based on definitions for terms and notations (Section 3.1.2.1), we suggest an approach that leverages events identified by *EventKG* [81, 82] to automatically retrieve an ontology that can be refined with little supervision (Section 3.1.2.2). Images for event types in the resulting VisE-O are crawled from the Web to create the VisE-D according to Section 3.1.2.3.

### 3.1.2.1 Notations & Definitions

In this section, we introduce definitions and notations that are used in the remainder of this section. Figure 3.1 contains supplementary visualizations to clarify the definitions.

**Event:** In computer vision, an event can refer to many things, e.g., specific (inter)actions in videos [259, 289] or general social events in everyday life [8, 287]. As in the *EventKG* [81], we define an *Event* as contemporary or historical happening of global importance (e.g., *2011 NBA Finals* in Figure 3.1) that is connected to one or multiple place(s) and time(s) or time period(s). This definition matches our overall goal to evaluate the cross-modal consistency of events in news that are important for a large target audience. Based on this definition of an event  $e$ , we create a set of events  $\mathbb{E}$  to construct an *Ontology*.

**Ontology, Root Node, Event Nodes, and Relations:** The *Ontology* is a directed graph  $\mathcal{G}(\mathbb{V}, \mathbb{R})$  composed of a set of *Event Nodes*  $\mathbb{V}$  as vertices and their corresponding *Relations*  $\mathbb{R}$  as edges. *Relations*  $\mathbb{R}$  are knowledge base specific properties such as "subclass of" in *Wikidata* [268] that describe the interrelations of *Event Nodes*  $\mathbb{V}$ . All parent nodes  $v \in \mathbb{V}$  that connect a specific *Event*  $e \in \mathbb{E}$  to the *Root Node* are denoted as *Event Nodes*. The *Root Node*  $v_R \in \mathbb{V}$  (e.g., *occurrence* in Figure 3.1) matches the general definition of an *Event* and represents a parent node shared by all *Events*.

**Leaf and Branch Event Nodes:** The *Leaf Event Nodes*  $\mathbb{V}_L \subset \mathbb{V}$  such as *basketball*, are the most detailed *Event Nodes* without children in the *Ontology*. They group *Events* of the same type, e.g., *2011 NBA Finals*  $\rightarrow$  *basketball* (Figure 3.1). *Event Nodes*, e.g., *ball game*, with at least one child node are referred to as *Branch Event Nodes*  $\mathbb{V}_B \subset \mathbb{V}$ .

**Subgraph:** A *Subgraph*  $\mathbb{S}_L$  is a set of all *Event Nodes*  $\mathbb{S}_L = \{v_L, \dots, v_R\} \subset \mathbb{V}$  related to a specified *Leaf Event Node*  $v_L \in \mathbb{V}_L$  while traversing to the *Root Node*  $v_R$ .

### 3.1.2.2 VisE-O: Visual Event Ontology

**Knowledge Base and Root Node Selection:** Several knowledge bases such as *DBpedia* [22], *YAGO* [241], or *Wikidata* [268] are available. We investigated them in terms of event granularity and correctness. The whole *DBpedia* ontology contains less than 1,000 classes. Thus, the granularity of potential event types is very coarse, and for instance, some types of natural disasters are either assigned to a wrong (*Tsunami*  $\rightarrow$  *television show*)<sup>15</sup> or generic

<sup>15</sup>Internet Archive snapshot for "Tsunami" from 14th February 2020: <https://web.archive.org/web/20200214202750/http://dbpedia.org/page/Tsunami>

### 3 Information Extraction from Photos

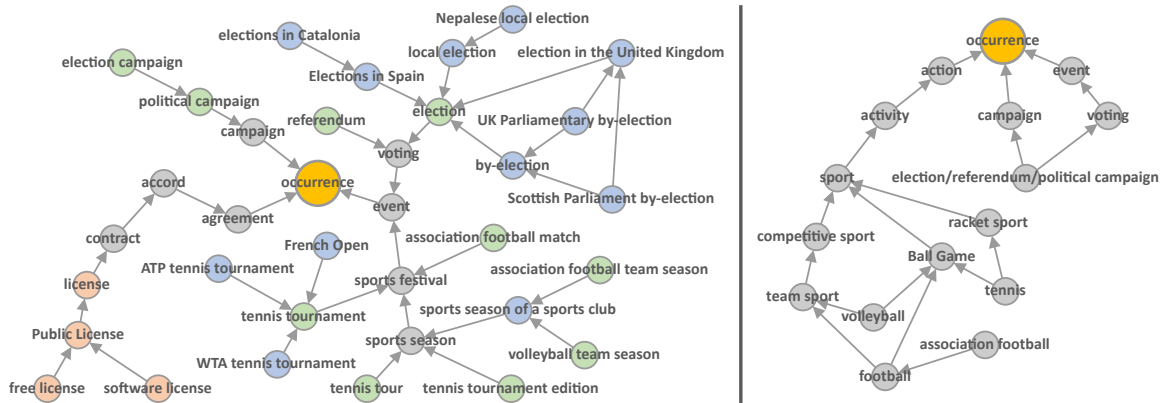


Figure 3.2: Exemplary subset of the initial *Ontology* after the extraction of all relations from *Wikidata* (left) and the respective final *Ontology* after applying the proposed approaches for event class disambiguation and refinement (right). Blue *Event Nodes* represent the same event type and might be too fine-granular to distinguish. Green nodes are semantically and visually similar to other *Event Nodes* in the *Ontology* and can therefore be ambiguous. Orange nodes do not represent an *Event* according to the definition in Section 3.1.2.1. Best viewed in color. Different versions of the *Ontology* can be explored on our *GitHub* page<sup>14</sup>.

classes (*Earthquake*  $\rightarrow$  *thing*)<sup>16</sup>. As mentioned by Gottschalk and Demidova [81], *YAGO* also contains noisy event categories. On the contrary, *Wikidata* offers fine-granular event types and relations, as shown in Figure 3.2, and is therefore used as the knowledge base in our approach. Gottschalk and Demidova [81] identified that *Events* are subclasses of *Wikidata*'s knowledge base entries *event* (*Wikidata* identifier *Q1656682*) and *occurrence* (*Q1190554*). We select *occurrence* as the *Root Node* of our *Ontology* since it is a parent of *event* (Figure 3.2) according to *Wikidata* and consequently covers more event instances.

**Automated Creation of an Initial Event Ontology:** A *bottom-up approach* is applied to create an event ontology automatically. Based on a large set of  $|\mathbb{E}| = 550,994$  real-world events<sup>17</sup> from *EventKG* [81, 82], we recursively obtain all parent *Event Nodes* from *Wikidata*. For *Event Nodes*, only relations of the type "subclass of" (*Wikidata* property *P279*) are considered since they already describe specific categories. For *Events*, we additionally allow for the properties "instance of" (*P31*) and "part of" (*P361*) as relations to increase the coverage because some events like *2018 FIFA World Cup Group A* are not a "subclass of" an *Event Node* but "part of" a superordinate event, in this case, *2018 FIFA World Cup*. Finally, we remove all *Event Nodes* that are not connected to the *Root Node*. As illustrated in Figure 3.1, the resulting *Subgraphs* of all *Leaf Event Nodes* are combined to generate the *Initial Event Ontology*.

<sup>16</sup>Internet Archive snapshot for "Earthquake" from 18th February 2020: <https://web.archive.org/web/20200218100604/http://dbpedia.org/page/Earthquake>

<sup>17</sup>List of events available at: [http://eventkg.l3s.uni-hannover.de/data/event\\_list.tsv](http://eventkg.l3s.uni-hannover.de/data/event_list.tsv)

We have also investigated a *top-down approach* where the child nodes of the *Root Node* are extracted recursively to obtain *Event Nodes*. Even though this might lead to a more complete event ontology, we found that the proposed *bottom-up approach* offers several advantages. Most importantly, the names of real-world events provide more sophisticated queries to retrieve images from the Web automatically, as explained in Section 3.1.2.3. These queries can also prevent possible selection bias of image search engines since, for example, queries constructed by names of *Event Nodes* such as "*election*" might only crawl images of the latest *U.S. elections*. In addition, a *top-down approach* obtains many irrelevant *Event Nodes* for event classification that were already neglected by *EventKG* [81, 82]. Given the large number of real-world events, we argue that the proposed *bottom-up approach* covers the most relevant *Event Nodes* while containing significantly fewer irrelevant ones.

However, we identified several problems in the *Initial Event Ontology* as illustrated in Figure 3.2: (1) There are differences in the granularity of *Leaf Event Nodes*, and some of them are too fine-grained, e.g., *ATP tennis tournament* or *Nepalese local election*, and might be hard to recognize; (2) in particular, sports-centric *Leaf Event Nodes* such as *association football match* and *association football team season* are ambiguous as they are semantically and visually similar; (3) some *Event Nodes*, e.g., *software license* do not represent an *Event* according to the definition in Section 3.1.2.1.

**Event Class Disambiguation** As pointed out above, most *Leaf Event Nodes* related to sports are ambiguous since they represent the same type of sport. The *Wikidata* knowledge base [268] distinguishes between *sports seasons*, *sports competitions*, etc. Although this structure might make sense for some applications, we aim to combine *Event Nodes* related to the same sports. However, this is not possible with the *Initial Event Ontology* that relies on "*subclass of*" *Relations*. As illustrated in Figure 3.2 (green nodes), *Event Nodes* of different sports domains (e.g., *volleyball team season* and *association football team season*) relate to a particular type of competition (here *team season*) before they relate to another *Event Node* of the same sports type (*association football match*). Value(s) for the *Wikidata* property "*sport*" (*P641*) for each *Event* and *Event Node* were extracted and used as *Relation* (if available) to solve this issue. As a result, sports events were combined according to their sports category rather than the type of the competition, as shown in Figure 3.2 (right). In addition, we delete all *Event Nodes* that are a parent of less than a minimum number of  $|\mathbb{E}|_{min} = 10$  *Events* to reduce the granularity of the resulting *Leaf Event Nodes*.

These strategies lead to the *Disambiguated Event Ontology* that, unlike datasets from related work [3, 8, 66, 136, 287] with manually selected event classes, can be constructed automatically. It can already be used for many applications and provides a hierarchical overview to more efficiently explore and select event classes for specific tasks. However, it can still contain irrelevant *Event Nodes*. Furthermore, expected (scheduled or regular) events such as *elections* or *sports festivals* occur more frequently than unexpected or rare events

### 3 Information Extraction from Photos

Table 3.1: Number of *Event Nodes*  $|\mathbb{V}|$ , *Leaf Event Nodes*  $|\mathbb{V}_L|$ , *Relations*  $|\mathbb{R}|$ , and images  $|\mathbb{I}|$  for training (T), validation (V), and test (B - *VisE-Bing*, W - *VisE-Wiki*). All ontologies can link approximately the same number of *Events*  $|\mathbb{E}|$  to any *Event Node*. However, the *Refined Event Ontology* can link the most *Events*  $|\hat{\mathbb{E}}|$  without ambiguities to a *Leaf Event Node* while reducing the complexity of the *Ontology*.

Ontology	Ontology Statistics					Dataset Statistics			
	$ \mathbb{E} $	$ \hat{\mathbb{E}} $	$ \mathbb{V} $	$ \mathbb{V}_L $	$ \mathbb{R} $	$ \mathbb{I}_T $	$ \mathbb{I}_V $	$ \mathbb{I}_B $	$ \mathbb{I}_W $
Initial	526,853	236,464	6,114	3,578	7,845	—	—	—	—
Disamb.	529,932	163,570	2,288	1,081	3,144	—	—	—	—
Refined	529,932	447,161	409	148	635	531,080	28,543	2,779	8,138

such as *epidemics* or *natural disasters*. *Leaf Event Nodes* representing expected event types more likely fulfill the filtering criteria  $|\mathbb{E}|_{min}$  and are consequently very fine-grained (e.g., elections in different countries), making them more difficult to distinguish. Thus, we decided to refine the *Disambiguated Event Ontology* manually.

**Event Ontology Refinement:** Two annotators manually refined the *Disambiguated Event Ontology* to create a challenging yet useful and fair ontology for image classification. To pursue this goal, the ontology was refined according to two criteria: (1) reject *Event Nodes* that do not match the *Event* definition in Section 3.1.2.1, and (2) select the most suitable *Leaf Event Nodes* to prevent ambiguities. For example, *election* was chosen as a representative *Leaf Event Node* since its children contain different types of elections (e.g., *by-election*) and elections in different countries (e.g., *elections in Spain*) that might be too hard to distinguish. We can use the hierarchical information to assign the children to the selected *Leaf Event Nodes* automatically and simultaneously remove all resulting *Branch Event Nodes* as candidates. Therefore, only around 500 annotations were necessary to label all 2,288 *Event Nodes* extracted from the previous steps. Finally, we manually merged 30 *Leaf Event Nodes* such as *elections*, *political campaign*, and *referendum* or *award* and *award ceremony* that are semantically similar but could not be fused using the *Disambiguated Event Ontology*. The statistics for all variants of the *Ontology* are shown in Table 3.1 and reveal that the *Refined Event Ontology* links the most *Events* to *Leaf Event Nodes*. In the preliminary *Ontologies*, many *Events* are children of *Branch Event Nodes*, and it is not possible to use them to query example images for a specific *Leaf Event Node*, as explained in the next section. The 148 *Leaf Event Nodes* used in this thesis can be found in Appendix A.1.1. The *Ontologies* presented in this section can be explored in the *GitHub* repository:

- Initial Ontology: [https://tibhannover.github.io/VisE/VisE-0\\_initial](https://tibhannover.github.io/VisE/VisE-0_initial)
- Disamb. Ontology: [https://tibhannover.github.io/VisE/VisE-0\\_disambiguated](https://tibhannover.github.io/VisE/VisE-0_disambiguated)
- Refined Ontology: [https://tibhannover.github.io/VisE/VisE-0\\_refined](https://tibhannover.github.io/VisE/VisE-0_refined)

### 3.1.2.3 Visual Event Classification Dataset

**Data Collection:** To create a large-scale dataset for the proposed *Ontology*, we defined different queries to crawl representative images from *Bing*. A maximum of 1,000 images (500 without restrictions and another 500 uploaded within the last year) using the names of the *Leaf Event Nodes* were crawled. In addition, the names of popular *Events* related to a *Leaf Event Node* that happened after 1900 were used as queries to increase the number of images and reduce ambiguities (e.g., *Skeleton at the 2018 Winter Olympics* for *Skeleton* in Figure 3.1). A sampling strategy was applied to set the number of images downloaded for an *Event* based on its *popularity*, i.e., number of *Wikipedia* page views, and date to prevent irrelevant images in the search results. Since less important events tend to contain more unrelated photos in the image search results, we only consider *Events* that were viewed at least 100 times per day on average. Images of historical events also typically contain less relevant images for news, e.g., drawings and scans. To further reduce the amount of irrelevant images, we emphasize significant events in the last decade(s). Thus, we used the page views  $v_e$  and the number of years  $a_e$  an *Event*  $e \in \mathbb{E}$  dates back to calculate the desired amount of images  $|\mathbb{I}_e|$  to crawl from *Bing* according to the following equation:

$$|\mathbb{I}_e| = \min \left( |\mathbb{I}_{\max}|, \frac{k_S \cdot v_e}{\max(1, a_e) \cdot v_s} \right). \quad (3.1)$$

The sampling parameter  $k_S$  controls the number of images to be crawled and  $v_s$  denotes the number of views of all *Events* that are children of the specified *Leaf Event Node* according to *Wikidata*. This normalization is used to achieve a more equal distribution of images crawled for *Leaf Event Nodes* because some event types are less popular than others, e.g., *skeleton* compared to *basketball*. We used  $k_S = 40,000$  for sampling and downloaded a maximum of  $|\mathbb{I}_{\max}| = 1,000$  images for the most popular events that represent a *Leaf Event Node*.

**Ground-truth Labels:** We provide two ground-truth vectors for each image based on the search query: (1) The **Leaf Node Vector**  $\mathbf{y}_L \in \{0, 1\}^{|\mathbb{V}_L|}$  indicates which of the  $|\mathbb{V}_L| = 148$  *Leaf Event Nodes* are related to the image and serves for classification tasks without using *Ontology* information. Note that  $\mathbf{y}_L$  is multi-hot encoded as a queried *Event* (e.g., *SpaceX Lunar Tourism Mission*  $\rightarrow$  *spaceflight* and *expedition*) can relate to multiple *Leaf Event Nodes*; (2) the multi-hot encoded **Subgraph Vector**  $\mathbf{y}_S \in \{0, 1\}^{|\mathbb{V}|}$  denotes which of the  $|\mathbb{V}| = 409$  *Event Nodes* (*Leaf* and *Branch*) are in the *Subgraphs* of all related *Leaf Event Nodes* and enables to learn from *Ontology* information.

**Splits:** We were able to download about 588,000 images, which are divided into three splits for training (90%), validation (5%), and test (5%). We only use images from *Events* related to exactly one *Leaf Event Node* for the test set. Test images that are a duplicate (us-

### 3 Information Extraction from Photos

ing the image hash) of a training or validation image are removed. Overall, the dataset provides an interesting challenge since it (1) contains the largest number of event types for event recognition to date, (2) can be considered large-scale and thus allows to train neural networks, (3) enables approaches that learn from structured ontology information, and (4) contains irrelevant web images for training which allows measuring the impact of self- or semi-supervised deep learning techniques.

**VisE-Bing Test Set:** Two annotators verified whether a test image depicts the respective *Leaf Event Node* or not. Each annotator labeled a maximum of ten valid images for each *Leaf Event Node* to prevent bias in the test dataset. The annotators received different sets of images to increase the number of test images. As a result, we were able to obtain 20 verified test images for most (109) of the 148 *Leaf Event Nodes*. The final dataset statistics are reported in Table 3.1, and the dataset distribution, including a list of all 148 *Leaf Event Nodes*, is reported in Appendix A.1.1.

**VisE-Wiki Test Set:** To create another larger test set, we downloaded all *Wikimedia* images for each *Leaf Event Node* and its child *Events* using the *Commons category* (*Wikidata* property *P373*) linked in *Wikidata*. Although *Wikimedia* is a trusted source, we noticed some less relevant images for news, e.g., historical drawings or scans. We applied a k-nearest-neighbor classifier based on the embeddings of a *ResNet-50* [92] trained on the ILSVRC 2012 dataset [58, 218]. For each manually verified test image in *VisE-Bing*, we selected the  $k = 100/|\mathbb{I}_a^v|$  nearest images, where  $|\mathbb{I}_a^v|$  is the number of annotated images of the *Leaf Event Node*  $v \in \mathbb{V}_L$  in *VisE-Bing*. The test set comprises 8,138 images for 146 of 148 classes. Detailed dataset statistics are reported in Appendix A.1.1.

#### 3.1.3 Ontology-Driven Event Classification

In this section, we propose a baseline classification approach (Section 3.1.3.1) and more advanced strategies as well as weighting schemes to integrate event type relations from the *Ontology* in the network training (Section 3.1.3.2). Section 3.1.3.3 introduces the inference strategies adopted in the testing scenario.

##### 3.1.3.1 Classification Approach

As shown in Table 3.1, the refined *Ontology* contains  $|\mathbb{V}_L| = 148$  *Leaf Event Nodes*. As a baseline classifier, we train a CNN that predicts *Leaf Event Nodes* without using ontology information. The *Leaf Node Vector*  $\mathbf{y}_L = \langle y_L^1, y_L^2, \dots, y_L^{148} \rangle \in \{0, 1\}^{|\mathbb{V}_L|=148}$  from Section 3.1.2.3 is used as the target for optimization. We add a fully-connected layer on top of a CNN architecture such as the *ResNet-50* [92] with  $|\mathbb{V}_L| = 148$  neurons. As an image can



depict multiple event types, a *sigmoid* activation function (Figure 2.2) is used that outputs a probability vector  $\hat{\mathbf{y}}_L = \langle \hat{y}_L^1, \hat{y}_L^2, \dots, \hat{y}_L^{148} \rangle$  where each entry  $\hat{y}_L^i$  ranges between  $0 \leq \hat{y}_L^i \leq 1$  to learn from the multi-hot encoded *Leaf Node Vector*. During training, the cross-entropy loss  $\mathcal{L}_c$  is optimized according to :

$$\mathcal{L}_c = - \sum_{i=1}^{|\mathbb{V}_L|} y_L^i \cdot \log \hat{y}_L^i \quad (3.2)$$

### 3.1.3.2 Integration of Ontology Information

In order to integrate information from the proposed *Ontology* in Section 3.1.2.2, we use the multi-hot encoded *Subgraph Vector*  $\mathbf{y}_S = \langle y_S^1, y_S^2, \dots, y_S^{409} \rangle \in \{0, 1\}^{|\mathbb{V}|=409}$  introduced in Section 3.1.2.3 that includes the relations to all  $|\mathbb{V}| = 409$  *Event Nodes* as a target. Consequently, a fully-connected layer with  $|\mathbb{V}| = 409$  neurons is added on top of a CNN architecture. As in the previous section, a *sigmoid* activation function is used to predict a probability vector  $\hat{\mathbf{y}}_S = \langle \hat{y}_S^1, \hat{y}_S^2, \dots, \hat{y}_S^{409} \rangle$  with each entry  $\hat{y}_S^i \in [0, 1]$ . Two different loss functions are considered. As for the classification approach, we apply the cross-entropy loss on the sigmoid activations  $\hat{\mathbf{y}}_S$  to define an ontology-driven loss function:

$$\mathcal{L}_o^{cel} = - \sum_{i=1}^{|\mathbb{V}|} y_S^i \cdot \log \hat{y}_S^i \quad (3.3)$$

As an alternative, we minimize the cosine distance of the predicted  $\hat{\mathbf{y}}_S$  and the ground truth  $\mathbf{y}_S$  *Subgraph Vector*:

$$\mathcal{L}_o^{cos} = 1 - \frac{\mathbf{y}_S \cdot \hat{\mathbf{y}}_S}{\|\mathbf{y}_S\|_2 \cdot \|\hat{\mathbf{y}}_S\|_2} \quad (3.4)$$

The granularity and the number of *Event Nodes* within the *Subgraphs* of *Leaf Event Nodes* varies for different domains, e.g., *sports*, *elections*, or *natural disasters*. As a consequence, the loss might be difficult to optimize. In addition, *Branch Event Nodes* such as *action* or *process* represent general concepts shared by many *Leaf Event Nodes*. Some *Branch Event Nodes* are also redundant since they do not include more *Leaf Event Nodes* than their children. Based on these observations, we suggest several improvements as described below.

**Redundancy Removal:** To remove the redundancy in the proposed *Ontology*, every *Branch Event Node* related to the same set of *Leaf Event Nodes* compared to its child nodes in the *Ontology* is deleted. These nodes are redundant since they do not include any new relationship information concerning the considered event types, i.e., *Leaf Event Nodes*. As a result, we are able to reduce the size of the *Subgraph Vector*  $\mathbf{y}_S \in \{0, 1\}^{|\mathbb{V}|}$  from  $|\mathbb{V}| = 409$  to  $|\mathbb{V}_{RR}| = 245$ .

### 3 Information Extraction from Photos

**Node Weighting:** We investigated two weighting schemes to encourage the neural network to focus on *Leaf Event Nodes* and more informative *Branch Event Nodes* in the *Ontology*. Based on *one* of the schemes, each entry in the ground-truth  $\mathbf{y}_S$  and predicted  $\hat{\mathbf{y}}_S$  *Subgraph Vector* is multiplied with its corresponding weight before the loss according to Equation (3.3) or Equation (3.4) is calculated.

We propose a **Distance Weight**  $\gamma^v$  based on the distance of an *Event Node*  $v \in \mathbb{V}$  to all connected *Leaf Event Nodes* in the *Ontology*. First, the length  $l^v$  of the shortest path, including self-loops (this means that a node is always in its own path; therefore  $l^v > 0$ ), to each connected *Leaf Event Node* is determined. The average length  $\bar{l}^v$  of these paths is used to calculate the weight:

$$\gamma^v = \frac{1}{2^{(\bar{l}^v - 1)}}. \quad (3.5)$$

This weighting scheme encourages the network to learn from *Event Nodes* that are close to the *Leaf Event Nodes*. They describe detailed event types that are harder to distinguish. Please note that the average length  $\bar{l}^v$  can change if the redundancy removal is applied.

Similarly, we calculate a **Degree of Centrality Weight**  $\omega_v$  for each *Event Node*  $v \in \mathbb{V}$  based on the number  $c^v$  of *Leaf Event Nodes* connected to an *Event Node*  $v$  and the total number of *Leaf Event Nodes*  $|\mathbb{V}_L| = 148$ :

$$\omega^v = 1 - \frac{c^v - 1}{|\mathbb{V}_L|}. \quad (3.6)$$

According to Equation (3.6), the weights of all *Leaf Event Nodes* are set to  $\omega^v = 1, \forall v \in \mathbb{V}_L$  (denoted as  $\omega_L$ ), while, for instance, the *Root Node*  $v_R$  is weighted with  $\omega^{v_R} \approx 0$  because it is connected to all *Leaf Event Nodes*. Thus, the network should focus on learning unique event types such as *tsunami* or *carnival* rather than coarse superclasses related to many *Leaf Event Nodes*.

While the maximum weight of *Branch Event Nodes* using the *Distance Weights* is 0.5 and defined by the nodes closest to the *Leaf Event Nodes* ( $\bar{l}^v = 2$ ), their corresponding *Degree of Centrality Weight* can be close to  $\omega_L$ . To put more emphasis on *Leaf Event Nodes*, we set their weights to  $\omega_L > 1$ . We set these weights to  $\omega_L = 6$ , as discussed in detail in Section 3.1.4.3.

#### 3.1.3.3 Inference

The classification approach predicts a *Leaf Node Vector*  $\hat{\mathbf{y}}_L$  that contains the probabilities of the  $|\mathbb{V}_L| = 148$  *Leaf Event Nodes* that can be directly used for event classification. On the other hand, the ontology-driven network outputs a *Subgraph Vector*  $\hat{\mathbf{y}}_S$  with probabilities for all  $|\mathbb{V}| = 409$  or  $|\mathbb{V}_{RR}| = 245$  (with redundancy removal) *Event Nodes* in the *Ontology*. There are several options to retrieve a *Leaf Node Vector*  $\hat{\mathbf{y}}_L$  for classification using  $\hat{\mathbf{y}}_S$ .

- (1) We retrieve the probabilities  $\hat{\mathbf{y}}_L^o$  that are part of the predicted *Subgraph Vector*  $\hat{\mathbf{y}}_S$ .
- (2) The cosine similarity of the predicted *Subgraph Vector*  $\hat{\mathbf{y}}_S$  to the multi-hot encoded *Subgraph Vector*  $\mathbf{y}_S^v$  of each *Leaf Event Node*  $v \in \mathbb{V}_L$  is measured to leverage the probabilities of *Branch Event Nodes* as follows:

$$\hat{\mathbf{y}}_L^{cos} = \frac{\mathbf{y}_S^v \cdot \hat{\mathbf{y}}_S}{\|\mathbf{y}_S^v\|_2 \cdot \|\hat{\mathbf{y}}_S\|_2} \quad \forall v \in \mathbb{V}_L \quad (3.7)$$

Note that the ground truth and predicted *Subgraph Vectors* are first multiplied with the weights used during network training. As a result, we obtain  $|\mathbb{V}_L| = 148$  similarities that are stored as  $\hat{\mathbf{y}}_L^{cos} \in \mathbb{R}^{|\mathbb{V}_L|}$ .

We decided to use the elementwise product of both strategies  $\hat{\mathbf{y}}_L = \hat{\mathbf{y}}_L^o \odot \hat{\mathbf{y}}_L^{cos}$  as the prediction for the ontology approach since we found that this combination worked best in most cases. Results using the individual probabilities are reported in Appendix A.1.2.

### 3.1.4 Experimental Setup & Results

In this section, the parameters (Section 3.1.4.1), evaluation metrics (Section 3.1.4.2), and experimental results are presented. The experimental evaluation includes a comparison of the ontology-driven approaches to the classification baseline (Section 3.1.4.3), an analysis of results for specific event types (Section 3.1.4.4), and an evaluation on other benchmark datasets (3.1.4.5).

#### 3.1.4.1 Network Parameters

We used a *ResNet-50* [92] as the basic architecture for the proposed approaches. They were optimized using *Stochastic Gradient Descent* (SGD) with *Nesterov momentum* term [249], weight decay of  $1 \times 10^{-5}$ , and a batch size of 128 images. The initial learning rate of 0.01 is increased to 0.1 using a linear ramp up in the first 10,000 iterations to speed up the training. Then, a cosine learning rate annealing [151] is applied to lower the learning rate to zero after a total of 100,000 iterations. The model that achieves the lowest loss on the validation set is used for the experiments.

#### 3.1.4.2 Evaluation Metrics

We report the *top-1*, *top-3*, and *top-5* accuracy using the top-k predictions in the *Leaf Node Vector*  $\hat{\mathbf{y}}_L$  (Section 3.1.3.3). However, the accuracy does not reflect the similarity of the predicted to the ground-truth *Leaf Event Node* concerning the *Ontology* information. For this reason, we create a multi-hot encoded *Subgraph Vector*  $\tilde{\mathbf{y}}_S \in \{0, 1\}^{|\mathbb{V}|}$  representing the whole *Subgraph* of the predicted (*top-1*) *Leaf Event Node*  $\hat{v}$ . Note that the full *Subgraph Vector* with dimension  $|\mathbb{V}| = 409$  is created to generate comparable results for models

### 3 Information Extraction from Photos

trained with and without redundancy removal and that  $\tilde{\mathbf{y}}_S \neq \hat{\mathbf{y}}_S$  since  $\hat{\mathbf{y}}_S$  corresponds to the predicted *Subgraph Vector* of an ontology-driven approach, whereas  $\tilde{\mathbf{y}}_S$  is the multi-hot encoded *Subgraph Vector* of the predicted *Leaf Event Node*  $\hat{v}$ . We propose to measure the cosine similarity (*CS*) and *Jaccard Similarity Coefficient* (*JSC*) according to Equation (3.8) and Equation (3.9) between the multi-hot encoded *Subgraph Vector*  $\tilde{\mathbf{y}}_S$  of the predicted class and the ground-truth *Subgraph Vector*  $\mathbf{y}_S$  of the test image to quantify the similarity based on all  $|\mathbb{V}| = 409$  *Event Nodes*.

$$CS = \frac{\mathbf{y}_S \cdot \tilde{\mathbf{y}}_S}{\|\mathbf{y}_S\|_2 \cdot \|\tilde{\mathbf{y}}_S\|_2} \quad (3.8)$$

$$JSC = \frac{\|\mathbf{y}_S \odot \tilde{\mathbf{y}}_S\|_1}{\|\mathbf{y}_S\|_1 \cdot \|\tilde{\mathbf{y}}_S\|_1 + \|\mathbf{y}_S \odot \tilde{\mathbf{y}}_S\|_1} \quad (3.9)$$

#### 3.1.4.3 Ablation Study

The results of an ablation study, including the various proposed approaches on *VisE-Bing*, are presented in Table 3.2. The results of the ontology-driven approaches (denoted as *O*) are significantly worse without applying any weighting scheme. The reason is that the correct prediction of the majority of *Event Nodes* in a *Subgraph* is already sufficient to achieve low loss signals. However, the ontology-driven approaches benefit from the weighting schemes and clearly outperform the classification baseline (denoted as *C*). As discussed in Section 3.1.3.2, a higher weight  $\omega_L$  for *Leaf Event Nodes* needs to be assigned using the

Table 3.2: Results (numbers are multiplied by 100) on *VisE-Bing* using different loss functions, weighting schemes (WS), and ontology redundancy removal (RR)

Model Notation	Loss	WS	RR	Accuracy			<i>JSC</i>	<i>CS</i>
				<i>Top-1</i>	<i>Top-3</i>	<i>Top-5</i>		
<i>C</i>	$\mathcal{L}_c$			77.4	89.8	93.6	84.7	87.7
<i>O<sup>cel</sup></i>	$\mathcal{L}_o$			67.5	83.3	88.5	81.1	85.4
<i>O<sub>ω</sub><sup>cel</sup></i>	$\mathcal{L}_o^{cel}$	$\omega, \omega_L = 1$		68.1	83.7	88.9	81.1	85.3
<i>O<sub>6ω</sub><sup>cel</sup></i>	$\mathcal{L}_o^{cel}$	$\omega, \omega_L = 6$		79.8	91.0	94.0	86.6	89.2
<i>O<sub>6ω</sub><sup>cel</sup> + RR</i>	$\mathcal{L}_o^{cel}$	$\omega, \omega_L = 6$	✓	81.7	91.5	<b>94.5</b>	<b>87.9</b>	90.3
<i>O<sub>γ</sub><sup>cel</sup></i>	$\mathcal{L}_o^{cel}$	$\gamma$		66.6	83.5	89.1	78.3	82.8
<i>O<sub>γ</sub><sup>cel</sup> + RR</i>	$\mathcal{L}_o^{cel}$	$\gamma$	✓	73.2	86.8	91.3	82.6	86.2
<i>O<sup>cos</sup></i>	$\mathcal{L}_o^{cos}$			67.6	77.8	81.8	82.6	86.7
<i>O<sub>ω</sub><sup>cos</sup></i>	$\mathcal{L}_o^{cos}$	$\omega, \omega_L = 1$		72.7	84.1	87.2	84.5	87.9
<i>O<sub>6ω</sub><sup>cos</sup></i>	$\mathcal{L}_o^{cos}$	$\omega, \omega_L = 6$		80.2	90.6	93.4	86.3	88.9
<i>O<sub>6ω</sub><sup>cos</sup> + RR</i>	$\mathcal{L}_o^{cos}$	$\omega, \omega_L = 6$	✓	80.8	90.1	93.1	86.9	89.4
<i>O<sub>γ</sub><sup>cos</sup></i>	$\mathcal{L}_o^{cos}$	$\gamma$		81.1	90.2	93.1	87.1	89.7
<i>O<sub>γ</sub><sup>cos</sup> + RR</i>	$\mathcal{L}_o^{cos}$	$\gamma$	✓	80.7	90.3	93.1	86.9	89.5
<i>CO<sub>6ω</sub><sup>cel</sup> + RR</i>	$\mathcal{L}_c + \mathcal{L}_o^{cel}$	$\omega, \omega_L = 6$	✓	81.5	<b>91.8</b>	94.3	87.5	90.0
<i>CO<sub>γ</sub><sup>cos</sup></i>	$\mathcal{L}_c + \mathcal{L}_o^{cos}$	$\gamma$		<b>81.9</b>	90.8	93.2	<b>87.9</b>	<b>90.4</b>

*Degree of Centrality Weights* to balance the impact of *Branch* and *Leaf Event Nodes* on the overall loss. Thus, we increased the weight to  $\omega_L = 6$  as it approximately corresponds to the average number of *Branch Event Nodes* in all  $|\mathbb{V}_L| = 148$  *Subgraphs*.

Models trained with the ontology-driven loss functions  $\mathcal{L}_o^{cel}$  and  $\mathcal{L}_o^{cos}$  achieve similar results in their best setups. The models trained with  $\mathcal{L}_o^{cos}$  work well with both weighting schemes, while models optimized with  $\mathcal{L}_o^{cel}$  are better with the *Degree of Centrality Weight*  $\omega$ . We argue they are more tailored towards single-label classification tasks and benefit from the higher weights  $\omega_L = 6$  of *Leaf Event Nodes*. We achieved slightly better results when combining the classification-based ( $\mathcal{L}_c$ ) and *one* of the ontology-driven loss functions (models denoted as *CO*). The combination emphasizes the prediction of *Leaf Event Nodes* while still considering ontology information.

The best results for *top-1 accuracy*, *JSC*, and *CS* were achieved when combining the classification ( $\mathcal{L}_c$ ) and ontology-driven cosine loss term ( $\mathcal{L}_o^{cos}$ ) with *Distance Weights*  $\gamma$ . The cosine loss is, in general, more stable when training with and without redundancy removal (RR), which could indicate that it is more robust to changes in depth and size of the *Ontology*. Furthermore, it works well with the *Distance Weights*  $\gamma$ , which does not require an extra weight parameter  $\omega_L$  for *Leaf Event Nodes*.

#### 3.1.4.4 Experimental Results for Individual Event Types

The *top-1 accuracy* for a selection of *Event Nodes* and qualitative results of the  $CO_\gamma^{cos}$  model (notation according to Table 3.2) are provided in Figure 3.3 and Figure 3.4. The proposed approach achieves good results for the majority of event types. Misclassification can be typically explained by the visual similarity of the respective events. For example, images for *tornado*, *tsunami*, and *earthquake* are often captured after the actual event, and the consequences of these natural disasters can be visually similar, as illustrated in Figure 3.1 and Figure 3.4(f). It also turned out that classes such as *protest*, *earthquake*, and *explosion* are predicted very frequently because they depict visual concepts that are also part of other events. For instance, images of the event types *police brutality*, *vehicle fire*, and *economic crisis* are frequently classified as a *protest* since they depict typical scenes of riots or demonstrations (Figure 3.4(e)). The best results were achieved for sports-centric event types, which is not surprising as they are usually unambiguous. In general, the performance for expected (scheduled or regular) event types such as *election* and *sport* is better compared to unexpected or rare events. We assume the main reason is that journalists usually broadcast live coverage of expected events. At the same time, photos of crimes (e.g., *robbery*, *terrorist attack*) and *natural disasters* are rare and captured mainly by amateurs. Thus, it is more likely that web images depict the consequences rather than the actual event.

### 3 Information Extraction from Photos

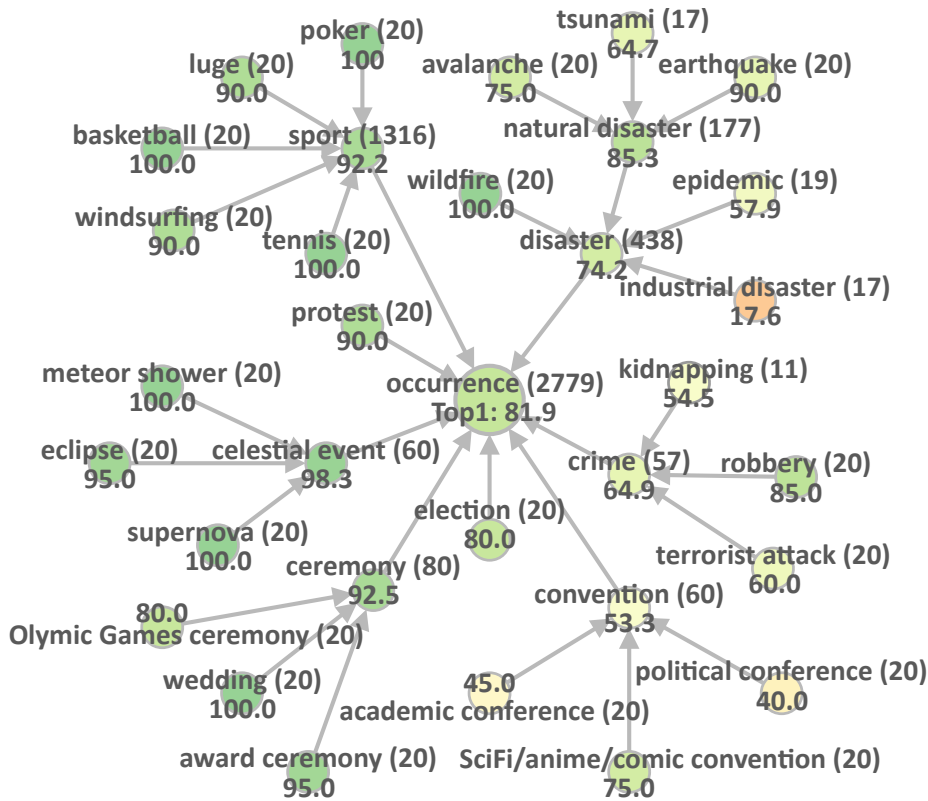


Figure 3.3: *Top-1 accuracy [%]* and number of images (in brackets) for a selection of *Event Nodes* on the *VisE-Bing* test set using the  $CO_{\gamma}^{cos}$  approach. The results correspond to the mean *top-1 accuracy* of all (also those that are not shown) related *Leaf Event Nodes*. The *Ontology* is simplified for better comprehensibility.

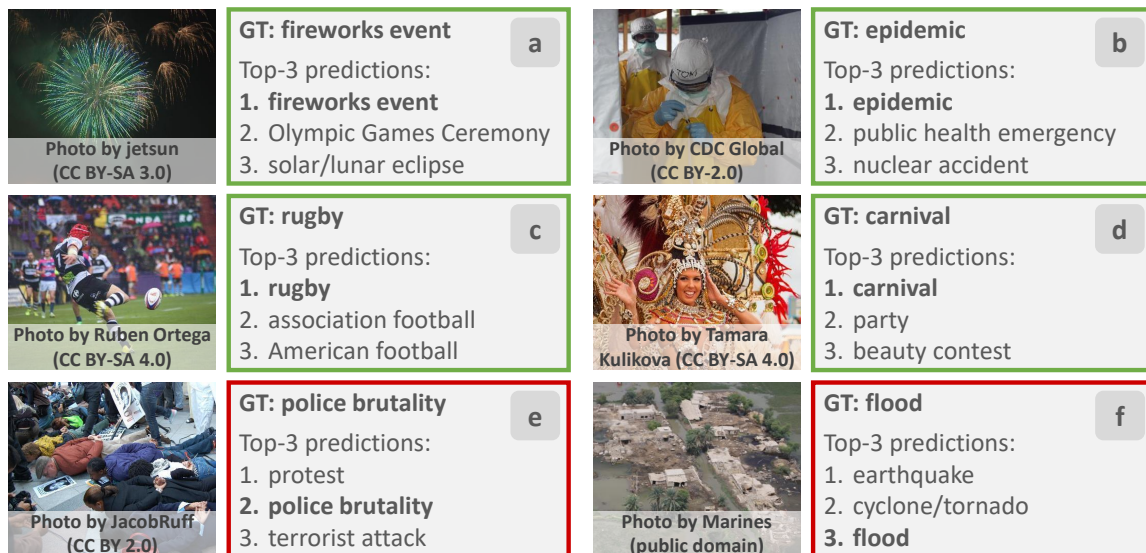


Figure 3.4: Correctly (a – d) and incorrectly (e, f) classified examples of the  $CO_{\gamma}^{cos}$  network model from the *VisE-Wiki* test set.

### 3.1.4.5 Comparisons on other Benchmarks

We considered several benchmarks, including the novel *VisE-Wiki* (Section 3.1.2.3) test dataset as well as WIDER [287], SocEID [8], and RED [8]. These benchmarks have different characteristics, which allows us to evaluate the ontology-driven approach in various setups. The WIDER dataset comprises 50,574 web images for 61 event types. To the best of our knowledge, it covers the most diverse event classes to date. Similar to our proposed dataset, the images are gathered from the Web and contain some irrelevant images, i.e., photos that do not depict the labeled event. Furthermore, some event classes also relate to actions (e.g., *Handshaking*) or occupations (e.g., *Surgeons*) rather than actual event types. The SocEID dataset consists of circa 37,000 images but contains only eight social event classes, while the RED dataset is comparatively small and contains around 7,000 images from 21 real-world events that were manually selected. We used the splits provided by the authors for the WIDER [287] and SocEID [8] datasets. For the RED dataset, we randomly used 70% of the images for training and the remaining 30% for testing, as suggested by Ahsan et al. [8]. The splits are provided on our *GitHub* page<sup>14</sup> for fair comparisons.

As the WIDER, SocEID, and RED datasets do not provide an *Ontology*, we have manually linked their classes to *Wikidata*, e.g., *soccer* in the WIDER dataset to the *Wikidata* item *association football* with the identifier *Q2736*, to define the set of *Leaf Event Nodes*. Then, we created the *Ontologies* according to Section 3.1.2.2. The *Ontologies* of the benchmark datasets are available on our *GitHub* page:

- *WIDER* ontology: <https://tibhannover.github.io/VisE/WIDER>
- *SocEID* ontology: <https://tibhannover.github.io/VisE/SocEID>
- *RED* ontology: <https://tibhannover.github.io/VisE/RED>

The models are trained with parameters similar to Section 3.1.4.1. Due to the smaller dataset sizes, the number of training iterations was reduced to 2,500 for the RED dataset and 10,000 for the SocEID and WIDER datasets. Cosine learning rate annealing [151] was applied from the beginning to decrease the learning rate from 0.01 to zero after the specified amount of iterations. The results for our approach and other comparable solutions from the related work [6, 8, 287] that use a single network model (and no ensemble) and the whole image as input are presented in Table 3.3.

The ontology-driven approaches (*CO*) clearly outperform the classification baseline (*C*) on the *VisE-Wiki*, WIDER, and RED test sets. As expected, the results on the SocEID dataset just slightly improved because less *Ontology* information is provided due to the lower number of eight classes, which leads to an ontology with fewer *Event Nodes* and relations. Results on *VisE-Wiki* are worse compared to *VisE-Bing* (reported in Table 3.2) since the test set is not manually annotated and contains unrelated or ambiguous images, particularly for rare event types such as *city fire*. The same applies to the WIDER dataset. Superior

Table 3.3: Results (numbers are multiplied by 100) on different benchmark datasets. While our results are superior on SocEID and RED, Ahsan et al. [8] achieved better results (77.9%) on WIDER using random splits (gray, not provided on request) also compared to other baselines by training an SVM on *AlexNet* embeddings, which is a similar approach for which Ahmad et al. [6] reported 41.9%. Their results for WIDER and RED are nearly identical, although WIDER contains more classes and is, in general, more challenging. We conclude that these results are not explainable and need to be verified in a reproducibility experiment. Model notations are according to Table 3.2.

Approach	VisE-Wiki 148 classes		WIDER [287] 61 classes		SocEID [8] 8 classes		RED [8] 21 classes	
	<i>Top-1</i>	<i>JSC</i>	<i>Top-1</i>	<i>JSC</i>	<i>Top-1</i>	<i>JSC</i>	<i>Top-1</i>	<i>JSC</i>
AlexNet [287]	—	—	38.5	—	—	—	—	—
AlexNet-fc7 [8]	—	—	77.9	—	86.4	—	77.9	—
WEBLY-fc7 [8]	—	—	77.9	—	83.7	—	79.4	—
Event conc. [8]	—	—	78.6	—	85.4	—	77.6	—
AlexNet [6]	—	—	41.9	—	—	—	—	—
ResNet-152 [6]	—	—	48.0	—	—	—	—	—
$C$	61.7	72.7	45.6	56.9	91.2	92.7	76.1	82.1
$CO_{6\omega}^{cel} + RR$	63.4	73.9	<b>51.0</b>	<b>61.6</b>	91.4	<b>92.9</b>	79.1	84.3
$CO_{\gamma}^{cos}$	<b>63.5</b>	<b>74.1</b>	49.7	60.3	<b>91.5</b>	<b>92.9</b>	<b>80.9</b>	<b>85.4</b>

results are achieved in comparison to similar solutions [6, 8, 287]. It is worth noting that the proposed ontology-driven approach can also be easily integrated into methods that use ensemble models [5, 6, 273] or additional image regions [86, 287].

### 3.1.5 Summary

In this section, we have presented a novel ontology, dataset, and ontology-driven deep learning approach to classify newsworthy event types in photos. A large number of events in conjunction with a knowledge base were leveraged to retrieve the *Visual Event Ontology* (VisE-O) that covers many possible real-world event types. The corresponding large-scale *Visual Event Classification Dataset* (VisE-D) with 570,540 photos allowed us to train powerful deep learning models and is, to the best of our knowledge, the most complete and diverse public dataset for event classification to date. We have proposed several baselines, including an ontology-driven deep learning approach that exploits event relations to integrate structured information from a knowledge graph. The results on several benchmarks have shown that the integration of structured information from an ontology can improve event classification. For this reason, we argue that the proposed approach provides discriminative semantic features that allow for the distinction between a majority of event types covered in the news.



In Chapter 4 of this thesis, we study the usefulness of the deep learning approach for the quantification of cross-modal event relations in news articles. It should be noted that the current ontology only distinguishes between event types (e.g., *association football*, *election*, *epidemic*) and, therefore, mainly provides semantic features. The classification of more fine-grained event classes or even concrete events (e.g., *FIFA World Cup Final 2014*, *2020 U.S. election*, *COVID-19 pandemic*) would allow for a more detailed analysis. In addition, concrete events are usually connected to one or multiple location(s), date(s), or time period(s) and can also provide geospatial and temporal information. Thus, the prediction of concrete events or fine-granular event types is another important research direction. Moreover, we plan to further explore strategies that leverage ontology information such as *Graph Convolutional Neural Network* (GCNN). Other interesting research directions are the combination of several knowledge bases and the investigation of semi-supervised approaches to learn from heterogeneous web sources that typically include also irrelevant images.

## 3.2 Geolocation Estimation of Photos

News articles often refer to specific locations to describe the geographic context. These locations can range from coarse entities such as continents and countries over specific urban environments (e.g., *cities, streets, buildings*, and *landmarks*) to natural environments (e.g., *mountains, seas, deserts*, and *forests*). In order to verify the cross-modal occurrences of these different types of locations between image and text, geographical information at a global scale needs to be extracted from the photos without any restrictions to certain environments.

As mentioned in Section 1.2, predicting the geographical location of photos taken all over the world without any prior knowledge is a very challenging task since they depict a huge amount of intra-class (e.g., different daytimes, objects, or camera settings) and extra-class variations (e.g., architecture, flora and fauna, or style of interior furnishings). Besides, the photos can be ambiguous or provide only very few visual clues about their respective capturing location. For these reasons, many approaches have simplified geolocation estimation and focused on photos depicting well-known landmarks and cities [20, 142, 226, 280, 302, 313] or natural areas like deserts or mountains [24, 225, 261]. Only a few proposals [89, 90, 229, 267, 279] treat the task at global scale without any prior assumptions. These approaches particularly benefit from the advancements in deep learning (Section 2.2.2) and the increasing number of publicly available large-scale image collections from platforms such as *Flickr*. Due to the complexity of the problem and the unbalanced distribution of photos taken from all over the world, methods based on CNNs [229, 267, 279] treat photo geolocation as a classification task by subdividing the Earth into geographical cells with a similar number of images. However, as also discussed in Section 1.2, the granularity of this partitioning is critical for the system performance and entails a trade-off problem [229]. A partitioning with more cells covering smaller geographic areas allows for more accurate predictions at a city (accuracy of about 25 km) or even street level (accuracy of about 1 km). However, it also reduces the number of training photos available for each cell, making models prone to overfitting. Models trained with fewer but larger cells, on the other hand, are less precise at these fine-granular levels but tend to generalize better and improve performance at coarser levels (e.g., country level with a geolocation accuracy of about 750 km).

Moreover, a single CNN consisting of tens of millions of parameters might struggle to memorize the visual appearance of locations around the world, according to Vo et al. [267]. In our opinion, one of the main reasons for this problem is the huge diversity in the photos caused by various environmental settings, which requires specific features to distinguish different locations. Referring to Figure 3.5, urban images mainly differ in, e.g., architecture, people, and specific objects like cars or street signs. On the contrary, natural scenes like forests or indoor scenarios are most likely defined by features encoding the flora and fauna or the style of the interior furnishings, respectively. Therefore, we argue that photo geolocal-

### 3.2 Geolocation Estimation of Photos

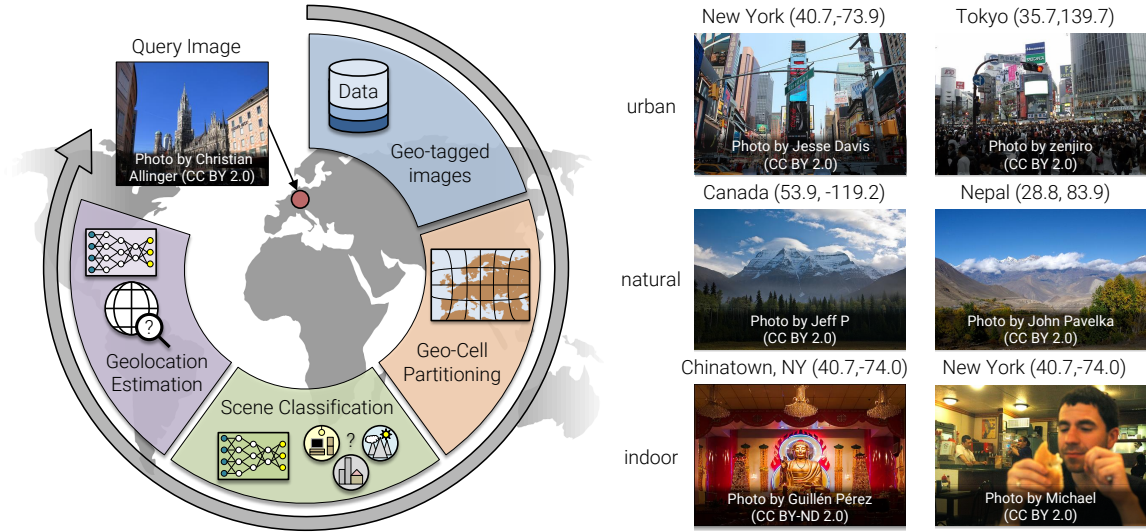


Figure 3.5: **Left:** Workflow of the proposed geolocation estimation approach. **Right:** Sample images of different locations for specific scene concepts.

ization can greatly benefit from contextual knowledge about the environmental scene since the diversity in the data space could be drastically reduced.

In this section, we present solutions for the problems mentioned above by (1) incorporating hierarchical knowledge at different geospatial resolutions in a multi-partitioning approach and (2) using information about the respective type of environmental settings (e.g., *indoor*, *natural*, and *urban*). We consider photo geolocalization as a classification task by subdividing the Earth into geographical cells with a balanced number of images (similar to *PlaNet* [279]). There are several contributions. We combine the outputs from all scales to exploit the hierarchical information of a CNN that is trained simultaneously with labels from multiple partitionings to encode local and global information. Furthermore, we suggest two strategies to include information about the respective scene type: (a) deep networks that are trained separately with images of distinctive scene categories, and (b) a multi-task network trained with both geographical and scene labels. This multi-task approach should enable the CNN to learn specific features for estimating the *Global Positioning System* (GPS) coordinate of images in different environmental surroundings. The workflow is illustrated in Figure 3.5.

To the best of our knowledge, this is the first approach that considers scene classification and exploits hierarchical (geo)information to improve unrestricted photo geolocalization. Furthermore, we have used a state-of-the-art CNN architecture, and our comprehensive experiments include an evaluation of the impact of different scene concepts. Experimental results on two different benchmarks demonstrate that our approach outperforms the state of the art without relying on image retrieval techniques (*Im2GPS* [89, 90, 267]) while using a significantly lower number of training images compared to *PlaNet* [279] and *CPlaNet* [229].

### 3 Information Extraction from Photos

The remainder of this section is organized as follows. In Section 3.2.1, related work on geolocation estimation is reviewed. The geographical cell partitioning of the Earth is explained in Section 3.2.2. The proposed framework to extract and leverage visual concepts of specific environmental settings and multiple partitionings of the Earth to estimate the GPS coordinates of images is introduced in Section 3.2.3. Experimental results on two different benchmarks are discussed in Section 3.2.4. A demonstrator of the system is presented in Section 3.2.5. Section 3.2.6 summarizes the work and outlines areas of future work.

#### 3.2.1 Related Work

Related work on image geolocalization can be roughly divided into two categories: (1) proposals that are restricted to specific environments or imagery, and (2) approaches at planet-scale without any restrictions. This section focuses on the second category since it is more closely related to the scope of this thesis. A more comprehensive survey is provided by Brejcha and Cadik [43].

Many proposals of the first category are introduced at city-scale resolution restricting the problem to specific cities or landmarks. The proposed methods mainly apply retrieval approaches to match a query image against a reference dataset [20, 80, 121, 122, 202, 226, 304]. Other approaches [23, 140, 200, 313] focus on landmark recognition and therefore either use a pre-defined set of landmarks or cluster a given photo collection in an unsupervised manner to retrieve the most interesting areas for geolocalization. Another line of works matches query images against 3D models of cities to enhance geolocation accuracy [53, 119, 141, 145, 206]. However, the underlying data collections of these methods are restricted to popular scenes and urban environments and therefore lack accuracy when predicting photos that do not have (many) instance matches. For this reason, some approaches additionally make use of satellite aerial imagery to enhance the geolocalization in sparsely covered regions [232, 266, 301, 302]. Solutions have been presented that match an aerial query image against a reference dataset containing satellite images in a wide baseline approach [14, 27, 280]. Some of these proposals [142, 143] address geolocation at planet-scale and extend the solution to rural areas. Only a minority of solutions have been suggested for natural geolocation estimation of images depicting beaches [46, 274], deserts [261], or mountains [24, 225]. Most of these approaches rely on extracted features from horizon lines to find the best matching locations [24, 225, 261].

All of the aforementioned proposals are restricted to well-covered regions of the Earth, specific imagery, or specific environmental scenes. Hays and Efros [89] have introduced *Im2GPS* as a first attempt for planet-scale geolocation estimation. They use a retrieval approach to match a given query image based on a combination of six global image descriptors to a reference dataset consisting of more than six million images with GPS coordinates. The authors extend *Im2GPS* [90] by incorporating information on specific geometrical classes like sky

and ground. Furthermore, they use an improved feature representation and retrieval technique. Weyand et al. [279] have been the first who applied deep learning to geolocalization. In their approach called *PlaNet*, the authors treat the task as a classification problem. For this reason, the Earth is subdivided into geographical cells with a similar number of images according to their GPS coordinates using a quad-tree approach. The resulting geographical cells are used as image labels to train a CNN. This approach noticeably outperformed *Im2GPS*, which encouraged Vo et al. [267] to learn a feature representation with a CNN to improve the *Im2GPS* framework. The features of a query photo extracted from the deep learning model are used to search for the (k)-nearest neighbors in the reference dataset based on a kernel density estimation. Moreover, a multi-partitioning approach is introduced to train photo-geolocation at different geospatial resolutions simultaneously.

The underlying quad-tree cell partitioning from *PlaNet* [279] that converts geolocation estimation to a classification problem introduces a critical trade-off problem. On the one hand, fewer but geographically larger cells are easier to distinguish, but they also lower the geospatial resolution of the outputs and consequently result in less accurate predictions. On the other hand, more but smaller cells that provide a good geospatial resolution are more difficult to distinguish and also lower the number of training examples per cell, making the model more prone to overfitting. To solve this issue, Seo et al. [229] proposed a combinatorial partitioning approach with multiple overlapping partitionings created based on the geographical and visual similarities of training images. It generates fine-grained output classes by intersecting overlapping coarse partitionings of the Earth. This allows estimating photo locations at a high geospatial resolution while maintaining a sufficient number of training examples per cell. Izbicki et al. [111] proposed a new loss function called *Mixture of von-Mises Fisher* (MvMF) that, unlike standard classification loss functions such as the *Cross-entropy Loss*, exploits the Earth’s spherical geometry and refines the geographical cell shapes in the partitioning. Similar to methods for face recognition (Section 3.4.1), Liu et al. [146] focus on a representation learning approach using a new *Stochastic Attraction and Repulsion* (SARE) loss function. They learn discriminative image representations by maximizing similarities among intra-place images while minimizing them among inter-place images. These representations are used to retrieve the most similar images in a reference database in order to determine the geolocation as proposed by *Im2GPS* [89, 90, 267].

### 3.2.2 Partitioning of the Earth Surface for Classification

We present a deep learning approach that aims to tackle the existing challenges by considering information about the environmental setting and exploiting hierarchical (geo)information using partitionings of the Earth at multiple geospatial resolutions. According to *PlaNet* [279], we treat the task as a classification problem by subdividing the Earth into geographical cells

### 3 Information Extraction from Photos

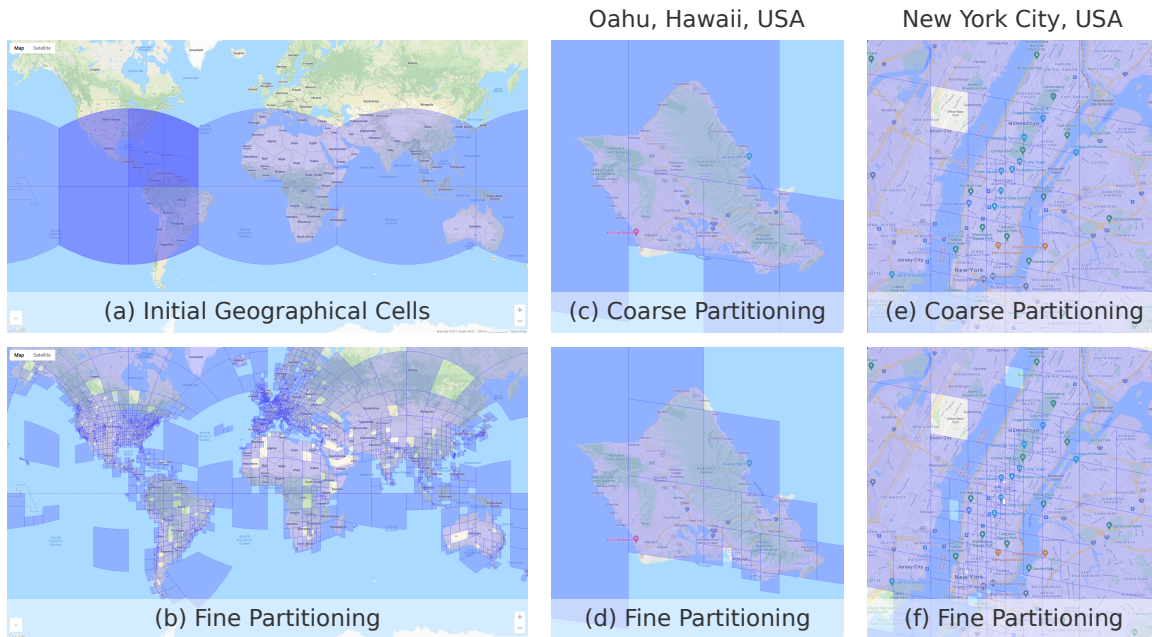


Figure 3.6: Partitioning of the Earth into geographical cells. (a) Four initial geographical cells (light blue) and exemplary cells in the second (blue) and third (dark blue) level of the quad-tree approach. Note that the top and bottom regions on the map represent the two remaining initial geographic cells of the Earth projected onto a cube with six sides and that the boundaries are spherical geodesics (i.e., straight lines on the sphere) that appear to be curved on the two-dimensional map. Coarse ( $\tau_{min} = 50; \tau_{max} = 5,000$ ) and fine partitioning ( $\tau_{min} = 50; \tau_{max} = 1,000$ ) of the Earth (b), the Hawaiian island *Oahu* (c, d), and *New York City* (e, f). Cell areas are smaller in regions that are photographed frequently, which allows for a more accurate geolocation estimation. The Screenshots are taken from: <https://s2.sidewalklabs.com/regioncoverer/>

that contain a similar number of images. The *S2 geometry library*<sup>18</sup> is used to generate a set of non-overlapping geographical cells  $\mathbb{C}$  used as classes. The Earth’s surface is projected on an enclosing cube with six sides representing the initial cells (illustrated in Figure 3.6). An adaptive hierarchical subdivision based on the GPS coordinates of the training images is applied [279], where each cell is the node of a quad-tree. Starting at the root nodes, the respective quad-tree is subdivided recursively until all cells contain a maximum of  $\tau_{max}$  images. Afterward, all resulting cells with fewer than  $\tau_{min}$  photos are discarded because they likely cover areas like poles or oceans, which are hard to distinguish.

This approach has several advantages compared to a subdivision of the Earth into cells with roughly equal areas. On the one hand, an adaptive subdivision prevents dataset biases and allows for classes with a similar number of images. On the other hand, fine cells are generated in photographically well-covered areas, allowing more accurate prediction of image locations that most likely represent regions of interest, such as landmarks or cities.

<sup>18</sup><https://code.google.com/archive/p/s2-geometry-library/>

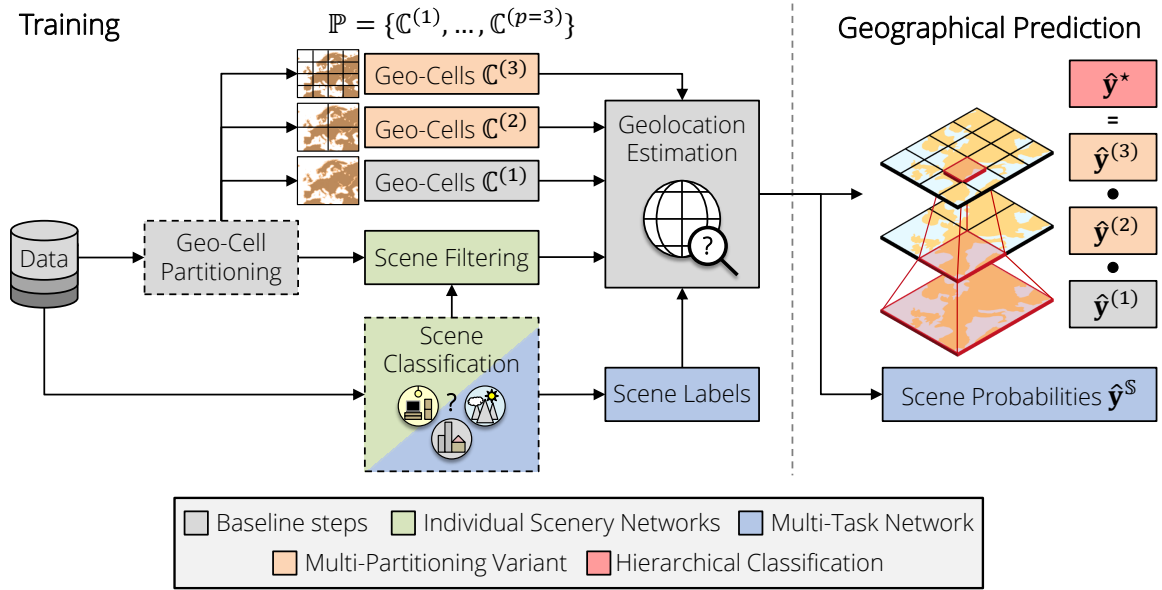


Figure 3.7: Pipeline of the proposed geolocation estimation frameworks. Gray: Baseline steps that are part of every network. Additional steps for various network setups are visualized in different colors. Steps in dashed rectangles are applied to all images before the training process takes place.

### 3.2.3 Geolocation Estimation using Contextual Information

In contrast to previous work, we exploit contextual information of the environmental scenario solely using the visual content of a given photo to improve the localization accuracy. Therefore, we predict the scene probabilities of all images based on the 365 categories of the *Places365* dataset [315] (Section 3.2.3.1). Several approaches that are aimed at integrating the extracted information about the given type of scene and multiple geographical cell partitionings are introduced in Section 3.2.3.2. Finally, we explain how the proposed approaches are applied to estimate the GPS coordinates of images based on the predicted geo-cell probabilities (Section 3.2.3.3). In this context, we introduce our hierarchical approach to combine the results of multiple geospatial resolutions. An overview of the proposed framework is presented in Figure 3.7.

#### 3.2.3.1 Environmental Scene Classification

A *ResNet* model [93] with 152 layers<sup>19</sup> provided by the authors of the *Places365* dataset [315], which is a subset of the *Places2* database, is applied to calculate the scene probabilities of a given image. The model has been trained with more than 16 million images from 365 different place categories. This scene classification fits nicely with our approach since the resulting classifier already distinguishes images that depict specific environments.

<sup>19</sup> *ResNet-152* model trained with *Caffe* on *Places365* [315]: <https://github.com/CSAILVision/places365>

### 3 Information Extraction from Photos

We consider three different sets  $\{\mathbb{S}_3, \mathbb{S}_{16}, \mathbb{S}_{365}\}$  of scene categories with different levels of granularity using the scene hierarchy<sup>20</sup> provided by the dataset. First, we compute the scene probabilities  $\hat{\mathbf{y}}^{\mathbb{S}_{365}}$  for all 365 scenes in  $\mathbb{S}_{365}$  using the classification output of the CNN. The scene hierarchy allows a mapping in order to additionally extract the probabilities  $\hat{\mathbf{y}}^{\mathbb{S}_{16}}$  and  $\hat{\mathbf{y}}^{\mathbb{S}_3}$  of the sets  $\mathbb{S}_{16}$  and  $\mathbb{S}_3$  containing 16 and three superordinate scene categories, respectively. For this purpose, we add the probabilities of all classes assigned to the same superordinate scene category to generate the corresponding probabilities. However, some place categories such as *barn* visually overlap and are consequently allocated to multiple superordinate categories, in this case to "*outdoor, natural*" and "*outdoor, man-made*" that are part of the scene set  $\mathbb{S}_3$ . For this reason, we first divide the probability of these classes by the number of assigned categories to maintain the normalization  $\sum_{i=1}^{|\mathbb{S}|} y_i^{\mathbb{S}} = 1$  with  $\mathbb{S} \in \{\mathbb{S}_3, \mathbb{S}_{16}\}$ . Please note that we use the terms *natural* for "*outdoor, natural*" and *urban* for "*outdoor, man-made*" in the remainder of this thesis.

#### 3.2.3.2 Geolocation Estimation

In this section, several approaches based on CNNs for unrestricted planet-scale geolocalization are introduced. First, we present a baseline approach that is trained without using scene information and multiple geographical partitionings. In the following, we describe how the information for different geospatial resolutions as well as environmental concepts are integrated into the training process. In this context, two different approaches using environmental scene labels are proposed. An overview is provided in Figure 3.7.

**Baseline:** We first introduce a baseline system that does not rely on information about the environmental setting and different geospatial resolutions to evaluate the impact of the suggested approaches for geolocalization. Therefore, we generate a single geo-cell partitioning  $\mathbb{C}$  according to Section 3.2.2. For classification, we add a fully-connected layer with a softmax activation function (Equation (2.5), page 32) on top of the "avg pool" layer (Table 2.2, page 37) of the *ResNet* architecture [93], where the number of output neurons corresponds to the number of geo-cells  $|\mathbb{C}|$ . The cross-entropy geolocalization loss  $\mathcal{L}_{geo}^{single}$  based on the ground-truth cell label encoded in a one-hot vector  $\mathbf{y} = \langle y_1, y_2, \dots, y_{|\mathbb{C}|} \rangle \in \{0, 1\}^{|\mathbb{C}|}$  and the predicted probability distribution  $\hat{\mathbf{y}} = \langle \hat{y}_1, \hat{y}_2, \dots, \hat{y}_{|\mathbb{C}|} \rangle \in \mathbb{R}^{|\mathbb{C}|}$  is minimized during training:

$$\mathcal{L}_{geo}^{single}(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_{i=1}^{|\mathbb{C}|} y_i \log \hat{y}_i \quad (3.10)$$

**Multi-Partitioning Variant:** We propose to simultaneously learn geolocation estimation at multiple geospatial resolutions, as also suggested by Vo et al. [267]. In contrast to the

<sup>20</sup>*Places365* scene hierarchy: <http://places2.csail.mit.edu/download.html>



baseline approach, we add a fully-connected layer with a softmax activation function for the geographical cells of all  $p$  partitionings  $\mathbb{P} = \{\mathbb{C}^{(1)}, \mathbb{C}^{(2)}, \dots, \mathbb{C}^{(p)}\}$ . The multi-partitioning classification loss  $\mathcal{L}_{geo}^{multi}$  is calculated using the mean of the loss values  $L_{geo}^{single}$  for every partitioning. Let  $\mathbf{y}^{(j)} \in \{0, 1\}^{|\mathbb{C}^{(j)}|}$  and  $\hat{\mathbf{y}}^{(j)} \in \mathbb{R}^{|\mathbb{C}^{(j)}|}$  be the one-hot encoded ground truth vector and the predicted probability distribution for geo-cell partitioning  $\mathbb{C}^{(j)}$ , then the multi-partitioning classification loss can be defined as follows:

$$\mathcal{L}_{geo}^{multi} = \frac{1}{p} \sum_{j=1}^p \mathcal{L}_{geo}^{single}(\mathbf{y}^{(j)}, \hat{\mathbf{y}}^{(j)}) \quad (3.11)$$

As a consequence, the CNN can learn geographical features at different scales resulting in a more discriminative classifier. However, in contrast to Vo et al. [267], we further exploit the hierarchical knowledge for the final prediction. The details are presented in Section 3.2.3.3.

**Individual Scenery Networks (ISNs):** Given a set of scenes  $\mathbb{S} \in \{\mathbb{S}_3, \mathbb{S}_{16}, \mathbb{S}_{365}\}$ , *Individual Scenery Networks* (ISNs) for every scene  $s \in \mathbb{S}$  are trained in a first attempt to incorporate context information about the environmental setting for photo geolocation. For each photograph, we extract the scene probabilities  $\hat{\mathbf{y}}^{\mathbb{S}}$  using the scene classification approach presented in Section 3.2.3.1. During the training, every image with a probability  $\hat{y}_i^{\mathbb{S}}$  for a scene  $s \in \mathbb{S}$  with index  $i$  greater than a threshold of  $\hat{y}_i^{\mathbb{S}} > \tau_{\mathbb{S}}$  is used as input for the respective ISN $_s$ . It can be optimized using a single or multiple geographical cell partitionings. Following this approach offers the advantage that the network is solely trained on photos depicting specific environmental scenarios. It significantly reduces the diversity in the underlying data space and enables the network to learn more specific features. On the contrary, it is necessary to train individual models for each scene concept, which is hard to manage if the number of different concepts  $|\mathbb{S}|$  is large. For this reason, we suggest fine-tuning a model, which was initially trained without scene restriction, with images of the respective environmental category.

**Multi-task Network (MTN):** Since the aforementioned method for geolocation estimation may become infeasible for many different environmental concepts, we aim for a more practicable approach using a network that treats photo geolocation and scene recognition as a multi-task problem. We simultaneously train two classifiers for these complementary tasks in order to encourage the network to distinguish between images of different environmental scenes. Adding another (complementary) task has proven to be efficient in improving the results of the main task [34, 112, 216, 310]. More specifically, an additional fully-connected layer with a softmax activation function on top of the "avg pool" layer (Table 2.2, page 37) of the *ResNet* architecture [93] is used. Given a set of scenes  $\mathbb{S} \in \{\mathbb{S}_3, \mathbb{S}_{16}, \mathbb{S}_{365}\}$ , the number of output neurons of this layer corresponds to the amount of scene categories  $|\mathbb{S}|$ . The weights of all other layers in the network are shared. In addition, the scene loss  $\mathcal{L}_{scene}$

### 3 Information Extraction from Photos

based on the ground-truth one-hot vector  $\mathbf{y}^{\mathbb{S}} \in \{0, 1\}^{|\mathbb{S}|}$  and the scene probabilities  $\hat{\mathbf{y}}^{\mathbb{S}} \in \mathbb{R}^{|\mathbb{S}|}$  is minimized using the cross-entropy loss. The total loss of the *Multi-Task Network* (MTN) is defined by the sum of the geographical and scene loss  $\mathcal{L}_{total} = \mathcal{L}_{scene} + \mathcal{L}_{geo}$ , where the geographical loss  $\mathcal{L}_{geo}$  can refer to both the single cross-entropy loss  $\mathcal{L}_{geo}^{single}$  or multi-partitioning cross-entropy loss  $\mathcal{L}_{geo}^{multi}$ .

#### 3.2.3.3 Predicting Geolocations using Hierarchical Spatial Information

In order to estimate the GPS coordinate from the classification output, we apply the trained models from Section 3.2.3.2 on three evenly sampled crops of a given query image according to its orientation. Afterward, the mean of the resulting class probabilities of each crop is calculated. Please note that an additional step for testing is necessary for the ISNs. In this case, the scene label  $s \in \mathbb{S}$  with the maximum probability is predicted to feed the image into the respective  $ISN_s$  for geolocalization.

**Standard Geo-Classification:** Without relying on hierarchical information, we solely use the probabilities  $\hat{\mathbf{y}}^{(i)}$  of one given geo-cell partitioning  $\mathbb{C}^{(i)}$ . In this regard, we assign the class label with the maximum probability to predict the geographical cell. Applying the multi-partitioning approach in Section 3.2.3.2, we can obtain  $p$  class probabilities at different geospatial resolutions. In our opinion, the probabilities at all scales should be exploited to enhance the geolocalization and to combine the capabilities of all partitionings.

**Hierarchical Geo-Classification:** A fixed threshold parameter  $\tau_{min}$  for the adaptive partitioning of the Earth explained in Section 3.2.2 is applied to ensure that every geographical cell in the finest representation can be uniquely connected to a larger parent area in an upper level. As a result, we are able to generate a geographical hierarchy from the partitionings of varying granularity. Inspired by the hierarchical object classification approach from *YOLO9000* [210], we multiply the respective probabilities at each level of the hierarchy. Consequently, the prediction for the finest subdivision can be refined by incorporating the knowledge of coarser representations.

**Class2GPS:** Depending on the predicted class, we extract the GPS coordinates of the given query image. In contrast to Weyand et al. [279], we use the mean location of all training images assigned to the predicted cell instead of the geographical center. This approach is more precise for regions containing an interesting area where the majority of photos are taken. Imagine a geographical cell centered around an ocean and a city that is located at the cell boundary. In this example, the error using the geographical center would be very high, even if it is clear that the photo was most likely taken in the city.

### 3.2.4 Experimental Setup & Results

In the following section, the experimental setup, system parameters, and implementation details are introduced (Section 3.2.4.1 to Section 3.2.4.5). Subsequently, the different system parameters are evaluated in detail (Section 3.2.4.6 to Section 3.2.4.8), and a comparison to the state of the art is conducted (Section 3.2.4.9).

#### 3.2.4.1 Training Data

A subset of the *Yahoo Flickr Creative Commons 100 Million dataset* (YFCC100M) [258] is used as input data for our approach. This subset was introduced for the *MediaEval Placing Task 2016* (MP16) [131] and includes around five million geo-tagged images<sup>21</sup> from *Flickr* without any restrictions. The dataset contains ambiguous photos of, e.g., indoor environments, food, and humans for which the location is difficult to predict. Like Vo et al. [267], we exclude images from the same authors as in the test datasets, which we use for evaluation. A *ResNet* model [93] is used, which has been pre-trained on the ILSVRC 2012 dataset [58, 218] to avoid duplicate images by comparing the resulting feature vectors from the last pooling layer. Overall, our training dataset consists of 4,723,695 images.

#### 3.2.4.2 Parameters for the Adaptive Partitioning using S2 Cells

As explained in Section 3.2.3.3, we choose a constant value of  $\tau_{min} = 50$  (according to *PlaNet* [279]) as the minimum threshold for the adaptive partitioning to enable the hierarchical classification approach. Our goal is to train the geolocation at multiple geospatial resolutions. Therefore, the following maximum thresholds  $\tau_{max} \in \{1,000; 2,000; 5,000\}$  are used. We select these thresholds because the MP16 dataset has approximately 16 times fewer images than *PlaNet* [279], and we aim to produce around  $\sqrt{16}$  fewer classes (*PlaNet* has 26,263 cells) at the middle representation. Since we want to show how fine and coarse representations can be efficiently combined, the other thresholds are specified to produce circa two times more and fewer classes than the middle representation. The resulting number of classes  $|\mathbb{C}|$  for different partitionings are shown in Table 3.4.

#### 3.2.4.3 Scene Classification Parameters

The performance of the environmental scene classification (Section 3.2.3.1) is evaluated on the *Places365* validation dataset [315] containing 36,500 images (100 for each scene). In Table 3.5, results for the different scene hierarchy levels are reported. The quality of the scene classification is crucial for the ISNs presented in Section 3.2.3.2 because it defines the underlying data space. Since the top-1 accuracy of 91.50 % already provides a good basis, we

<sup>21</sup>Available at: <http://multimedia-commons.s3-website-us-west-2.amazonaws.com>

Table 3.4: Number of geographical cells  $|\mathcal{C}|$  for Earth partitionings with different thresholds  $\tau_{min}$  and  $\tau_{max}$ 

Partitioning $\mathcal{C}$	$\tau_{min}$	$\tau_{max}$	$ \mathcal{C} $
coarse	50	5,000	3,298
middle	50	2,000	7,202
fine	50	1,000	12,893

Table 3.5: Top-1 and Top-5 accuracy on the *Places365* validation set [315] using scenes of different granularity

Scene Set	Top-1	Top-5
$\mathcal{S}_3$	91.5 %	—
$\mathcal{S}_{16}$	72.1 %	97.1 %
$\mathcal{S}_{365}$	45.7 %	77.3 %

focus on a set of three scene concepts  $\mathcal{S}_3 = \{indoor, natural, urban\}$ . Furthermore, this limits the number of ISNs to a feasible number of three concepts. We suggest applying a small threshold of  $\tau_S = 0.3$ . Admittedly, this selection is somewhat arbitrary, but we intend to use images with similar scene probabilities as input for each ISN. This parameter selection can be especially useful for images depicting rural areas because they share visual information like architecture as well as flora and fauna that are beneficial for both environmental categories, *urban* and *natural*. The scene filtering yields a total of around 1.80 million, 1.42 million, and 2.34 million training images for the concepts *indoor*, *natural*, and *urban*, respectively.

### 3.2.4.4 Network Training

The proposed approaches are trained using a *ResNet* architecture [93] with 101 convolutional layers (Table 2.2, page 37). The weights are initialized by a model pre-trained on the ILSVRC 2012 dataset [58, 218]. The data is augmented by randomly selecting an area covering at least 70% of the image with an aspect ratio  $R$  between  $3/4 \leq R \leq 4/3$  to avoid overfitting. Furthermore, the input images are randomly flipped and subsequently cropped to  $224 \times 224$  pixels. We use the SGD optimizer with an initial learning rate of 0.01, a momentum of 0.9, and a weight decay of 0.0001. The learning rate is exponentially lowered by a factor of 0.5 after every five training epochs. We initially train the networks for 15 epochs and a batch size of 128. We validate the CNNs on 25,600 images of the YFCC100M dataset [258].

As described in Section 3.2.3.2, it could be beneficial to fine-tune the ISNs based on a model that was initially trained without scene restriction. For a fair comparison, all models are therefore fine-tuned for five epochs or until the loss on the validation set converges. In this regard, the initial learning rate is decreased to 0.001. Finally, the best model on the validation set is used for conducting the experiments. The implementation is realized using the *TensorFlow* library [1] in *Python*. The trained models and all necessary data to reproduce our results are available at: <https://github.com/TIBHannover/GeoEstimation>

Table 3.6: Notations of the geolocalization approaches.  $T$  denotes whether the network was trained with a single/lone (L) or multiple (M) partition(s).  $\mathbb{C} \in \{c, m, f\}$  indicates which cell partition (coarse ( $c$ ), middle ( $m$ ), fine ( $f$ )) is used for classification. If  $\mathbb{C}$  is denoted with a star (\*), the hierarchical classification is utilized.

Notation	Description
$base(T, \mathbb{C})$	<i>Baseline</i> trained without scene information
ISNs ( $T, \mathbb{C}, \mathbb{S}_3$ )	<i>Individual Scenery Networks</i> using the scene set $\mathbb{S}_3$
MTN ( $T, \mathbb{C}, \mathbb{S}$ )	<i>Multi-Task Network</i> using a scene set $\mathbb{S} \in \{\mathbb{S}_3, \mathbb{S}_{16}, \mathbb{S}_{365}\}$

### 3.2.4.5 Test Setup

We evaluate our approaches on two public benchmarks datasets for geolocation estimation. The *Im2GPS* test dataset [89] contains 237 photos, where 5% depict specific tourist sites and the remaining are only recognizable in a generic sense. Because this benchmark is very small, Vo et al. [267] introduced a new dataset called *Im2GPS3k* that contains 3,000 images from *Im2GPS* (2,997 images are provided with a GPS tag and used for testing). The *Great Circle Distance* (GCD) between the predicted and ground-truth image location is calculated for evaluation. As suggested by Hays and Efros [89], we report the geolocalization accuracy as the percentage of test images predicted within a certain distance to the ground-truth location. The notations of the proposed approaches are presented in Table 3.6. The most significant results using the suggested multi-partitioning and scene concepts for geolocalization, as well as a comparison to the state-of-the-art methods, are given in the related sections. A complete list of results is provided in Appendix A.2.

### 3.2.4.6 Evaluating the Multi-Partitioning Approach

The results for the baseline and the multi-partitioning approach are displayed in Figure 3.8. Surprisingly, no significant improvement using multiple partitionings can be observed for the *Im2GPS* test dataset. However, it is clearly visible that the results, especially for the *fine* partitioning, have improved for the *Im2GPS3k* dataset, which is more representative due to its larger size. The results demonstrate that the network is able to incorporate features at different geospatial resolutions and use this knowledge to learn a more discriminative classifier. A similar observation was made in the latest *Im2GPS* approach [267]. However, by exploiting the hierarchical knowledge at different geospatial resolutions, the localization accuracy can be indeed further increased. Figure 3.9 shows that the geolocation of the photo is predicted with a higher accuracy using the coarse and middle partitioning compared to the finest representation. Unfortunately, these coarser partitionings do not fully exploit the network capabilities in terms of geospatial resolution. However, the use of hierarchical information can refine the prediction at the finest resolution leading to a more accurate estimation of the photo’s GPS position. Referring to the supplemental material and the

### 3 Information Extraction from Photos

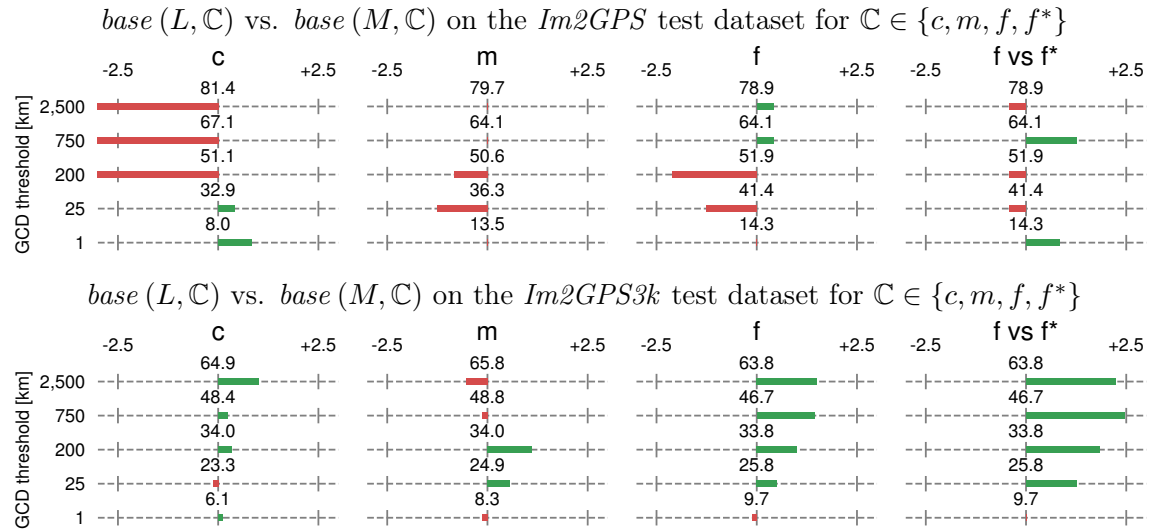


Figure 3.8: Comparison of the geolocation approaches trained with and without multiple partitions for geo-cell partitionings  $\mathbb{C}$  of varying granularity on the *Im2GPS* (top) and *Im2GPS3k* (bottom) test set. First mentioned approach  $base(L, \mathbb{C})$  is used as reference and its accuracy [%] is denoted in the middle of the  $x$ -axis.

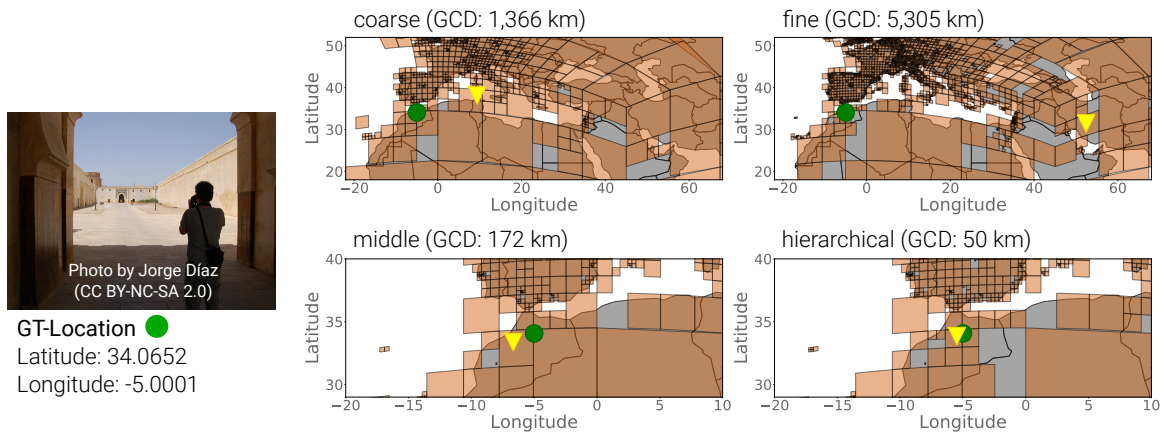


Figure 3.9: Prediction (yellow triangle) using outputs of different partitionings as well as the hierarchical result compared to the ground truth (GT) location (green circle).

next section, it is worth mentioning that the ISNs greatly benefit from the knowledge at multiple geospatial resolutions. The results on both datasets improve drastically while using the multi-partitioning approach.

#### 3.2.4.7 Evaluating the Individual Scenery Networks

We apply the scene classifier introduced in Section 3.2.3.1 to extract the scene labels for all test images to evaluate the performance for specific environmental settings. The resulting number of images for every scene is presented in Table 3.7. Due to the low number of im-

Table 3.7: Number of images for the *Im2GPS* and *Im2GPS3k* geolocalization benchmarks depicting different scene concepts in  $\mathbb{S}_3$

Scenes	Im2GPS	Im2GPS3k
all	237	2,997
indoor	19	545
natural	80	845
urban	138	1,607

Table 3.8: Top-1 and Top-5 scene classification accuracies on the validation set of the *Places365* benchmark [315] for different MTNs

Network	Top-1	Top-5
MTN ( $L, f, \mathbb{S}_3$ )	92.0 %	—
MTN ( $L, f, \mathbb{S}_{16}$ )	71.7 %	97.5 %
MTN ( $L, f, \mathbb{S}_{365}$ )	46.0 %	76.5 %

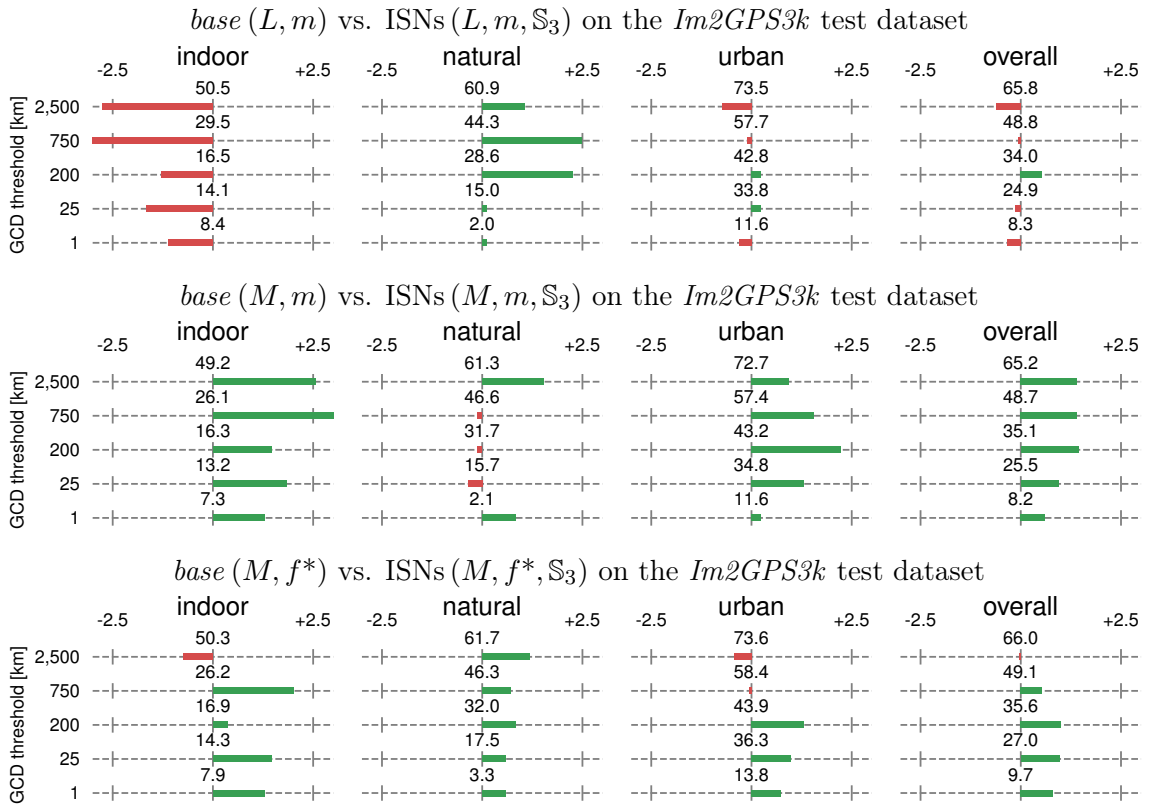


Figure 3.10: Comparison of the *Individual Scenery Networks* (ISNs) to the baseline approaches for different environmental scene concepts. First mentioned approach is used as reference and its accuracy [%] is denoted in the middle of the  $x$ -axis.

ages in the *Im2GPS* test dataset, we analyze the performance of the ISNs on the *Im2GPS3k* dataset. However, referring to Table 3.10 and the supplemental material, similar observations can be made for *Im2GPS*. The geolocation results do not improve when restricting a single-partitioning network to specific concepts (Figure 3.10). On the other hand, using a multi-partitioning approach with scene restrictions noticeably improves geolocation estimation, particularly for *urban* and *indoor* photos. One possible explanation is that the intra-class variation for coarser subdivisions with more images in larger areas is reduced.

### 3 Information Extraction from Photos

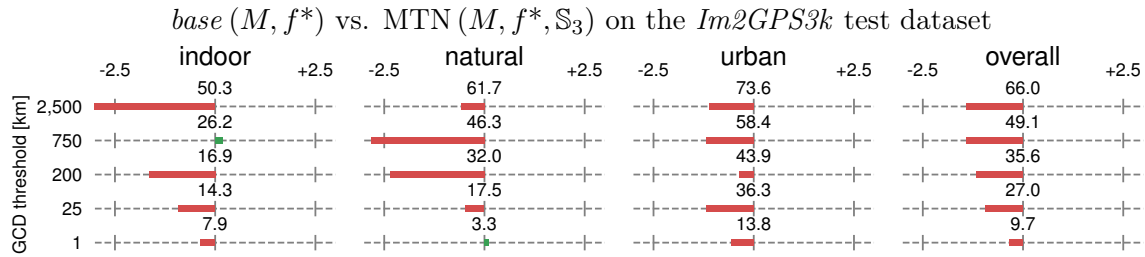


Figure 3.11: Comparison of the *Multi-Task Network* (MTN) to the baseline approach for different scene concepts. First mentioned approach is used as reference and its accuracy [%] is denoted in the middle of the  $x$ -axis.

Therefore, the network can learn specific features for the respective scene concept. The best results are achieved for urban images, which is intuitive since they often contain relevant cues for geolocation. It is also not surprising that the performance of indoor photos is the lowest among all scene concepts since the images can be ambiguous. For this reason, Weyand et al. [279]) (*PlaNet*) even disregard *indoor* images for geolocation estimation. Despite only 1.42 million *natural* images are available to cover the huge diversity of very different scenes like *beaches*, *mountains*, and *glaciers*, we were able to improve the performance for this concept. We argue that the respective ISN mainly benefits from the hierarchical information because it enables the encoding of more global features such as different climatic zones. Overall, the results for nearly all GCD thresholds and scene categories show that geolocation estimation benefits from training with specific scene concepts.

#### 3.2.4.8 Evaluating the Multi-Task Network

We investigate the performance of the MTN regarding environmental scene classification (Table 3.8) and geolocation estimation (Figure 3.11). Although the results show that the MTN is able to learn both tasks simultaneously, geolocalization does not benefit from learning an additional task no matter which model configuration we analyze. The results indicate that reducing the diversity in the underlying data space is more important for the estimation of GPS coordinates of photos. Regarding environmental scene classification, similar results are achieved compared to the provided model of the *Places365* dataset (Table 3.5).

#### 3.2.4.9 Comparison to the State of the Art

In this section, we compare our proposed solutions to state-of-the-art baselines from the literature. However, these baselines use different network architectures and (number of) training images (*Im2GPS* [267] also use additional retrieval datasets) that significantly impact the performance. For a fair comparison, we have summarized the most important parameters in Table 3.9. Regarding the number of training images, our approaches are comparable to



Table 3.9: Parameters used by approaches for geolocation estimation, including the number of training images ( $I_T$ ), reference images ( $I_R$ ), and CNN architecture. The results of the respective CNN architectures for object recognition on ILSVRC-2012 serve as a reference to evaluate the overall network performance and are taken from: <https://pytorch.org/vision/stable/models.html>

Method	$I_T$	$I_R$	CNN	ILSVRC 2012	
				Top-1	Top-5
Im2GPS [267]					
• [L] 7011C	6M [89]	–	VGG-16 [235]	71.6 %	63.7 %
• [L] kNN, $\sigma = 4$	6M [89]	6M [89]	VGG-16 [235]	71.6 %	90.4 %
• ... 28m database	6M [89]	22M [258]	VGG-16 [235]	71.6 %	90.4 %
PlaNet (6.2M) [279]	6.2M [279]	–	Inception v3 [251]	77.3 %	93.5 %
PlaNet (91M) [279]	91M [279]	–	Inception v3 [251]	77.3 %	93.5 %
PlaNet (rep. by [229])	30.3M [229]	–	Inception v3 [251]	77.3 %	93.5 %
CPlaNet (best) [229]	30.3M [229]	–	Inception v3 [251]	77.3 %	93.5 %
MvMF (best) [111]	6M [230]	–	Wide ResNet-50 [300]	78.5 %	94.1 %
Our models	4.7M [131]	–	ResNet-101 [93]	77.4 %	93.5 %

Im2GPS [L] 7011C [267], PlaNet (6.2M) [279], and MvMF [111]. The remaining PlaNet variants and CPlaNet [229] can be considered as equivalent at a larger scale. Unlike Im2GPS, which uses a less powerful VGG-16 [235], the baselines use CNN architectures with comparable performances for object recognition on the ILSVRC 2012 dataset [58, 218].

The results for geolocation estimation on the *Im2GPS* and *Im2GPS3k* test datasets are presented in Table 3.10. It is evident that our proposed solutions outperform the current state-of-the-art methods. Interestingly, our baseline approach *base* ( $L, m$ ) already significantly outperforms its equivalents, i.e., Im2GPS [L] 7011C and PlaNet (6.2M), which are trained with a similar number of images and classes using a single partitioning of the Earth. For this reason, we investigate the influence of the *ResNet* architecture [93] used in our approach. Therefore, we train the system *base* ( $L, m$ ) with the VGG-16 network [235] used in Im2GPS [267]. The result is denoted with *base-vgg* ( $L, m$ ) and shows that the main improvement is explained by the more powerful *ResNet* architecture. As in PlaNet and Im2GPS [L] 7011C, the system *base-vgg<sub>c</sub>* ( $L, m$ ) uses the geographical center of the predicted cell as location instead of the mean GPS coordinate of all images that we suggested in Section 3.2.3.3. Using the mean coordinate already improves the performance on *street* and *city* levels noticeably. As described in the previous sections, the geolocalization can be further increased by training the CNN with multiple partitionings and exploiting the hierarchical knowledge at all geospatial resolutions. Best results were obtained when combining the ISNs with the hierarchical approach trained with images of a specific environmental scene concept. Overall, we achieved state-of-the-art results even compared to baselines that use significantly more training images or additional retrieval datasets on both benchmarks.

### 3 Information Extraction from Photos

Table 3.10: Results on the *Im2GPS* (top) and *Im2GPS3k* (bottom) test datasets. The fraction of photos localized within various distances to the actual photo location using the *Great Circle Distance* (GCD) are reported [%]. According to Vo et al. [267], Human\* performance was averaged from 30 *Amazon Mechanical Turk* workers over 940 trials and might not be directly comparable.

Im2GPS Test Dataset – 237 Photos					
Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2,500 km
Human* [267]			3.8 %	13.9 %	39.3 %
Im2GPS [267]					
• [L] 7011C	6.8 %	21.9 %	34.6 %	49.4 %	63.7 %
• [L] kNN, $\sigma = 4$	12.2 %	33.3 %	44.3 %	57.4 %	71.3 %
• ... 28m database	14.4 %	33.3 %	47.7 %	61.6 %	73.4 %
PlaNet (6.2M) [279]	6.3 %	18.1 %	30.0 %	45.6 %	65.8 %
PlaNet (91M) [279]	8.4 %	24.5 %	37.6 %	53.6 %	71.3 %
PlaNet (reprod. by [229])	11.0 %	31.2 %	37.6 %	64.6 %	<b>81.9 %</b>
CPlaNet (best) [229]	16.5 %	37.1 %	46.4 %	62.0 %	78.5 %
MvMF (best) [111]	8.4 %	32.6 %	39.4 %	57.2 %	80.2 %
<i>base-vgg<sub>c</sub>(L, m)</i>	7.6 %	22.8 %	35.0 %	50.6 %	66.7 %
<i>base-vgg(L, m)</i>	8.9 %	26.6 %	36.7 %	50.6 %	65.8 %
<i>base(L, m)</i>	13.5 %	36.3 %	50.6 %	64.1 %	79.7 %
<i>base(M, m)</i>	13.5 %	35.0 %	49.8 %	64.1 %	79.7 %
<i>base(M, f*)</i>	15.2 %	40.9 %	51.5 %	65.4 %	78.5 %
ISNs ( $M, f^*, \mathbb{S}_3$ )	<b>16.9 %</b>	<b>43.0 %</b>	<b>51.9 %</b>	<b>66.7 %</b>	80.2 %

Im2GPS3k Test Dataset – 2,997 Photos					
Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2,500 km
Im2GPS [267]					
• [L] 7011C	4.0 %	14.8 %	21.4 %	32.6 %	52.4 %
• [M] 7011C	3.7 %	14.2 %	21.3 %	33.5 %	52.7 %
• kNN, $\sigma = 4$	7.2 %	19.4 %	26.9 %	38.9 %	55.9 %
PlaNet (reprod. by [229])	8.5 %	24.8 %	34.3 %	48.4 %	64.6 %
CPlaNet (best) [229]	10.2 %	26.5 %	34.6 %	48.6 %	64.6 %
<i>base-vgg<sub>c</sub>(L, m)</i>	4.2 %	14.6 %	22.2 %	34.4 %	54.2 %
<i>base-vgg(L, m)</i>	4.8 %	16.5 %	22.6 %	34.5 %	54.4 %
<i>base(L, m)</i>	8.3 %	24.9 %	34.0 %	48.8 %	65.8 %
<i>base(M, m)</i>	8.2 %	25.5 %	35.1 %	48.7 %	65.2 %
<i>base(M, f*)</i>	9.7 %	27.0 %	35.6 %	49.2 %	<b>66.0 %</b>
ISNs ( $M, f^*, \mathbb{S}_3$ )	<b>10.5 %</b>	<b>28.0 %</b>	<b>36.6 %</b>	<b>49.7 %</b>	<b>66.0 %</b>

### 3.2.5 Demonstrator

We have developed a web demonstrator of our system to make our approach accessible to a broad audience. Figure 3.12 shows a screenshot of the demonstrator. It is publicly available as a lab service of the *Leibniz Information Centre for Science and Technology* (TIB) under the following link: <https://labs.tib.eu/geoestimation>

The demonstrator allows users to either upload a photo to estimate its geolocation or compete with the proposed approach for geolocation estimation on a subset of images from *Im2GPS* [89] with *Creative Commons* licenses. Once an image has been selected, the user can place a marker on the world map to guess the geolocation of the photo. After pressing the button "*Guess Location*", the results of the proposed ISNs ( $M, f^*, S_3$ ) are computed and presented on the world map. In this context, markers of different colors (see legend below the world map) are used to indicate the prediction of the user and the proposed approach as well as to indicate the ground-truth location. The distances of the user and model prediction to

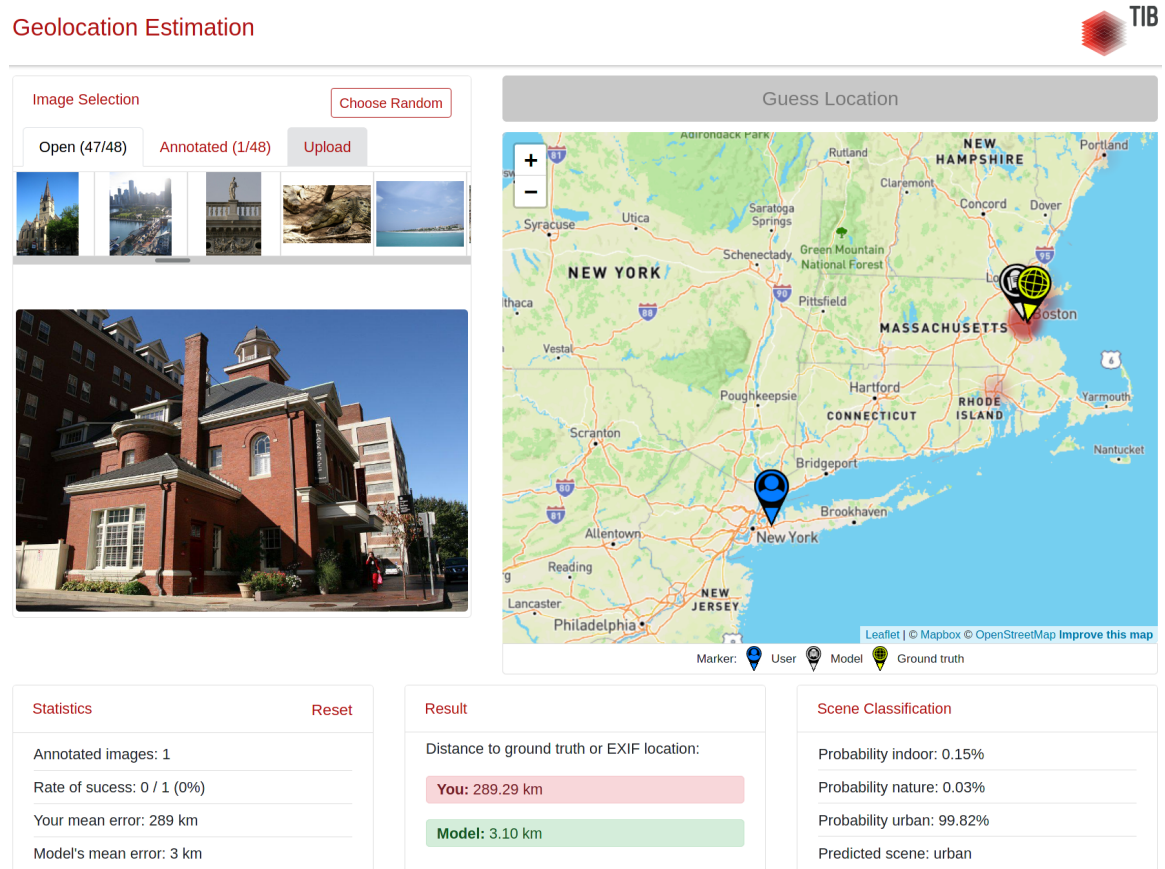


Figure 3.12: Screenshot of the demonstrator for geolocation estimation. The location of the photo shown on the left side is estimated by the user (blue marker) and the proposed ISNs ( $M, f^*, S_3$ ) (gray marker). The GCD to the ground truth location (yellow marker) is shown in the "Result" box.

the ground-truth location are presented once the calculation has finished. Besides, the world map is overlaid with a geographical heat map, where the regions colored in red represent the most likely locations of the photo according to the proposed model. This visualization allows some degree of explainability of the model’s output. In addition, the class activation map of the predicted geographical cell is computed using the approach from Zhou et al. [314] to visualize the photo parts that contributed most to the decision (not shown in the screenshot for better visibility of the photo).

As mentioned in Section 1.6, the geolocation approach has attracted attention in the media and was or will be presented as an exhibit at several exhibitions.

#### 3.2.6 Summary

In this section, several deep learning approaches for planet-scale photo geolocation estimation have been presented. As suggested by previous solutions, we have treated geolocation estimation as a classification task by subdividing the Earth into geographical cells. A multi-partitioning approach has been proposed that combines hierarchical information from various geospatial resolutions. Moreover, scene information has been exploited to incorporate context about the environmental setting (e.g., *indoor*, *natural*, and *urban*) into a CNN model. Experimental results on two benchmarks have demonstrated that our framework improves the state of the art in estimating the GPS coordinates of photos. We have shown that the CNN can learn specific features for the different environmental settings and geospatial resolutions, yielding a better classifier for geolocation. Best results were achieved when the hierarchical approach was combined with scene classification. In contrast to previous work, the proposed framework neither relies on an exemplary dataset for image retrieval [89, 90, 267] nor a training dataset consisting of several tens of millions of images [229, 279].

Overall, the experiments demonstrated that deep learning approaches can accurately estimate photo locations, especially for outdoor photos, when given enough and unambiguous geographical cues. For this reason, we conclude that CNNs provide rich geographic features that can be used to measure the cross-modal consistency of locations in news articles and other multimedia content. A corresponding system and study are provided in Chapter 4.

In the future, we intend to investigate how other contextual information like cultural, climatic, or economic aspects as well as from specific objects, daytimes, and seasons can be exploited to improve geolocation. Moreover, we plan to leverage information from geographic databases such as *OpenStreetMap* (<https://www.openstreetmap.org>) to create partitions based on territorial borders (e.g., countries, cities), natural geological boundaries (e.g., rivers, mountains), or man-made barriers (e.g., roads, railways, or buildings) that better reflect location entities mentioned in the media [257].

### 3.3 Date Estimation of Historical Photos

Date estimation of photos is an interesting and challenging task with many applications. For example, semantic search and multimedia retrieval can be leveraged by historians, archivists, or even to sort (digitized) personal photo collections chronologically. Moreover, news articles on the World Wide Web typically refer to specific events, points in time, or time periods to provide temporal information to the reader. Consequently, temporal information offers essential cues to quantify relations between image and text. However, as mentioned in Section 1.2, only a few approaches [72, 77, 189, 221] have been presented that aim to estimate the capturing time of (historical) images, but they simplify the task of date estimation. Some approaches focus on specific concepts like cities [227], cars [133], persons [77, 221], or historical documents [94, 139] and therefore cannot learn the temporal differences of the wide variety of motifs. Other approaches use color features [72, 154, 189] to model the developments in color photography. Thus, they rely on historical color photographs, which were uncommon before the 1970s. Alternatively, solutions on timestamp verification [52, 117, 137, 187, 220] check the month and daytime information claimed in the metadata of a photograph. However, these solutions cannot predict in which year a photo was taken and use meteorological features [52], sun azimuth angles [117, 137], or satellite images [187, 220] for verification, which limits them to outdoor photos. These restrictions limit potential applications, and due to the absence of large-scale training datasets for date estimation, previous work has not yet exploited the potential of deep learning models to solve this task.

In this section, we introduce a novel dataset called *Date Estimation in the Wild* (Section 3.3.2). Unlike previous datasets, it contains more than one million photos (black-and-white and color) from *Flickr* captured in the period from 1930 to 1999. As shown in Figure 3.13, the dataset covers a broad range of domains, e.g., city scenes, family photos, nature, and historical events. Two baseline approaches are proposed using a deep CNN (in this case, a *GoogLeNet* architecture [251]), treating the task of dating images as a classification and regression problem, respectively. Experimental results have shown the feasibility of the suggested approaches, which are superior to annotations of untrained humans.

The remainder of this section is structured as follows. Related work on date estimation is reviewed in Section 3.3.1. The proposed *Date Estimation in the Wild* dataset is presented in Section 3.3.2. In Section 3.3.3, two baseline approaches are proposed. Experimental results on date estimation and a comparison to human annotations are provided in Section 3.3.4. Section 3.3.5 contains a summary and outlines potential directions of future work.

#### 3.3.1 Related Work

In this section, we briefly review related work on date estimation. As previously mentioned, solutions for metadata verification use meteorological features (e.g., temperature, humidity,

### 3 Information Extraction from Photos



Figure 3.13: Some example images from the *Date Estimation in the Wild* dataset

or weather conditions) [52], the sun azimuth angle estimated from shadow angles as well as the appearance of the sky [117, 137], or visual attributes from satellite images [187, 220] to verify the claimed timestamp of a photo. However, these approaches are limited to outdoor photos and only verify the month and time of day information without considering the year the photo was taken. The first work that deals with dating historical images from different decades has been introduced by Schindler et al. [227]. The authors present an approach to sort a collection of city-scape images temporally by reconstructing the 3d-world, requiring many overlapping images of the same location. Lee et al. [133] identify style-sensitive groups of patches for cars and street view images in order to model stylistic differences across time and space. He et al. [94] and Li et al. [139] address the task of estimating the age of historical documents. While He et al. [94] explore contour and stroke fragments, Li et al. [139] apply CNNs in combination with optical character recognition. Ginosar et al. [77] and Salem et al. [221] model the differences in human appearance and clothing style in order to predict the year of photos that depict people in school yearbooks.

More closely related to the scope of this thesis, Palermo et al. [189] suggest an approach to automatically estimate the age of historical color photos without restrictions to specific concepts. They combine different color descriptors to model the historical color film processes. The results on the proposed dataset, which contains 1,375 images from 1930 to 1980, are further improved by Fernando et al. [72] by including color derivatives and angles. Martin et al. [154] treat date estimation as a binary task by deciding whether an image is older or newer than a reference image. Ahmed et al. [7] have proposed a similar approach and trained a CNN that estimates the acquisition date by predicting a timeline based on images

with known temporal order from the same source (i.e., digital camera). Ashida et al. [21] have used our *Date Estimation in the Wild Dataset* presented in this section and proposed a rank-consistent ordinal classification scheme. In addition, they showed the effectiveness of combining global image features from the entire input image and object-specific temporal features from frequent concepts, such as cars or people, for date estimation. This direction was also suggested as future work (Section 3.3.5) of our previous publication [179].

Overall, the works proposed for this research area are limited to dating historical color photos [72, 154, 189], require pictures from the same source [7], or simplify the task of date estimation to photographs of specific concepts [77, 94, 133, 139, 221, 227], which limits potential applications.

### 3.3.2 Date Estimation in the Wild Dataset

To overcome the existing limitations, we introduce the *Date Estimation in the Wild* dataset that is neither restricted to specific concepts nor to historical color photographs. The *Flickr Application Programming Interface* (API) was used to download photos for each year from 1930 to 1999. We have observed that many historical images are supplemented with time information, either in the title or in the related tags and descriptions. Therefore, we used the specified year as an additional query term to reduce the number of irrelevant images that were not captured in the queried year. The only kind of filtering that we applied was restricting the web search to photos. As a consequence, the dataset contains some irrelevant photos, for example, close-ups of plants or animals as well as historical documents. In order to avoid a bias towards more recent images, the maximum number of images per year was limited to 25,000. Finally, the dataset consists of 1,029,710 images with a high diversity of concepts, as shown in Figure 3.13. Information about the granularity  $g \in \{0, 4, 6, 8\}$  according to the *Flickr* annotation of the date entry is stored as well. The distribution of images per year and the related granularity of dates are visualized in Figure 3.14.

A maximum number of 75 *unique images* for 1930 to 1954 and 150 *unique images* for the remaining years were extracted to obtain reliable validation and test sets that match the dataset distribution. A *unique image* is defined as an image with a date granularity of  $g = 0$  (Y-m-d H:i:s) or  $g = 4$  (Y-m), for which no visual near-duplicates exist in the entire dataset. The near-duplicates are detected by comparing the visual features extracted from the last pooling layer of a *GoogLeNet* pre-trained on the ILSVRC 2012 dataset [58, 218] of all images using the *Euclidean distance*. We consider all images with a *Euclidean distance* of 15 or lower to another image as near-duplicates. Subsequently, 8,495 *unique images* were extracted for the validation set, and another 16 per year were selected manually to obtain a reliable test dataset comprising 1,120 images. The remaining 1,020,095 images constitute the training set. The dataset<sup>22</sup> is available at <https://doi.org/10.22000/0001abcde>.

<sup>22</sup>Images or links (depending on the copyright status) and metadata are provided.

### 3 Information Extraction from Photos

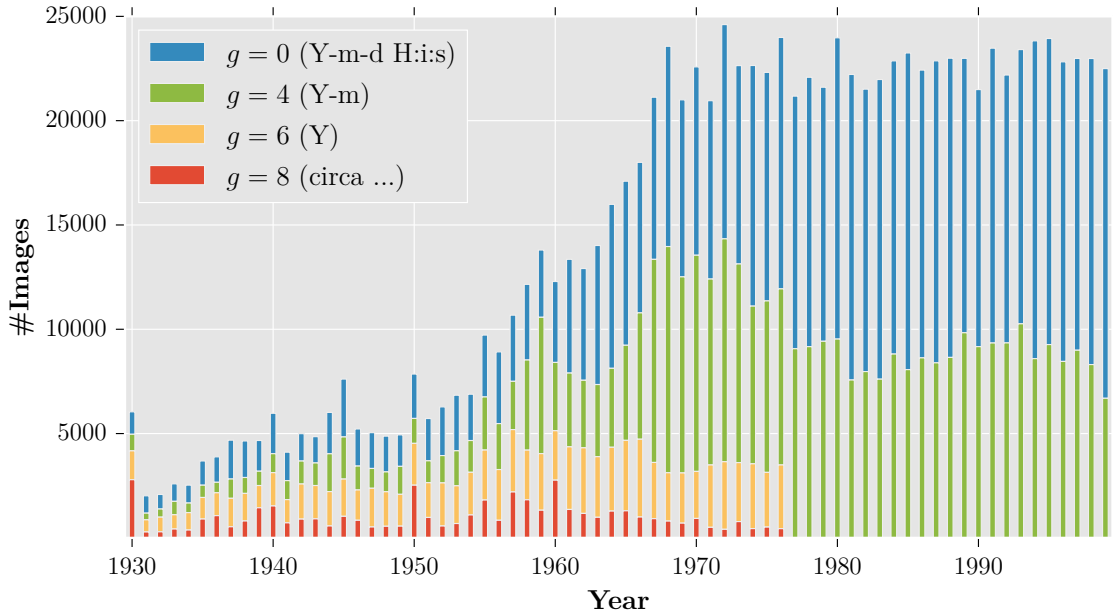


Figure 3.14: Number of crawled images and the accuracy (granularity  $g$ ) of the provided timestamps for each year in the *Date Estimation in the Wild* dataset

### 3.3.3 Deep Learning Models for Date Estimation

Two baseline approaches are realized that train a CNN architecture by treating image date estimation as a regression (Section 3.3.3.1) or classification problem (Section 3.3.3.2).

#### 3.3.3.1 Regression Model

Intuitively, date estimation is a regression task where the network should predict the ground truth acquisition year of a photograph. Therefore, a fully-connected layer with a single neuron (no activation function applied) is added on top of a CNN architecture that outputs an estimated acquisition year. As the dataset contains images captured between 1930 and 1999, the bias of this neuron is initialized with 1975, which corresponds to the middle year of this time period. During training, the *Euclidean distance* between the predicted  $\hat{a}$  and ground-truth acquisition year  $a$  is minimized to learn the network weights:

$$\mathcal{L}_{euc} = \sqrt{(a - \hat{a})^2} \quad (3.12)$$

However, as shown in Figure 3.14, the dataset contains fewer photographs for some acquisition years, particularly before 1960. Besides, images of a particular year can relate to specific historical events, which can introduce a dataset bias leading to less accurate models. To alleviate this problem, we propose a classification model that uses a larger number of



images for different time periods. Besides, it has been proven that transforming a regression problem into a classification problem can yield better results, e.g., for depth [75] or geolocation estimation [279].

### 3.3.3.2 Classification Model

As mentioned in the last section, CNNs benefit from a larger number of images per class or, in our case, acquisition year to learn appropriate models and prevent possible bias. However, the proposed *Date Estimation in the Wild* dataset lacks images for the 1930s to 1960s (Figure 3.14). For this reason, we assign the image acquisition years to 5-year periods (1930 – 1934, 1935 – 1939, . . . , 1995 – 1999) to treat date estimation as a classification problem with lower complexity and more samples per class while still maintaining a good temporal resolution. As a result,  $c = 14$  time periods are extracted that are used as classes. Based on the ground truth acquisition year  $a$ , we generate a one-hot encoded vector  $\mathbf{y} = \langle y_1, y_2, \dots, y_{14} \rangle \in \{0, 1\}^{14}$  that indicates the corresponding time period. A fully-connected layer with *softmax* activation function (Equation (2.5)) and  $c = 14$  neurons is added on top of a CNN architecture to calculate a probability vector  $\hat{\mathbf{y}} = \langle \hat{y}_1, \hat{y}_2, \dots, \hat{y}_{14} \rangle$  with each entry  $\hat{y}_i \in [0, 1]$  and the same dimension. During training, the cross-entropy loss between the ground-truth  $\mathbf{y}$  and predicted vector  $\hat{\mathbf{y}}$  is optimized:

$$\mathcal{L} = - \sum_{i=1}^c y_i \log \hat{y}_i \quad (3.13)$$

**Inference:** Unlike the regression-based approach, the classification model does not directly predict an acquisition year  $\hat{a}$ . To estimate the acquisition year, the averaged network outputs  $\hat{\mathbf{y}} = \langle \hat{y}_1, \hat{y}_2, \dots, \hat{y}_{14} \rangle \in \mathbb{R}^c$  for  $c = 14$  classes are interpolated by:

$$\hat{a} = 1930 + \left\lfloor 0.5 + \frac{1999 - 1930}{c - 1} \cdot \sum_{i=1}^c (i - 1) \cdot \hat{y}_i \right\rfloor, \quad \text{with } \sum_{i=1}^c \hat{y}_i = 1. \quad (3.14)$$

### 3.3.4 Experimental Setup & Results

**Network parameters:** A *GoogLeNet* [251] (explained in Section 2.2.4), which was initialized with pre-trained weights learned on the ILSVRC 2012 dataset [58, 218], was used as network architecture for the proposed approaches. We randomly selected 128 images per batch for training, which were scaled by the ratio  $256/\min(w, h)$ , where  $w$  and  $h$  correspond to the width and height of the image. For data augmentation, the training images were horizontally flipped and cropped randomly to  $224 \times 224 \times 3$  pixels to match the input resolution of the *GoogLeNet*. The SGD optimizer was employed using 1 million iterations with a momentum of 0.9. For the classification approach, a learning rate of 0.001 was used. For the

### 3 Information Extraction from Photos

regression-based approaches, the base learning rate was reduced to 0.0001 to stabilize the training. The learning rates were decreased by a factor of 2 every 100,000 iterations. The weights of the fully-connected classification layer (see Table 2.1, page 34) were re-initialized, and their corresponding learning rates were multiplied by 10 for faster training.

While testing, the images are scaled by the ratio  $224/\min(w, h)$ , and three evenly sampled regions (crops) of size  $224 \times 224$  pixels depending on the images' orientations are passed to the trained model. The averaged network outputs of the three crops are used as the prediction  $\hat{a}$  for regression or  $\hat{y}$  for classification.

**Metrics:** In the experiments, the trained *GoogLeNet* models were applied to the test set. In contrast to Palermo et al. [189], we do not report the classification accuracy for predicting the correct 5-year period. For example, imagine that the ground-truth date of an image is 1989, and the model predicts the class 1990 – 1994. Although the difference is possibly only one year, the prediction would be false in this case. For this reason, we argue that the absolute mean error (ME) and the number of images with an absolute estimation error of at most  $n$  years ( $EE_n$ ) are more meaningful for evaluation.

**Human Performance:** We conducted a user study to compare our approach to human performances. Seven untrained annotators of different ages (from 26 to 58) were asked to label all 1,120 test images and to take a break after each batch of 100 images. The average human performance and the results of our baseline approaches are displayed in Table 3.11.

**Results:** The results clearly show the feasibility of our baselines, outperforming human annotations in nearly all periods and reducing the mean error by more than three years on the entire test set. Another observation is that there is a correlation between the number of images and the results for each 5-year period. For this reason, an increased mean error for images between 1930 to 1964 is noticeable. Besides, the potential error can be higher for classes at the interval boundaries (1930 and 1999), which explains the slightly worse results from 1990 to 1999. A similar observation can be made for human annotations since they are more familiar with images, TV material, and their own experiences starting from 1960. Interestingly, the human error is noticeably lower for images covering the period from 1940 and 1944, which frequently show scenes from *World War II*.

Despite the problem caused by the interval bounds of the entire time period, which affects the interpolation step, the classification approach provides slightly better results than the regression approach. This improvement is attributed to the easier task of minimizing the classification loss of  $c = 14$  classes compared to minimizing the *Euclidean loss* for regression. Overall, the results confirm related studies in other computer vision areas such as depth [75] or geolocation estimation [279] that have shown the superiority of classification approaches.

Table 3.11: Absolute mean error (ME; lower is better) in years and number of images estimated with an absolute estimation error [%] of at most  $n$  years ( $EE_n$ ; higher is better). Results are reported for human annotators and the *GoogLeNet* classification (cls) and regression (reg) baselines on the *Date Estimation in the Wild* test dataset regarding different 5-year periods from 1930 to 1999.

Period	human performance				<i>GoogLeNet</i> cls				<i>GoogLeNet</i> reg			
	ME	EE <sub>0</sub>	EE <sub>5</sub>	EE <sub>10</sub>	ME	EE <sub>0</sub>	EE <sub>5</sub>	EE <sub>10</sub>	ME	EE <sub>0</sub>	EE <sub>5</sub>	EE <sub>10</sub>
30 – 34	15.7	3.0	24.8	40.7	15.0	0.0	5.0	37.5	14.4	0.0	7.5	41.3
35 – 39	12.2	2.7	34.1	53.2	11.1	2.5	23.8	52.5	10.7	3.8	26.3	58.8
40 – 44	9.6	4.1	43.2	66.6	8.8	2.5	40.0	67.5	9.1	7.5	42.5	66.3
45 – 49	11.7	3.9	31.1	54.3	8.2	6.3	51.3	71.3	8.5	3.8	43.8	70.0
50 – 54	12.2	2.5	29.6	49.8	7.5	3.8	47.5	77.5	7.3	2.5	52.5	73.8
55 – 59	13.3	1.4	27.1	49.5	6.1	6.3	60.0	86.3	7.0	7.5	50.0	77.5
60 – 64	13.6	1.4	24.1	43.0	7.3	5.0	51.3	73.8	7.2	1.3	47.5	75.0
65 – 69	12.5	2.7	24.6	46.4	5.4	12.5	63.8	82.5	6.0	1.3	52.5	83.8
70 – 74	10.5	4.8	33.2	55.9	5.6	3.8	58.8	85.0	5.4	8.8	61.3	85.0
75 – 79	9.4	4.1	37.9	62.1	4.7	8.8	71.3	90.0	5.0	7.5	63.8	90.0
80 – 84	7.5	5.2	45.5	76.1	4.4	8.8	62.5	95.0	4.5	6.3	61.3	93.8
85 – 89	7.6	5.0	49.6	77.3	4.8	10.0	71.3	83.8	4.9	8.8	68.8	90.0
90 – 94	7.5	5.9	51.3	76.1	5.6	5.0	66.3	85.0	5.7	6.3	61.3	83.8
95 – 99	9.4	6.1	39.5	62.9	7.5	11.3	52.5	75.0	8.7	1.3	36.3	73.8
overall	10.9	3.8	35.4	58.1	<b>7.3</b>	6.2	<b>51.8</b>	<b>75.9</b>	7.5	<b>4.7</b>	48.2	<b>75.9</b>

### 3.3.5 Summary

This section has introduced a novel dataset entitled *Date Estimation in the Wild* to foster research regarding the challenging task of photo date estimation. In contrast to previous work, the dataset is neither restricted to color photographs nor specific concepts but includes photos with a broad range of motifs for the period from 1930 to 1999. In a first attempt to tackle this challenging problem, we have proposed two approaches relying on deep CNNs to predict an image’s acquisition year, considering the task as a classification as well as a regression problem. Both approaches achieved a mean error of fewer than eight years and were superior to annotations of untrained humans.

Although the results have shown the superiority of deep learning approaches compared to human annotators for the task of (historical) date estimation, the average error is quite significant, and the exact date information (day, month, time of the day) cannot be extracted with the proposed system. The system is also limited to photographs that were taken until the year 1999. The low temporal resolution and year restriction might limit the potential system applications, including the prediction of temporal information in contemporary news articles. However, the approach can be used to assess temporal information in news articles and multimedia content that cover historical events.

### 3 *Information Extraction from Photos*

In the future, it is planned to exploit different specific classifiers for frequent concepts, such as persons or cars, to enhance the performance of our systems further. Besides, the estimation of more recent acquisition years (after 1999) and fine-granular image dates (e.g., calendar day and daytime) will be investigated. However, this probably requires geographic, economic, and cultural features since images taken at the same date can look quite different depending on their location. For example, photos taken in high-populated cities such as *New York City* require different temporal cues for date estimation compared to images captured in rural areas of, e.g., *South America*.

## 3.4 Person Identification in News Articles of the Internet Archive

Person entities are an important aspect of news and other multimodal documents. Appropriate deep learning approaches are necessary to extract rich facial feature vectors to verify the cross-modal occurrence of persons in joint placements of image and text. This section presents a novel approach for person identification in image data of news articles extracted from the *Internet Archive* ([www.archive.org](http://www.archive.org)) to evaluate the capabilities of deep learning models for facial feature extraction.

The World Wide Web contains billions of web pages and related multimedia content. These web pages include valuable information for many academic and non-academic applications. Therefore, the *Internet Archive* and national (digital) libraries have been capturing the (multimedia) web pages with time-stamped snapshots in large-scale archives since the mid-1990s. The *Internet Archive* serves as a playground for researchers and analysts in different domains such as *digital humanities*, *politics*, *economics*, and *entertainment*. One of the main challenges is to make the available unstructured data, which is rarely enriched with appropriate metadata, accessible and explorable by the users. For this reason, it is necessary to develop (semi)-automatic content analysis systems to extract metadata that can be subsequently used for semantic search and information visualization in order to provide users with relevant information about a given topic.

As discussed in Section 2.3.3, many tools like *AIDA* [98], *Agdistis* [263], *Babelfy* [167, 181], and *BLINK* [281] for *Named Entity Disambiguation* (NED) have been introduced that can be used to generate meta information from textual web content in order to, e.g., track entities and their relations in web archives. Although these tools achieve good results, online news articles are often complemented with photos. These photos potentially show additional entities that might not be mentioned in the text. Furthermore, possible ambiguities could be resolved using the photo content. Thus, photo and text are complementary, and their combination can serve as a basis for a more complete and robust detection of entities. While some approaches aim to find efficient solutions for person identification and object classification in large-scale datasets [71, 169, 170, 269], approaches that exploit image or video data in the *Internet Archive* are rare [171, 172].

This section presents an approach (illustrated in Figure 3.15) to identify persons and their joint occurrences in the image content of news articles in large-scale web archives such as the *Internet Archive*. It can be used by researchers and analysts to find and explore media coverage and relations of persons of interest in a given domain, e.g., *politics*, *sports*, or *entertainment*. We address various problems, such as how to automatically define which entities should be considered in such a system and how they can be automatically verified in large web collections. Example images are crawled for every entity using an image search

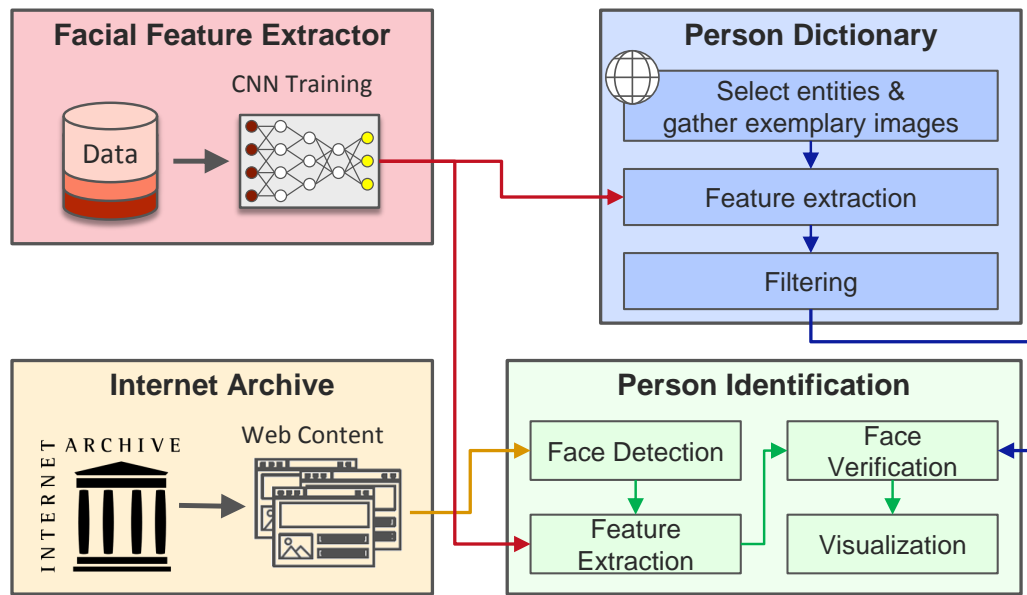


Figure 3.15: Workflow of the proposed person identification approach. A CNN serves as a generalized feature descriptor to obtain and filter features for a specified dictionary of persons. These features are used to identify persons as well as their relations in news articles gathered from the *Internet Archive*.

engine like *Google Images*. Due to irrelevant photos in the retrieved data of this web-based approach, we investigate three strategies to improve the quality of the example dataset. A state-of-the-art CNN is used to learn a robust feature representation to describe the facial characteristics of the entities. Based on the example dataset, the deep learning model is applied to identify the selected entities in the image content of news articles extracted from the *Internet Archive*. The CNN for facial feature extraction is evaluated on the *Labeled Faces in the Wild (LFW)* dataset [107]. Finally, we evaluate the performance of our system by presenting two use cases along with appropriate graphical representations that visualize the person relations extracted from the news articles. To the best of our knowledge, this is the first approach to identify entities in the *Internet Archive* solely using image data.

The remainder of this section is organized as follows. Section 3.4.1 reviews related work on face recognition. In Section 3.4.2, we introduce our deep learning system to identify persons in the image content of the *Internet Archive*. Experimental results for the face recognition approach and some use cases are presented in Section 3.4.3. Section 3.4.3.4 summarizes this work and outlines areas of future work.

### 3.4.1 Related Work

Face recognition has been a well-studied computer vision task for decades, and the performance has significantly improved since deep learning (Section 2.2.2) as well as huge public data collections like *CASIA-WebFace* [295], *Microsoft-Celebrity-1M* (MS-Celeb-1M) [87], or

*VGGFace* [47, 190] have been introduced. As stated by Jin and Tan [116], a face recognition system relies on three modules for face detection, face alignment using facial landmark detectors, and facial feature extraction. The remainder only focuses on facial feature extraction based on deep learning approaches. Surveys including face and facial landmark detection can be found in Jin and Tan [116], Ranjan et al. [208], and Wu and Ji [282].

Unlike many other image classification tasks, face recognition applications differentiate between a high number of identities that are rarely covered completely by publicly available training datasets. Thus, related work based on deep neural networks introduced new loss functions like the contrastive loss [243, 244, 246, 247, 272, 295], triplet loss [62, 63, 190, 224, 228], center loss [60, 197, 278, 283, 308], and large margin loss [59, 147, 148, 271] to learn robust face representation. These representations are used for face verification by comparing facial features extracted from an image (or video) to exemplary images of each identity of interest. Pose-variations, occlusions, and aging are among the main challenges in face recognition as they drastically increase intra-class variations. Some approaches use 3d face reconstruction [15, 157–159, 161], autoencoders [199, 291, 296, 319], or more recently *Generative Adversarial Networks* (GANs) [28, 29, 48, 234, 311] to synthesize new views (poses) of the face in order to augment the training dataset and improve the robustness against pose variations. Others instead frontalize faces using autoencoders [309, 318, 320] or GANs [108, 260, 297] to generate a normalized view, or alternatively use CNNs to directly learn a mapping for normalization [104, 316, 317]. Another widely applied technique to increase the robustness against poses and occlusions is to use multi-input networks [63, 243–246, 292] that use several image patches around facial landmarks as input to extract and combine their features. As training data of individuals at a different age is rare, approaches on cross-age face recognition suggest to, e.g., generate a face for a given age [17, 18, 275, 290] or decompose aging and identity components to obtain age-invariant features [270, 277, 312]. Masi et al. [160] provide a more detailed overview that also includes many other variations affecting face recognition, such as make-up or low-resolution images.

#### 3.4.2 Person Identification in Archived Web News

In this section, a system for the identification of interesting persons in photos of archived web news is introduced. First, a CNN is trained to learn robust facial representations (Section 3.4.2.1). Subsequently, we describe a way to define a lexicon of persons and to automatically gather example images for them from the Web to build an entity dictionary for a given domain like *politics* or *entertainment* (Section 3.4.2.2). For this purpose, we explain how to reduce the amount of irrelevant data, i.e., photos that do not depict the queried person, in the example dataset caused by the web-based image retrieval. Finally, the proposed approach that identifies persons in image data retrieved from the *Internet Archive* (Section 3.4.2.3). The workflow is illustrated in Figure 3.15.

#### 3.4.2.1 Learning a Feature Representation for Faces

A CNN is trained to learn facial representations for person identification in the subsequent steps. Given a dataset of face images, such as MS-Celeb-1M [87] or *CASIA-WebFace* [295], covering  $n$  individual persons, a model with  $n$  classes is trained for classification. During training, the *cross-entropy loss* is minimized given the one-hot encoded vector  $\mathbf{y} \in \{0, 1\}^n$  for the ground-truth class and predicted probability distribution  $\hat{\mathbf{y}}$  of the model:

$$\mathcal{L} = - \sum_{i=1}^n y_i \log(\hat{y}_i) . \quad (3.15)$$

Removing the fully-connected layer that assigns probabilities to the pre-defined classes of faces transforms the model into a generalized feature extractor. Thus, for a query image, the model outputs a compact vector of facial features  $\mathbf{f}$ . In this way, a query image can be compared with the facial features of entities in the pre-defined dictionary, which is presented in the next section.

#### 3.4.2.2 Creating a Dictionary of Persons for a Domain

First, it is explained how to automatically define entities and gather example images for them from the Web. Second, the process of defining a compact representation for every entity is described. In this context, three strategies for filtering inappropriate facial features are introduced and discussed.

**Selection of Relevant Persons:** As a first step, it is necessary to define entities of interest that the approach should identify in the archived web news collections. There are several options available to define a dictionary of relevant persons. (1) The person dictionary can be manually defined by the user(s) according to the specific needs and goals. (2) NER & NED approaches can be applied to extract mentions of people from the corresponding textual content automatically. (3) External sources such as the *Wikipedia* encyclopedia can be leveraged to identify which people are relevant for a general audience. We follow the latter approach to automatically choose a set of relevant persons  $\mathbb{P}$  whose *Wikipedia* pages were viewed most frequently in a given year and who were born after 1920. Only persons  $p \in \mathbb{P}$  associated with the target group, such as *politicians*, are considered to specify a target domain. However, the person dictionary can be modified according to specific user needs since example image material is gathered automatically using a web-based image retrieval.

**Web-based Retrieval of Example Images:** Since the person dictionary might contain a large number of entities, a manual selection of representative example images is, in general, not feasible. Instead, we propose an automatic web-based approach to retrieve exemplary



images for each person. Given the names of the selected entities, an image search engine such as *Google Images* is crawled to find a given number of  $k$  example images for each person  $p \in \mathbb{P}$ . However, the collected images do not necessarily always or only depict the target person  $p$ , and irrelevant photos should be eliminated for the subsequent steps.

**Extraction and Filtering of Feature Vectors:** In order to distinguish between the target person from other persons, it is necessary to compare feature vector representations describing the characteristics of all facial regions in the retrieved image material of a specific person  $p \in \mathbb{P}$ . First, a face detection approach is applied to retrieve the facial regions in a photo. We have used the *dlib* [123] face detector based on *Histogram of Oriented Gradients* (HOG) features [57] and an SVM classifier. Though not able to detect extreme facial poses, this face detector ensures efficiency in terms of computational speed when it comes to the large-scale image data of news pages gathered by the *Internet Archive*. For all  $v$  faces (face areas) detected in the photos crawled for a person  $p \in \mathbb{P}$ , a set of feature vectors  $\mathbb{F}_p = \{\mathbf{f}_p^{(1)}, \mathbf{f}_p^{(2)}, \dots, \mathbf{f}_p^{(v)}\}$  is computed using the CNN model presented in Section 3.4.2.1.

Since the detected faces can depict the target person  $p$  but also other individuals, a filtering step on the extracted facial features  $\mathbb{F}_p$  is conducted. For this purpose, it is necessary to determine a target feature vector  $\mathbf{f}_p^*$  representing the individual  $p$ . For the choice of this vector, we propose three strategies: (1) a **manual selection** of one or multiple representative face region(s) within the example material, (2) calculating the **mean vector** of all facial representations, or (3) applying a **clustering** approach to calculate the mean of all facial representations within the majority cluster that most likely represents the queried person. We have applied an agglomerative clustering approach (using Ward Jr [276]’s minimum variance method for linkage) based on the cosine similarity between all feature vectors of an individual. The cosine similarity  $s(\mathbf{f}_p^{(i)}, \mathbf{f}_p^{(j)})$  between a feature vector  $\mathbf{f}_p^{(i)} \in \mathbb{F}_p$  and another feature vector  $\mathbf{f}_p^{(j)} \in \mathbb{F}_p$  with  $i \neq j$  from images crawled the same person  $p$  is defined as:

$$s(\mathbf{f}_p^{(i)}, \mathbf{f}_p^{(j)}) = \frac{\mathbf{f}_p^{(i)} \cdot \mathbf{f}_p^{(j)}}{\|\mathbf{f}_p^{(i)}\|_2 \cdot \|\mathbf{f}_p^{(j)}\|_2} . \quad (3.16)$$

The feature vectors are assigned to the same cluster as long as their similarity is above a similarity threshold  $\tau_c$  that is used as the clustering stopping criteria. Note that we use the normalized cosine similarity (Equation (3.17)) for comparison throughout this section:

$$\rho(\mathbf{f}_p^{(i)}, \mathbf{f}_p^{(j)}) = \frac{s(\mathbf{f}_p^{(i)}, \mathbf{f}_p^{(j)}) + 1}{2} \in [0, 1] . \quad (3.17)$$

The manual selection of one or multiple representative face region(s) is the most reliable option since it unambiguously represents the target entities and ensures more robust filtering. However, in contrast to both other unsupervised approaches, it does require human

### 3 Information Extraction from Photos

supervision and might not be viable if a large number of entities is considered. Taking the mean of all facial representations relies on the assumption that a majority of facial regions in the retrieved exemplary material already depict the target person. While this is usually the case for popular or famous people (public figures), this approach might fail for less popular persons containing more irrelevant images, i.e., photos that do not depict the specified person, in the exemplary material. Thus, a clustering approach seems to be more robust since the facial features within the majority cluster more likely represent the queried person.

Finally, the target feature vector  $\mathbf{f}_p^*$  determined by one of the aforementioned approaches is compared to all facial representations  $\mathbf{f}_p^{(i)} \in \mathbb{F}_p$  of the queried person using the normalized cosine similarity  $\rho(\mathbf{f}_p^{(i)}, \mathbf{f}_p^*)$  (Equation (3.17)). We keep each facial representation  $\mathbf{f}_p^{(i)} \in \mathbb{F}_p$  with a normalized cosine similarity greater than a threshold  $\rho(\mathbf{f}_p^{(i)}, \mathbf{f}_p^*) > \tau_c$  to create a filtered set of feature vectors  $\mathbb{F}_p^* \subset \mathbb{F}_p$ . The evaluation of the proposed approaches for filtering as well as the choice of threshold  $\tau_c$  are discussed in Section 3.4.3.2.

After the filtering step is applied, we calculate the mean feature vector  $\bar{\mathbf{f}}_p$  of the remaining facial representations  $\mathbb{F}_p^*$  for each person  $p \in \mathbb{P}$ . As a result, the number of comparisons for each face found in a web archive is reduced to the number of persons  $|\mathbb{P}|$  in the person dictionary. Although a comparison to each remaining facial representation might lead to better results, it is much more computationally expensive.

#### 3.4.2.3 Person Identification Pipeline

The components introduced in the previous sections enable automatic identification of persons in the image data of the *Internet Archive*. Given a photo, the face detector (same as used in Section 3.4.2.2) is applied to extract face regions. Facial representations for these regions are computed using the CNN described in Section 3.4.2.1 and subsequently compared to the representative feature vector  $\bar{\mathbf{f}}_p$  of each person  $p \in \mathbb{P}$  in the dictionary using the normalized cosine similarity. This comparison allows for determining the most similar (likely) person shown in each image region. Given the similarity value  $\hat{\rho}$  of the most likely person  $\hat{p} \in \mathbb{P}$ , the identification threshold  $\tau_{id}$  decides whether the face region depicts this person ( $\hat{\rho} \geq \tau_{id}$ ) or an unknown (out-of-dictionary) person ( $\hat{\rho} < \tau_{id}$ ). Based on the results of the person identification, visualizations for single and joint occurrences of persons of interest in news articles of the *Internet Archive* can be created.

#### 3.4.3 Case Study & Qualitative Results

In this section, we evaluate the components of the proposed person identification approach. First, the dictionary of persons and dataset is introduced (Section 3.4.3.1). We present details of the technical realization and experimental results on the learned face representation as well as parameter selection in Section 3.4.3.2. Without loss of generality, the feasibility

### 3.4 Person Identification in News Articles of the Internet Archive

of our system is demonstrated on image data of the *Internet Archive* concentrating on a selection of *German* web content (Section 3.4.3.3). For this purpose, visualizations for relations among the persons of interest in the selected data are shown.

#### 3.4.3.1 Person Dictionaries & News Dataset

**Person Dictionaries:** In our experimental setting, the goal is to recognize persons in the German web content of the *Internet Archive* published in 2013 and visually infer relations among them. Hence, people of public interest have to be selected for the dictionary. We choose the groups of *politicians* and *actors* for each of whom we create a dictionary according to the description in Section 3.4.2.2. We query the German *Wikipedia* for persons according to the selected occupations to obtain German as well as international personalities. The entity names are fetched via *SPARQL* queries to the *Wikidata* knowledge base [268], along with the number of page views. Since *Wikidata* provides page views from mid-2015, we fetch the numbers for the year 2016. This results in a minor mismatch in terms of time concerning our search space for the *Internet Archive* data containing articles published in 2013. However, the extracted persons are still identifiable and relevant, as shown in Section 3.4.3.3. The number of page views determines the relevance of the collected entities. Thus, the ranked list of entity names is reduced to the first 100 most relevant entries. Given the sets of persons for the selected occupational groups, we crawl the *Google Images* search engine for a maximum of  $k = 100$  images per entity using the entity name.

**News Dataset:** The *Internet Archive* contains an enormous amount of multimedia data that can reveal dependencies between entities in various fields. Looking only at the collection of web pages, a large part of the multimedia content is irrelevant for person search, e.g., shopping websites. For this reason, we aim at selecting useful and interesting domains in which the entities from the dictionary are depicted. To demonstrate the feasibility of image analytics in web archives, we have selected two popular German news websites `welt.de` and `bild.de`. While `welt.de` addresses *political subjects*, `bild.de` has a stronger focus on *entertainment news* as well as *celebrity gossip*. We select image data published in the year 2013, in which the German *elections* took place. The number of analyzed news photos and corresponding faces extracted from them is shown in Table 3.12.

#### 3.4.3.2 Parameter Selection

**Network Training:** Several publicly available datasets exist to train a CNN model for the person recognition task. We use the large-scale MS-Celeb-1M [87] dataset comprising 8.5 million images of around 100 thousand different persons as input data to learn the network weights. A classification model considering all the available identities of the dataset is trained using the *ResNet-101* architecture [92] with 101 convolutional layers (Section 2.2.5).

### 3 Information Extraction from Photos

Table 3.12: Number of photos and faces extracted from archived news articles of the selected domains published in 2013

Domain	Images	Faces
welt.de	648,106	205,386
bild.de	566,131	243,343

Table 3.13: Results of methods for the filtering step of the entity dictionary on a subset of 20 politicians

Method	Precision	Recall	$F_1$
No filtering	0.669	1	0.802
Manual selection	0.977	0.922	0.949
Mean vector	0.993	0.449	0.618
Clustering	0.985	0.912	0.947

The network weights are initialized by pre-trained weights learned on the ILSVRC 2012 dataset [58, 218]. Furthermore, we augment the data by randomly selecting an area covering at least 70% of the image. The input images are then randomly cropped to  $224 \times 224$  pixels. The SGD optimizer is used with a momentum of 0.9. The initial learning rate of 0.01 is exponentially decreased by a factor of 0.5 after every 100,000 iterations. The model is trained for 500,000 iterations with a batch size of 64. The trained model is available at <https://github.com/TIB-Visual-Analytics/PIIA>.

**Evaluation of the Model Performance and the Clustering Threshold:** The trained model is evaluated on the LFW benchmark [107] to measure its performance for face verification and evaluate the threshold  $\tau_c$  applied for face clustering and filtering introduced in Section 3.4.2.2. As suggested by Huang et al. [107], a ten-fold cross-validation is conducted, where each fold consists of 300 matched and mismatched face pairs. The normalized cosine similarity (Equation (3.17)) between the feature vectors of two face images is calculated to determine whether the images depict the same or different individuals. For each subset, the best threshold maximizing the accuracy of the remaining nine subsets is calculated. Finally, the yielded accuracy, as well as threshold values, are averaged for the ten folds.

We obtained an accuracy of 98.0% with a threshold of  $\tau_c = 0.757$  (for normalized cosine similarities  $\rho \in [0, 1]$ ). Compared to approaches with state-of-the-art results on the LFW benchmark reviewed in Section 3.4.1, our model yields competitive results using a base architecture and loss function and provides a solid basis for our system. Moreover, the estimated threshold can be considered stable as it has a standard deviation of only 0.002.

**Evaluating the Methods for Feature Vector Filtering:** In Section 3.4.2.2, three methods for selecting a target vector  $\mathbf{f}_p^*$  for each person  $p \in \mathbb{P}$  were introduced to filter irrelevant faces in the example material. We manually annotated whether 1,100 facial regions detected in example photos crawled for 20 politicians depict the queried person or not. A comparison of the proposed options for filtering regarding mean precision, recall, and  $F_1$ -score values of the filtered faces to the ground truth annotations obtained for the 20 politicians is shown in Table 3.13. The results demonstrate that the best performance

regarding the  $F_1$ -score was achieved using manually selected feature vectors for data filtering. However, the results of the agglomerative clustering approach are comparable and were achieved without any supervision. Thus, this method allows for fully-automatic person identification and is used in the remainder of the case study in Section 3.4.3.3.

**Evaluating an Identification Threshold for Face Verification:** After filtering irrelevant example photos, each entity is described by its mean vector  $\bar{\mathbf{f}}_p$ . The use of a mean vector is plausible for our approach since we do not detect and identify faces in extreme poses. A ten-fold cross-validation is performed based on the manually annotated subset of politicians from the previous paragraph to evaluate the identification threshold  $\tau_{id}$  that determines whether or not a face depicted in a news image represents a person in the person dictionary. An accuracy of 96% is obtained. The threshold results in  $\tau_{id} = 0.833$  (for the normalized cosine similarities  $\rho \in [0, 1]$ ) and shows a standard deviation of 0.002. In particular, the small standard deviation implies that the mean entity vector works very stable for the face verification task of our approach.

#### 3.4.3.3 Face Recognition in Image Collections of the Internet Archive

We conducted a case study based on news articles extracted from the *Internet Archive* (Section 3.4.3.1). To quantify the relevance of individuals and their relation to other entities, we count how often the specific entities were identified in the selected image data and how often they are portrayed with other persons in the dictionary. Exemplary results are shown in Figure 3.16 and can be interactively explored in our demo available at <https://github.com/TIB-Visual-Analytics/PIIA>.

Figure 3.16 (top) visualizes relations between well-known heads of states and other politicians in 2013 inferred by our analysis system for the German news website *welt.de*. The graph shows that *Angela Merkel*, the German chancellor, and the former German minister of foreign affairs, *Guido Westerwelle*, appear most frequently in the image data. They also often seem to appear together, as indicated by the strong relationship in Figure 3.16, which is reasonable given their political role. The most relevant international politician detected in the news photos is *Barack Obama*, who also shares a strong relation to *Angela Merkel*. The relation of *Guido Westerwelle* to *Frank-Walter Steinmeier* is due to the transition of power in late 2013. Besides, relations between former and new heads of states of *Germany* and the *USA* have been revealed.

Figure 3.16 (bottom) visualizes relationships between different actors in 2013. For example, the graph indicates that the actors *George Clooney* and *Sandra Bullock*, who have both acted in the movie *Gravity*, often appear together. Moreover, actors of the sitcom *The Big Bang Theory* (*Kaley Cuoco*, *Jim Parsons*, *Johnny Galecki*) share relations with each other. The strongest relation has been discovered between *Angelina Jolie* and *Brad Pitt*, which is

### 3 Information Extraction from Photos

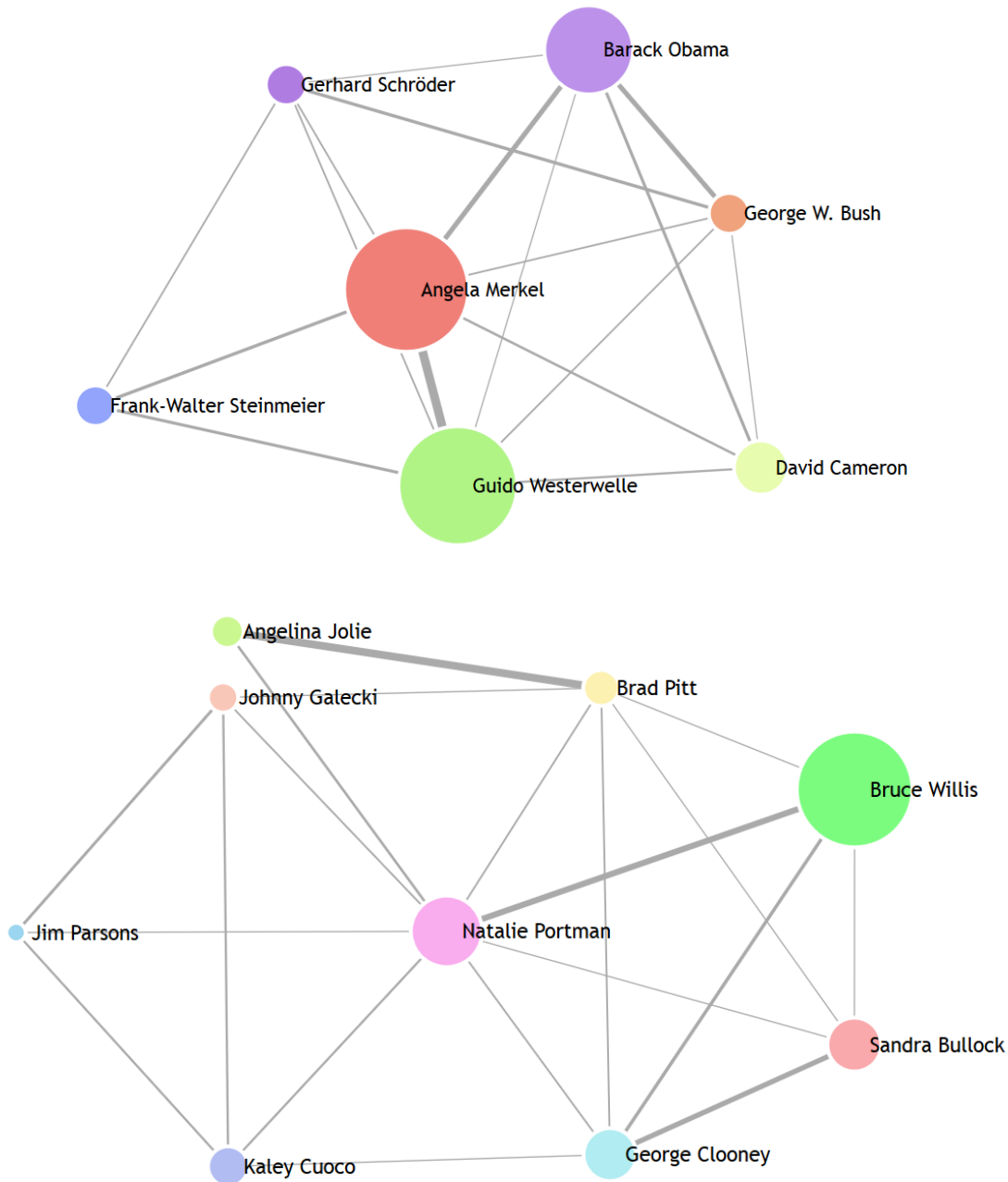


Figure 3.16: Graph showing relationships among an exemplary set of international *politicians* (top) and *actors* (bottom) using the domains `welt.de` and `bild.de`, respectively. The size of vertices encodes the occurrence frequency of the entity. The strength of edges denotes the frequency of joint occurrences.

reasonable as they are a famous actor couple. The actress *Natalie Portman* provides connections to all actors of the graph having the second strongest appearance frequency. These relations imply that there must be several images published in `bild.de` which depict her with colleagues, maybe due to a celebrity event like the *Academy Awards*.

#### 3.4.3.4 Summary

In this section, we have presented a system for the identification of persons of interest in the image content of web news in the *Internet Archive*. For this task, a CNN-based feature representation for faces was trained and evaluated on the LFW benchmark set. Moreover, we introduced a semi-automatic web-based method to create a dictionary of persons of interest, given a domain of interest. In addition, methods for filtering inappropriate images in the example data were introduced and evaluated, including a robust and fully-automatic filtering based on an agglomerative clustering approach. In order to cope with the enormous amount of image content the *Internet Archive* provides, a constrained search domain was defined. The proposed system reliably detects dictionary entities and reveals relations between the entities by means of joint occurrences. For this reason, we argue that deep learning approaches for face identification provide rich image features that should be well suited to quantify cross-modal person relations between image and text, as proposed in Chapter 4.

In the future, we plan to improve individual steps of the pipeline further. In particular, we aim to improve our deep learning model using a more sophisticated loss function or preprocessing steps to increase the robustness for pose and age variation. The process of determining a representative feature vector for individual persons can be enhanced by querying *Wikipedia* or *Wikidata* images that likely contain less to no irrelevant images. Finally, the approach will be extended to allow the exploration of relations of persons across different domains.





## 4 Multimodal Analytics using Measures of Cross-modal Consistency

In the previous chapter, deep learning approaches for event classification (Section 3.1), geolocation estimation (Section 3.2), date estimation (Section 3.3), and face recognition (Section 3.4) were introduced to extract rich information from photos. This information is a vital prerequisite to quantify entity relations between photos and text. Referring to Section 1.3, the goal of this thesis is to present an unsupervised and fully-automated approach applicable to real-world news articles and other multimodal documents that provides differentiated cross-modal relations for specific named entities such as (public) persons, locations, dates, and events (research question 1). While results for event classification, geolocation estimation, and person recognition are promising, the date estimation approach presented in Section 3.3 is not applicable to photographs taken after the year 1999, and the average error of about seven years is quite significant. Thus, we focus on measuring the cross-modal consistency of persons, locations, and events in this chapter.

As discussed in Section 1.1.2, previous solutions that quantify cross-modal relations between photos and text can be divided into two categories and have several limitations. Part of the related work [96, 127, 185, 294, 306] has suggested computational models to quantify image-text relations, such as the *Cross-modal Mutual Information* (CMI) [96, 185] that indicates the number of shared concepts between both modalities. However, they do not explicitly consider cross-modal relations of named entities that are relevant in the news. On the other hand, solutions on image repurposing detection [114, 115, 219] aim to verify the cross-modal occurrence of such entities, e.g., persons, locations, or organizations. In a more general sense, these kinds of approaches also quantify *Cross-modal Mutual Information* (CMI) in terms of shared entities between photos and text. But these solutions rely on multimodal deep learning techniques that require appropriate datasets with non-manipulated pairs of photos and text, which are hard to acquire automatically as they need to be verified for valid cross-modal relations. In addition, these methods cannot cope well with the ever-growing amount and diversity of entities covered in the news since they are restricted to the verification of entities that appear in the datasets used for training or retrieval. Experimental evaluation has been performed on images with relatively short image captions [114, 219] or existing metadata [115] using closely related reference data, which do not reflect real-world characteristics as illustrated in Figure 4.1.

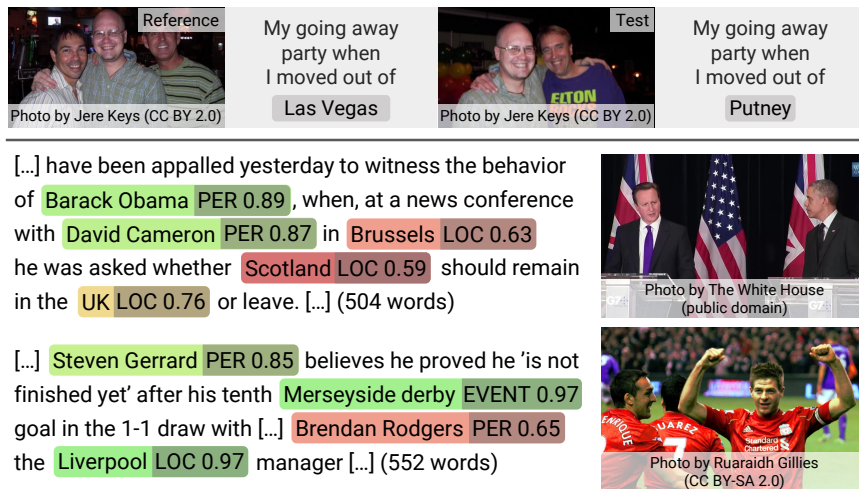


Figure 4.1: **Top:** Test and reference images of the MEIR dataset [219] and corresponding texts with the original and manipulated entity. **Bottom:** Two real-world news from *BreakingNews* [207] and outputs of our system for locations (LOC), persons (PER), and events (EVENT). The examples show that real-world news articles have much longer text and refer to many entities. Photos are replaced with similar ones depicting the same entity relations due to image copyright restrictions. Links to the original documents can be found on: [https://github.com/TIBHannover/cross-modal\\_entity\\_consistency/tree/master/supplemental\\_material](https://github.com/TIBHannover/cross-modal_entity_consistency/tree/master/supplemental_material)

In this chapter, we propose an unsupervised approach that quantifies the cross-modal consistency of entity relations. In contrast to previous work, the approach allows for a more fine-grained quantification of *Cross-modal Mutual Information* (CMI) and is completely unsupervised as it does not rely on any pre-defined reference or training data. To the best of our knowledge, we present a first system *applicable to real-world news articles* by tackling several news-specific challenges such as the arbitrary length of news documents, entity diversity, and irrelevant reference images. We automatically crawl reference images for entities extracted from the text by *Named Entity Recognition and Disambiguation* (NER & NED). These images serve as input for the verification of the entities to the accompanying news image. For this purpose, the proposed computer vision approaches from Chapter 3 are used as generalized feature extractors. Finally, novel measures for different entity types (persons, locations, events) as well as for the more general news context are introduced to quantify the cross-modal similarity of image and text. As mentioned in Chapter 1, the applications are manifold, ranging from a retrieval system to find articles with low or high cross-modal correlations to an exploration tool that reveals the relations between image and text (Figure 4.1). The feasibility of our approach is demonstrated on a novel large-scale dataset for cross-modal consistency verification derived from *BreakingNews* [207]. The dataset contains real-world news articles in English and covers different topics and domains. In addition, we have collected articles from *German* news sites to verify the performance in another language. In

contrast to previous work on image repurposing detection [114, 219], the entities are manipulated with more sophisticated strategies in order to obtain challenging datasets. Source code, web application, and datasets are publicly available<sup>23</sup>. The web application has been published as a demo paper at the *ACM SIGIR Conference on Research and Development in Information Retrieval 2021* [238].

The remainder of this chapter is organized as follows. Related work is reviewed in Section 4.1. The unsupervised system to quantify cross-modal entity and context relations between photos and text is described in Section 4.2 and Section 4.3. Section 4.4 introduces two benchmarks datasets that allow for measuring the performance of the proposed solutions for document verification and collection retrieval. The experimental results on these datasets are discussed in Section 4.5. A demonstrator of the proposed system is presented in Section 4.6. Section 4.7 summarizes this chapter and outlines areas of future work.

## 4.1 Related Work

The analysis of multimodal information such as image and text has attracted researchers from both communication and computational science for many years. According to Bateman [31], the consideration of multimodal relationships is crucial for understanding the overall multimodal message. As a comprehensive survey [26] on multimodal machine learning reveals, computer scientists have introduced a variety of novel multimodal approaches for specific tasks like image captioning [118, 265, 299, 307] or visual question answering [16, 19, 105, 236]. However, these applications typically disregard the deeper meaning between multimodal information and do not allow any form of (human-like) interpretation. On the other hand, communication scientists [30, 88, 153, 155, 262] attempt to assign joint placements of image and text to distinct image-text classes in order to define the interrelations using suitable taxonomies. However, only recently, few works attempted to build computational models to quantify the cross-modal relations between image and text. They can be divided into two categories and are described in more detail in the remainder of this section.

### 4.1.1 Quantification of Image-Text Relations

Several works [96, 127, 185, 294, 306] have recently been proposed that introduce computational concepts and models to quantify mutual information and semantic correlations between image and text. They aim to bridge the semantic gap [237] between both modalities in order to build more powerful models, e.g., for information retrieval. Henning and Ewerth [96, 97] suggested two computable metrics, namely *Cross-modal Mutual Information* (CMI) and *Semantic Correlation* (SC), to quantify the relations between image and text. CMI describes the number of shared concepts (e.g., objects) in both modalities. SC,

<sup>23</sup>[https://github.com/TIBHannover/cross-modal\\_entity\\_consistency](https://github.com/TIBHannover/cross-modal_entity_consistency)

on the other hand, aims to quantify the shared meaning of both modalities by considering the overall context regardless of the shared concepts. Due to the absence of large-scale training datasets, Henning and Ewerth [96, 97] train an autoencoder to reconstruct the multimodal input. The encoder learns a compact, low-dimensional representation of the input used by the decoder to reproduce the salient parts of the image and text. Finally, the encoder-network is used in conjunction with a limited amount of training data to train the final classifier that outputs scores for CMI and SC. Otto et al. identified that the *Semantic Correlation* (SC) is comparable to the semantic relations from Marsh and White [153]. Furthermore, they analyzed established taxonomies from communication science [153, 155, 262] and observed that some image-text classes entail a difference in the abstractness level between image and text. They applied the autoencoder approach from Henning and Ewerth [96, 97] to identify whether the image is an abstraction of the text or vice versa. In another work, Otto et al. [185, 186] identified the *Status* relation proposed by Barthes [30] as an essential image-text relation that has been adopted by the majorities of taxonomies [155, 262] established by communication scientists. Thus, they extended the initial set containing CMI and SC proposed by Henning and Ewerth [96, 97] with the *Status* relation. It describes the hierarchical relation between an image and text with respect to their relative importance. In this way, it can be quantified whether image and text are equally important for conveying the entire multimodal message or whether one modality (text or image) is "subordinate" to the other. Furthermore, they proposed a novel dataset to directly train a multimodal deep learning approach that outputs scores for three image-text dimensions: CMI, SC, and *Status*. These scores are used to characterize eight specific image-text classes (Figure 1.2, page 4), which are partially compliant to classes in existing taxonomies.

Other works take a more differentiated approach to image-text relationships. Zhang et al. [306] investigate image-text relations in advertisements. They claim that image-text alignment methods alone are insufficient to detect parallel relationships between both modalities because the information from text and image do not always align but can still convey the same message. Zhang et al. [306] use a variety of features from both modalities, as well as methods that analyze the semantics within and across channels to predict parallel or non-parallel relationships between image and text. Ye et al. [294] further extended this approach by interpreting the rhetoric of advertisements using cross-modal embeddings and image embeddings for symbol regions. Both aforementioned approaches define their own types of image-text relations and do not leverage established relations from previous work, e.g., from the field of linguistics. Kruk et al. [127] provided some additions to the taxonomy of Marsh and White [153] to determine the author's intent in *Instagram* posts. In compliance with Bateman [31], they realized that the combination of different modalities could create a new meaning that needs to be modeled more carefully. Thus, Kruk et al. [127] modeled contextual and semiotic relationships between the literal and signified meanings of the image and caption, respectively.

### 4.1.2 Image Repurposing Detection

Solutions on image repurposing detection [114, 115, 219] intend to reveal inconsistencies between image-text-pairs concerning more concrete entities (persons, locations, organizations, etc.), mainly to identify repurposed multimedia content that might indicate misinformation.

Jaiswal et al. [114] presented an assessment system that considers a multimedia package containing an image and a corresponding caption to verify its semantic integrity. They train a multimodal deep learning model (e.g., a multimodal autoencoder [182]) that jointly encodes features from photos and text to output consistency scores. A VGG-19 [235] is applied to generate image features, and the average *Word2Vec* embeddings [164] of the caption is used as a textual representation. Finally, an outlier detection model (SVM or isolation forest [144]) is trained on the reference dataset and used to output the inlierness of an image-caption pair, which is considered as the semantic integrity by the authors. To evaluate the system, they constructed a synthetic dataset with manipulated image-text-pairs by completely replacing one modality (image or text). However, this results in semantically inconsistent image-caption-pairs that are relatively easy to detect. Sabir et al. [219] improved this dataset and carefully replaced specific entities with entities of the same type (persons, locations, and organizations) to generate semantically consistent altered packages. They have also refined the multimodal model using a multitask learning approach that further incorporates geographical information. Alternatively, Jaiswal et al. [115] presented an adversarial neural network that simultaneously trains a bad actor who intentionally counterfeits metadata and a watchdog who verifies the multimodal semantic consistency. The counterfeiter selects manipulated metadata for a given image by analyzing the similarity to images of different entities in the reference dataset. On the other hand, the watchdog uses evidence from the reference dataset to assess the credibility of the claimed metadata. The system was tested for person verification, location verification, and painter verification of artworks. However, in contrast to the aforementioned approaches, the system is more closely related to work on metadata verification [51, 52, 117, 137] as it only verifies the consistency between pairs of images and metadata and does not incorporate any textual information.

## 4.2 Cross-modal Entity Consistency

This section presents an unsupervised system that goes beyond existing approaches and automatically verifies the cross-modal relations in terms of shared entities between pairs of photos and text. Verification is realized through measures of cross-modal similarity for different entity types (persons, locations, and events). Based on NER & NED (Section 4.2.1), example photos for the detected entities are collected from the Web. Features are obtained from the photos by appropriate computer vision approaches (Section 4.2.2), which are used

## 4 Multimodal Analytics using Measures of Cross-modal Consistency

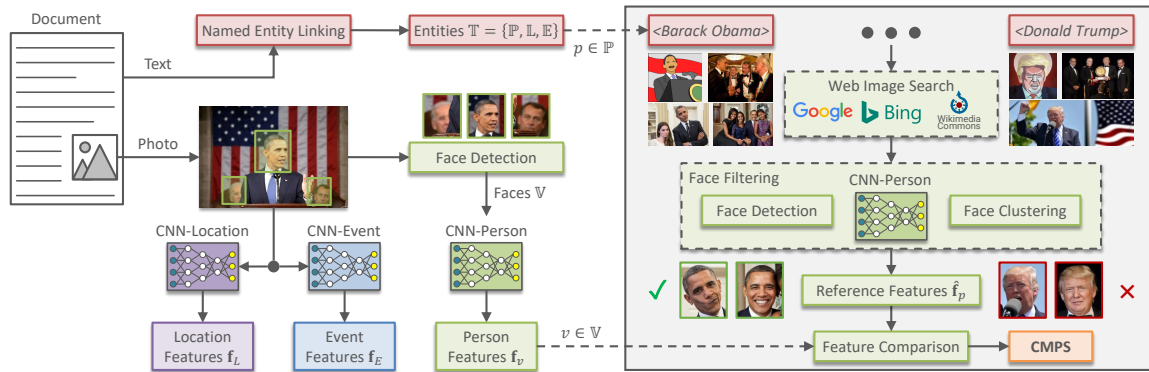


Figure 4.2: Workflow of the proposed system to quantify cross-modal entity similarities. **Left:** Extraction of entities  $\mathbb{T}$  from the text according to Section 4.2.1, as well as features for persons  $\mathbb{P}$  (green), locations  $\mathbb{L}$  (purple), and events  $\mathbb{E}$  (blue) from the document’s photo (Section 4.2.2). **Right:** Workflow to measure the *Cross-modal Person Similarity* (CMPS) between photo and text (Section 4.2.3) based on example images crawled from the Web. The same pipeline but without filtering is used for locations and events.

in conjunction with measures of cross-modal similarity (Section 4.2.3) to quantify the cross-modal consistency. The workflow is illustrated in Figure 4.2.

### 4.2.1 Extraction of Entities from the Text

In order to quantify cross-modal relation for specific types of entities, namely persons, locations, and events, *Named Entity Recognition and Disambiguation* (NER & NED) are applied to extract a set of named entities  $\mathbb{T}$  from the text. We have tried several tools such as *AIDA* [98], Rizzo and Troncy [212], or Kolitsas et al. [125]’s approach. In an initial experiment, we found that combining the output of *spaCy* [99]<sup>24</sup> for *Named Entity Recognition* (NER) and *Wikifier* [40, 41] (explained in Section 2.3.3) for *Named Entity Disambiguation* (NED) provide the best results for different languages. Given an NER system for a specific language, *Wikifier* enables our system to support a large number of 100 languages. Furthermore, it can dynamically detect entities in the text covered in *Wikipedia* information used at inference time, while learning-based approaches are limited to the entities mentioned in the training data and require fine-tuning to adapt to new entities. We link the entity candidate with the highest *PageRank* according to *Wikifier* for every named entity recognized by *spaCy* to the *Wikidata* knowledge base [268]. Linked entities with a *PageRank* below  $1 \cdot e^{-4}$  are neglected due to their low confidence. If *Wikifier* does not provide a linked entity for a given string, the *Wikidata* API function "*wbsearchentities*" is used for disambiguation.

<sup>24</sup>*spaCy* version 2.2.4 was used in this thesis

As shown in Figure 4.2, suitable computer vision approaches based on deep learning are applied to extract features from photos used to quantify the cross-modal entity consistency. The computer vision model is selected based on the type (person, location, or event) of the named entity. Therefore, it is necessary to assign each named entity to one of these entity types. Although some NER tools such as *spaCy* [99] automatically predict entity types, they do not make use of valuable knowledge base information provided by NED. To handle mistakes of entity type classification by *spaCy* and to discard irrelevant entities such as given names that cannot be linked to a knowledge base, the entity types are re-evaluated using the *Wikidata* information of the linked entities based on the following requirements. For persons, only entities that are an instance (*Wikidata* property *P31*) of *human* (*Wikidata* identifier *Q5*) according to *Wikidata* are considered, while for locations, a valid *coordinate location* (*P625*) is set as a requirement. This allows us to extract a variety of locations ranging from *continents*, *countries*, and *cities* to specific *landmarks*, *streets*, or *buildings*. For events, we instead require an entity to be in a verified list of events<sup>17</sup> according to *EventKG* [81, 82], which was also used to create the *Visual Event Ontology* (VisE-O) presented in Section 3.1.2.2. Entities that do not fulfill any of the aforementioned criteria are neglected. As a result, distinct sets of persons  $\mathbb{P}$ , locations  $\mathbb{L}$ , and events  $\mathbb{E}$  are extracted from the text that are verified with example images from the Web, as explained in Section 4.2.3.

#### 4.2.2 Extraction of Features from Photos

Our approach is applicable to articles with multiple images, but we assume that only a single image is present for simplicity. Suitable models are applied to obtain image representations.

**Person Features:** Although the proposed model used in Section 3.4 achieves good results for person identification in news images, we have applied an implementation<sup>25</sup> of *FaceNet* [228] as it provides slightly better results on the LFW benchmark [107]. The *FaceNet* model is used to calculate the individual feature vectors  $\mathbf{f}_v$  of all faces  $v \in \mathbb{V}$  found in the image by the face detection approach from Zhang et al. [305].

**Location Features:** We employ the *base* ( $M, f^*$ ) model<sup>26</sup> for geolocalization [173] presented in Section 3.2 to obtain a geospatial representation of the article’s photo. It provides good results across different environmental settings (*indoor*, *natural*, and *urban*) using a single CNN model. In contrast to the original method, we treat geolocalization as a verification approach and use the feature vector  $\mathbf{f}_L$  from the last pooling layer (Table 2.2, page 37) of the *ResNet-101* model [92, 93].

<sup>25</sup> *FaceNet* implementation from David Sandberg: <https://github.com/davidsandberg/facenet>

<sup>26</sup> *base* ( $M, f^*$ ) model for geolocation estimation: <https://github.com/TIBHannover/GeoEstimation>

**Event Features:** As discussed in Section 3.1, related approaches for event classification [3, 8, 287] have not considered many event types relevant to news and are consequently not capable of distinguishing between them. Thus, we have used a more general image descriptor for place classification (trained on the *Places365* dataset [315]) to extract features for events in our initial method presented in Müller-Budack et al. [175]. As explained in Section 3.1, we recently introduced a dataset and ontology-driven deep learning models for event classification [174]. Unlike previous work, these models distinguish between the majority of newsworthy event types such as *natural disasters*, *epidemics*, and *elections*. For this reason, we use the ontology-driven  $CO_\gamma^{cos}$  model<sup>27</sup> in the approach described in this thesis. The event features  $\mathbf{f}_E$  are extracted from the last pooling layer of the *ResNet-50* architecture [92, 93]. A comparison to the previous approach [175] is conducted in Section 4.5.4.

### 4.2.3 Verification of Shared Cross-modal Entities

In this section, we present measures of *Cross-modal Similarity* (CMS) for different entity types, namely persons, locations, and events. It should be emphasized that we treat each verification task independently. The CMS results for different entity types are *not combined*, which allows a more detailed and realistic analysis. Referring to Figure 4.1 (bottom), please imagine a news article where the image depicts one or several person(s) talking at a conference. While multiple events and locations might be mentioned in the corresponding text, the news image does not provide any visual cues for their verification. This aspect is typical for news articles since the text usually contains more entities and information. In the case of *fake news*, it is expected that only certain entities of one entity type are manipulated to maintain credibility.

#### 4.2.3.1 Verification of Persons

As illustrated in Figure 4.2, we first crawl a maximum of  $I$  example images using image search engines such as *Google* or *Bing* for each person  $p \in \mathbb{P}$  extracted from the NER & NED approach presented in Section 4.2.1. However, as also discussed in Section 3.4, these images can be misleading as they may depict multiple or different persons than the queried one. Thus, a filtering step is necessary. Feature vectors are extracted for each detected face in the reference images according to Section 4.2.2. These features are compared with each other to perform a single-linkage hierarchical clustering with a minimum similarity threshold  $\tau_{\mathbb{P}}$  as a termination criterion. The *normalized cosine similarity* between two feature vectors  $\mathbf{f}_1$  and  $\mathbf{f}_2$  is used for comparison to output a probability distribution  $\rho(\mathbf{f}_1, \mathbf{f}_2) \in [0, 1]$ :

$$\rho(\mathbf{f}_1, \mathbf{f}_2) = 0.5 \cdot \left( \frac{\mathbf{f}_1 \cdot \mathbf{f}_2}{\|\mathbf{f}_1\|_2 \cdot \|\mathbf{f}_2\|_2} + 1 \right) \quad (4.1)$$

---

<sup>27</sup> $CO_\gamma^{cos}$  model for event classification: <https://github.com/TIBHannover/VisE>



The feature vectors of the majority cluster are averaged to create a reference vector  $\hat{\mathbf{f}}_p$  because they most likely represent the queried person  $p$ .

In order to quantify the image-text relation of person  $p$ , its reference vector  $\hat{\mathbf{f}}_p$  is compared to the feature vector  $\mathbf{f}_v$  of each face  $v \in \mathbb{V}$  detected in the document image. We define the entity similarity  $ES_p$  as the maximum among all comparisons since it is sufficient if one face depicts the queried entity:

$$ES_p = \max_{v \in \mathbb{V}} \left( \rho(\mathbf{f}_v, \hat{\mathbf{f}}_p) \right) . \quad (4.2)$$

Using this approach, we can extract the cross-modal similarity of each individual mentioned in the text, as shown in Figure 4.1, allowing for fine-grained document analysis. However, many multimedia and information retrieval applications require an overall cross-modal similarity for the entire multimedia document. Several options are available to calculate the overall *Cross-modal Person Similarity* (CMPS) based on the individual entity similarities  $ES_p$  for all persons  $p \in \mathbb{P}$ , such as the mean,  $n\%$ -quantile, or the maximum of all comparisons. Yet, as mentioned above, the text usually contains more entities than the image, and already a single correlation can theoretically ensure credibility. Since the mean or quantile would require the presence of several or all entities mentioned in the text, we define the CMPS as the maximum similarity among all comparisons according to Equation (4.3)

$$CMPS = \max_{p \in \mathbb{P}} (ES_p) \quad (4.3)$$

#### 4.2.3.2 Verification of Locations and Events

In general, we follow the pipeline of person entity verification. The feature vectors of a maximum of  $I$  reference images for each location and event mentioned in the text are calculated using the CNN of the respective entity type according to Section 4.2.2. However, while some entities are very specific (e.g., *landmarks, sports finals*), others are more general (e.g., *countries, international crises*) and can therefore contain diverse example data. This makes filtering based on clustering complicated as these entities can already contain many visually different sub-clusters due to high intra-class variations.

Thus, the entity similarity ( $ES_t$ ) of an entity  $t \in \mathbb{T}$ ;  $\mathbb{T} \in \{\mathbb{L}, \mathbb{E}\}$  is calculated by comparing the feature vector of the news photo (Section 4.2.2) to the feature vector  $\hat{\mathbf{f}}_{t,i}$  of each reference image  $i \in \mathbb{I}_t$  crawled for the given entity. The entity similarity for a location  $t = l \in \mathbb{L}$  and for an event  $t = e \in \mathbb{E}$  is defined according to the following equations:

$$ES_l = \Psi_{i \in \mathbb{I}_l} \left( \rho(\mathbf{f}_L, \hat{\mathbf{f}}_{l,i}) \right) \quad (4.4)$$

$$ES_e = \Psi_{i \in \mathbb{I}_e} \left( \rho(\mathbf{f}_E, \hat{\mathbf{f}}_{e,i}) \right) \quad (4.5)$$

Since the comparison of all reference images to the news photo leads to a similarity vector  $\mathbf{s}$ , a function  $\Psi : \mathbf{s} \rightarrow [0, 1]$  (e.g., the maximum operator) maps these similarities to a scalar. In the experiments (Section 4.5.2), we evaluate the maximum and several  $n\%$ -quantiles as potential functions. We argue that using an  $n\%$ -quantile is more robust against incorrect or unrelated entity images in the retrieved reference data.

As explained for person verification (Section 4.2.3.1), a single correlation might already ensure the credibility of the whole document, and typically only a single location or event is portrayed in a news photo. Therefore, we decided to use the maximum CMS among all entities of a given type for both the overall *Cross-modal Location Similarity* (CMLS) and *Cross-modal Event Similarity* (CMES).

$$\text{CMLS} = \max_{l \in \mathbb{L}} (\text{ES}_l) \tag{4.6}$$

$$\text{CMES} = \max_{e \in \mathbb{E}} (\text{ES}_e) \tag{4.7}$$

### 4.3 Cross-modal Context Consistency

In the previous section, we have presented an approach that quantifies the cross-modal consistency for each entity based on reference images crawled from the Web. This approach is not feasible for the quantification of the contextual semantic relation since web queries are hard to define automatically based on the entire news content. For this reason, we pursued a different direction. We extracted word embeddings from the article text (Section 4.3.1) as well as the probabilities of general environmental settings (e.g., *beach*, *conference center*, or *church*) from the photo along with the respective word embeddings of the environment names (Section 4.3.2) to quantify the *Cross-modal Context Similarity* (CMCS) according to Section 4.3.3. An overview is illustrated in Figure 4.3.

#### 4.3.1 Text Context

To retrieve suitable candidates representing the context  $\mathbb{C}$  of the text, the part-of-speech tagging from *spaCy* [99] is applied to extract all nouns  $c \in \mathbb{C}$ . They can represent general concepts (e.g., *politics* or *sports*), *places*, or *actions* that might correlate to specific classes, e.g., of a place classification dataset such as *Places365* [315]. We calculate the word embedding  $\mathbf{w}_c$  for each candidate  $c \in \mathbb{C}$  using *fastText* [83] (explained in Section 2.3.1.2) as a prerequisite for the cross-modal comparison explained in Section 4.3.3.

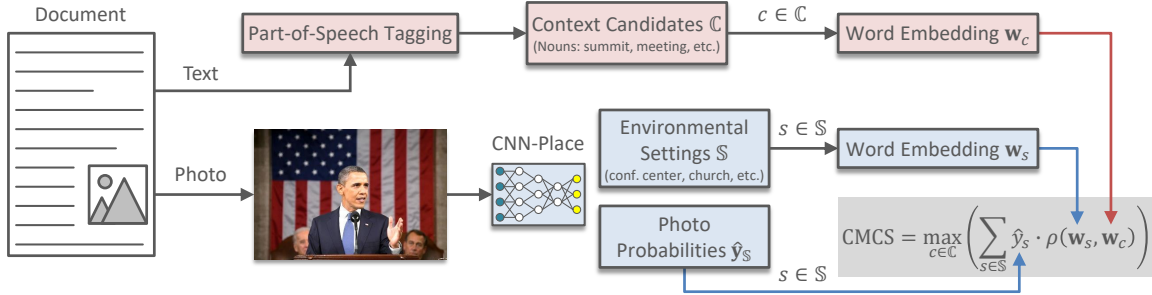


Figure 4.3: Workflow for the quantification of the *Cross-modal Context Similarity* (CMCS). Red boxes indicate the text context (Section 4.3.1), while blue boxes provide the photo context (Section 4.3.2). Part-of-speech tagging is applied to extract context candidates  $\mathbb{C}$ , i.e., nouns from the text. The word embedding  $\mathbf{w}_c$  of each noun  $c \in \mathbb{C}$  is compared to the word embedding  $\mathbf{w}_s$  of 365 environmental settings  $s \in \mathbb{S}$  from the *Places365* dataset [315] using the normalized cosine similarity  $\rho(\mathbf{w}_s, \mathbf{w}_c)$  (Equation (4.1)). The resulting similarity values are weighted with the probabilities  $\mathbf{y}_{\mathbb{S}}$  of the environmental settings extracted from the photo using a CNN model to calculate the CMCS according to Section 4.3.3.

### 4.3.2 Photo Context

A *ResNet-50* model<sup>28</sup> [92, 93] for place classification trained on the *Places365* dataset [315] is applied to predict the probabilities  $\hat{y}_{\mathbb{S}}$  of 365 environmental settings  $\mathbb{S}$  from the photo. As for the text context (Section 4.3.1), *fastText* [83] is employed to extract the corresponding word embeddings  $\mathbf{w}_s$  for the labels of each environmental setting  $s \in \mathbb{S}$ . While environments such as *beach*, *conference center*, or *church* are rather generic, their word embeddings can also be associated with specific news topics such as *travel*, *politics*, or *religion*. Both the probabilities of the environmental settings and their word embeddings are used as photo context. The class labels in the *Places365* dataset were manually translated to *German* for the experiments on *German* news articles.

### 4.3.3 Cross-modal Context Similarity

Unlike the quantification of cross-modal entity consistency (Section 4.2), calculating the *Cross-modal Context Similarity* (CMCS) does not require any reference images. We compare the individual word embeddings  $\mathbf{w}_c$  of each noun  $c \in \mathbb{C}$  to the word embeddings  $\mathbf{w}_s$  of all 365 environmental settings  $s \in \mathbb{S}$  covered by the *Places365* dataset [315] using the *normalized cosine similarity*  $\rho(\mathbf{w}_s, \mathbf{w}_c)$  (Equation (4.1)). Since only certain environmental settings are depicted in a news photo, these similarities are weighted with the respective probability  $\hat{y}_s$  of an environmental setting  $s \in \mathbb{S}$  to integrate the image information. Finally, the *Cross-modal Context Similarity* (CMCS) is defined as the maximum similarity among all comparisons according to the equation in Figure 4.3.

<sup>28</sup> *ResNet-50* model trained with *PyTorch* on *Places365*: <https://github.com/CSAILVision/places365>

## 4.4 Datasets

Two real-world news datasets that cover different languages, domains, and topics are used for the experiments. They are both manipulated to perform experiments for cross-modal consistency verification. Experiments and comparisons to related work [114, 219] on datasets such as MEIR [219] are not reasonable since (1) they do not contain public persons or events, and (2) rely on *pre-defined* reference or training data for *given* entities. These restrictions severely limit the application in practice. We propose an automated solution for real-world scenarios that can handle the vast and ever-growing amount of entities represented in a knowledge base. Source code and datasets to reproduce our results are publicly available<sup>23</sup>.

In the remainder of this section, the manipulation strategies (Section 4.4.1), as well as two novel datasets called *TamperedNews* (Section 4.4.2) and *News400* (Section 4.4.3) are introduced, which contain articles written in English and German, respectively.

### 4.4.1 Manipulation Techniques

We create multiple sets of manipulated entities for each document in our datasets. Similar to Sabir et al. [219], we replace entities extracted from the text at random with another entity of the same type to change semantic relations as little as possible. We also present more sophisticated manipulation techniques as follows. Three additionally manipulated person sets are created by replacing each original person with another person of the same gender (PsG), the same country of citizenship (PsC), or matching both aforementioned criteria (PsCG). Locations are replaced by other locations that share at least one parent class (e.g., *country* or *city*) according to *Wikidata* [268] and are located within a *Great Circle Distance* (GCD) of  $d_{min}$  to  $d_{max}$  kilometers ( $GCD_{d_{min}}^{d_{max}}$ ). Three intervals are used to experiment with different geospatial resolutions at region-level ( $GCD_{25}^{200}$ ), country-level ( $GCD_{200}^{750}$ ), and continent-level ( $GCD_{750}^{2500}$ ). Events that share the same parent class (e.g., *sports competition* or *natural disaster*) with the original event are used for a second set (EsP) of manipulated events. When no valid candidate for a manipulation strategy was available, we have used a random candidate that matched most of the other criteria.

The contextual verification is based on the nouns in the text. Thus, textual manipulation techniques are not applicable. We instead replaced the image with a random image from all other documents for a first manipulated set. We randomly selected similar images (from Top- $k\%$  with  $k \in \{5, 10, 25\}$ ) to maintain semantic relations to create three more sets. The similarity was computed using feature vectors extracted from a *ResNet* model [92, 93] trained on the ILSVRC 2012 dataset [58, 218] for object recognition.

Table 4.1: Number of test documents  $|\mathbb{D}|$ , unique entities  $\mathbb{T}^*$  in all articles, and mean amount of unique entities  $\bar{\mathbb{T}}$  in articles containing a given entity type (for *context*, this is the mean amount of nouns as explained in Section 4.3.1) for *TamperedNews* (top) and *News400* (bottom). Valid image-text relations for *News400* were first manually verified according to Section 4.4.3.

<b>TamperedNews dataset</b>			
<b>Documents</b>	$ \mathbb{D} $	$\mathbb{T}^*$	$\bar{\mathbb{T}}$
All (context)	72,561	—	121.40
With person entities	33,695	4,772	4.01
With location entities	66,484	3,464	4.78
With event entities	15,467	875	1.32

<b>News400 dataset</b>			
<b>Documents</b>	$ \mathbb{D} $	$\mathbb{T}^*$	$\bar{\mathbb{T}}$
All (thereof with manually verified context)	397 (91)	—	137.35
With person entities (thereof manually verified)	320 (116)	413	5.31
With location entities (thereof manually verified)	389 (69)	434	8.83
With event entities (thereof manually verified)	166 (31)	39	1.84

#### 4.4.2 TamperedNews Dataset

To the best of our knowledge, *BreakingNews* [207] is the largest available corpus with news articles that contain both images and text. It originally covered approximately 100,000 news articles published in 2014 written in English across different domains and a huge variety of topics (e.g., *sports*, *politics*, *healthcare*). We created a subset called *TamperedNews* for cross-modal consistency verification of 72,561 articles for which the news text and image were still available. The entities in these articles were additionally manipulated according to Section 4.4.1. Only persons and locations mentioned in at least ten documents and events that occur in at least three documents are considered to discard most irrelevant entities. Detailed dataset statistics are reported in Table 4.1.

#### 4.4.3 News400 Dataset

To show the capability of our approach for another language and time period, we have used the *Twitter* API to obtain the web links (*Uniform Resource Locators* (URLs)) of news articles from three popular *German* news websites (*faz.net*, *haz.de*, *sueddeutsche.de*). The texts and main images of the articles were crawled from the URLs. We have gathered 397 news articles covering four different topics (*politics*, *economy*, *sports*, and *travel*) in the period from August 2018 to January 2019. The smaller dataset size allowed us to conduct a

manual annotation with three experts to ensure valid relationships between image and text. For each document, the annotators verified the presence of at least one person, location, or event in the image as well as in the text and whether the context was consistent in both modalities. Experiments were conducted exclusively on data with valid relations. Again the manipulation techniques presented in Section 4.4.1 are applied to create the test sets. Due to its smaller size, every entity is considered regardless of how often it appears in the entire dataset. The resulting statistics are shown in Table 4.1.

## 4.5 Experimental Setup & Results

In this section, we introduce the tasks and metrics for evaluation (Section 4.5.1) and explain the parameter selection (Section 4.5.2). The performance of the proposed system on real-world news articles is evaluated in Section 4.5.3, and two different deep learning approaches for the quantification of cross-modal event relationships are compared in Section 4.5.4. Finally, the limitations and dependencies of our proposed approach are discussed in Section 4.5.5.

### 4.5.1 Evaluation Tasks and Metrics

The evaluation tasks are motivated by potential real-world applications of our system. We propose to evaluate the system regarding two tasks: (1) document verification and (2) collection retrieval. As illustrated in Figure 4.1, the system can also be used as an analytics tool to explore cross-modal relations within a document efficiently.

**Document Verification:** Please imagine a set of two or more news articles with similar text content and images but differences in the mentioned entities that might have been manipulated by an author with harmful intents. The idea behind this task is to decide which pair of image and entities extracted from the news text provides a higher cross-modal consistency. Thus, document verification can help users to detect the most or least suitable document. We address this task using the following strategy. For each individual document in the dataset, we compare the cross-modal similarities between the news image and the respective set of original entities as well as *one* set of manipulated entities (e.g., PsG) according to the strategies proposed in Section 4.4.1. This separate analysis allows us to evaluate the impact of different manipulation strategies. We report the *Verification Accuracy* (VA) that quantifies how often the original entity set has achieved the higher cross-modal similarity to the document’s image. Some qualitative examples are shown in Figure 4.5 (page 125). Please note that the image is manipulated for the context evaluation instead and that the nouns in the text are considered as "entities".

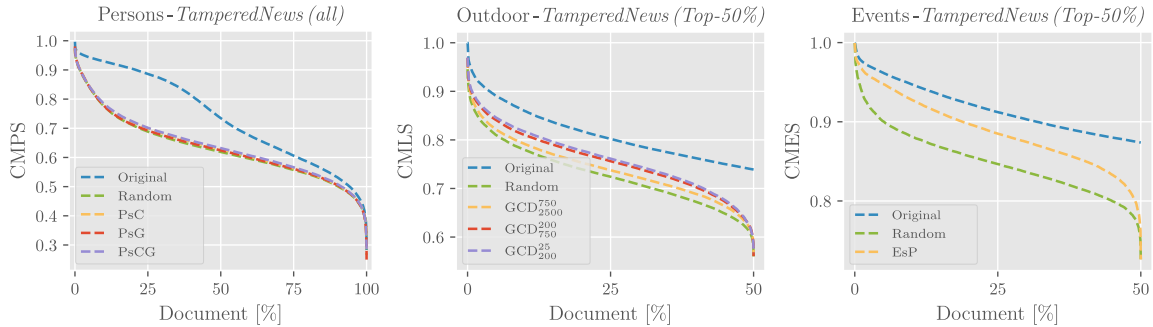


Figure 4.4: Cross-modal similarity values of all (or a subset of) *TamperedNews* documents sorted in descending order for person, location (outdoor), and event entities using different manipulation techniques (notations according to Section 4.4.1)

**Collection Retrieval:** The system can also be leveraged in news collections to retrieve news articles with high or low cross-modal relations to support human assessors to gather the most credible news or possibly *fake news* (in extreme cases). In contrast to document verification, we consider all  $|\mathbb{D}|$  original documents as well as  $|\mathbb{D}|$  manipulated documents applying *one* manipulation strategy. The cross-modal similarities are calculated and used to rank all  $2 \cdot |\mathbb{D}|$  documents. As suggested by previous work [114, 219], the *Area Under Receiver Operating Curve* (AUC) is used for evaluation. We also propose to calculate the *Average Precision* (AP) for retrieving original (AP-original) or manipulated (AP-manipulated) documents at specific recall levels  $R$  according to Equation (4.8). In this respect,  $TP^i$  is the number of relevant documents at position  $i$ . For example, AP-manipulated@25% describes the average precision when  $|\mathbb{D}_R| = 0.25 \cdot |\mathbb{D}|$  of all manipulated documents are retrieved.

$$AP@R = \frac{1}{|\mathbb{D}_R|} \sum_{i=1}^k \frac{TP^i}{i}, \quad (4.8)$$

**Test Document Selection for TamperedNews:** Although the large size of the *TamperedNews* dataset allows for a large-scale analysis of the results, unfortunately, a manual verification of cross-modal relations as for *News400* is infeasible. Thus, reporting the proposed metrics for the whole dataset can be misleading since it turned out during the annotation of *News400* that only a fraction of the documents has cross-modal entity correlations (Table 4.1). As discussed at the beginning of Section 4.2.3, not a single entity mentioned in a news text may be depicted in the corresponding image. To address this issue, we suggest measuring the metrics for specific subsets. We consider the Top-25% and Top-50% documents (denoted as *TamperedNews (Top-d%)*) concerning their cross-modal similarity of original entities since they more likely contain relations between image and text. This selection is also supported by the CMPS values for person verification (Figure 4.4), which decrease more significantly after 25%–50% of all documents and correspond to the percentage of manually verified documents in the *News400* dataset.

Table 4.2: AUC for different functions  $\Psi$  (AC - agglomerative face clustering,  $Q_n$  -  $n\%$  similarity quantile, max - max similarity) to calculate the cross-modal similarity for each entity of a given type (Section 4.2.3). Results are reported for the number of  $|\mathbb{D}|$  documents in the respective *TamperedNews (Top-50%)* dataset with the hardest manipulation strategy (notations according to Section 4.4.1).

Test set	$ \mathbb{D} $	AC	$Q_{75}$	$Q_{90}$	$Q_{95}$	max
Persons: PsCG	16,848	0.93	0.92	<b>0.94</b>	<b>0.94</b>	0.90
Location-Outdoor: GCD <sub>25</sub> <sup>250</sup>	14,113	—	0.71	0.73	0.74	<b>0.77</b>
Location-Indoor: GCD <sub>25</sub> <sup>250</sup>	19,129	—	0.64	0.66	0.67	<b>0.69</b>
Events: EsP	7,734	—	0.72	0.73	0.74	<b>0.75</b>

Please note that experiments on Top- $d\%$  subsets limit the comparability between two approaches to some degree. Depending on the specified parameters (e.g., feature descriptor, function  $\Psi$ ), the Top- $d\%$  subsets comprise different documents. In Section 4.5.4, we explain how a meaningful comparison between two different approaches can be conducted.

#### 4.5.2 Parameter Selection

**Face Clustering Threshold:** The threshold  $\tau_{\mathbb{P}}$  impacts the clustering approach that filters retrieved face candidates for a person explained in Section 4.2.3.1. For this reason, we have tested the *FaceNet* model [228] on the LFW [107] benchmark and evaluated an optimal cosine similarity (normalized to the interval  $[0, 1]$ ) threshold of  $\tau_{\mathbb{P}} = 0.65$ .

**Operator Function for Cross-modal Similarities:** In Section 4.2.3, we mentioned a number of possible functions  $\Psi$  such as the  $n\%$ -quantile or maximum to compute cross-modal similarity based on the comparisons of all reference images of a specific entity to the news image. The AUC using different operators and a maximum of  $I = 10$  reference images for all image sources (*Google*, *Bing*, and *Wikidata*) on the respective *TamperedNews (Top-50%)* subsets are presented in Table 4.2. For comparison, we also tested the face verification using the approach applied for event and location entities described in Section 4.2.3.2. Surprisingly, results for 90% and 95% quantiles are on par with the proposed person clustering. Also, contrary to our assumption that a quantile is more robust against irrelevant photos in the example images for locations and events, it turned out that the maximum operator provides slightly better results for these entity types. These results indicate that irrelevant examples in the reference data have no significant impact on the performance since they are less likely to match the location or event depicted in the news image. In the remainder of this section, results for persons are reported using the clustering strategy, as it yields similar results to the other operators and is likely more robust to larger amounts of unrelated reference images. For locations and events, the maximum operator is applied.



Table 4.3: AUC using different image sources ( $W$  - Wikidata,  $G$  - Google,  $B$  - Bing) and maximum number of  $I$  images on the respective *TamperedNews* (*Top-50%*) subsets. Results are reported for the hardest manipulation strategy (notations according to Section 4.4.1).

Source	#Images			Persons PsCG	Locations		Events EsP
	$I_W$	$I_G$	$I_B$		Outdoor GCD $_{200}^{25}$	Indoor GCD $_{200}^{25}$	
Google	—	20	—	<b>0.95</b>	0.76	0.68	0.73
Bing	—	—	20	0.90	0.76	<b>0.71</b>	<b>0.77</b>
All-10	all	10	10	0.93	0.77	0.69	0.75
All-20	all	20	20	0.93	<b>0.78</b>	<b>0.71</b>	0.76

**Amount and Sources of Reference Images:** In total, we collected a maximum of  $I = 20$  images from the image search engines of *Google* and *Bing* as well as all  $I_W$  available images on *Wikidata* (mostly one *Wikimedia* image) for each entity recognized in the text. We have used multiple sources to prevent possible selection biases of a specific image source and investigated the performance for different image sources and number of images. Since *Wikidata* usually only provides a single or sometimes no image for the linked entities, we exclude it from the comparison. The results on the respective *TamperedNews* (*Top-50%*) subsets for the AUC metric using the hardest manipulation strategies are presented in Table 4.3. They demonstrate that the performance using a single or all image sources is very similar. Also, the results using  $I = 10$  reference images are almost identical compared to the maximum of  $I = 20$  images. Hence, for the rest of our experiments, we use all available image sources with a maximum of  $I = 10$  images per source since this provides a good trade-off between performance and speed and prevents possible selection biases.

### 4.5.3 Experimental Results

In this section, we present the baseline results of the proposed system for cross-modal consistency verification on the *TamperedNews* (Section 4.5.3.1) and *News400* dataset (Section 4.5.3.2). Unfortunately, a comparison to previous work such as Jaiswal et al. [114] or Sabir et al. [219] is not reasonable since these approaches cannot handle the significantly longer news texts and need to be trained with labeled reference data that are much closer related to the source images. As discussed above, these approaches cannot deal with real-world news in contrast to our approach.

## 4.5.3.1 Results on TamperedNews

Qualitative and quantitative results are presented in Table 4.4, Figure 4.4 (page 121), and Figure 4.5 (page 125). Results for all *TamperedNews* documents as well as the Top-25% subset allow for similar conclusions and are reported in Appendix A.3.

Table 4.4: Results for document verification (DV) and collection retrieval for the *Tampered-News (Top-50%)* dataset for different entity test sets (notations according to Section 4.4.1)

Test set	DV		Collection Retrieval					
	VA	AUC	AP-original [%]			AP-manipulated [%]		
			@25%	@50%	@100%	@25%	@50%	@100%
<b>Persons (16,848 documents)</b>								
Random	0.94	0.95	96.08	95.45	92.64	100.0	100.0	96.16
PsC	0.93	0.94	95.53	94.67	91.68	100.0	100.0	95.61
PsG	0.94	0.95	95.77	95.07	92.27	100.0	100.0	96.00
PsCG	0.93	0.94	95.04	94.70	91.70	100.0	100.0	95.56
<b>Locations - Outdoor (14,113 documents)</b>								
Random	0.88	0.85	92.57	88.02	81.71	100.0	100.0	88.82
GCD $_{750}^{2500}$	0.86	0.81	88.04	83.65	77.25	100.0	100.0	85.45
GCD $_{200}^{750}$	0.79	0.74	82.85	76.96	70.56	100.0	96.98	79.38
GCD $_{25}^{200}$	0.77	0.72	80.50	74.23	68.30	100.0	95.19	77.42
<b>Locations - Indoor (19,129 documents)</b>								
Random	0.74	0.72	68.47	66.53	65.34	100.0	99.01	79.62
GCD $_{750}^{2500}$	0.73	0.70	63.62	62.86	62.72	100.0	97.57	77.80
GCD $_{200}^{750}$	0.74	0.71	66.93	65.10	63.97	100.0	97.70	78.14
GCD $_{25}^{200}$	0.69	0.68	55.99	57.74	59.48	100.0	95.97	76.04
<b>Events (7,734 documents)</b>								
Random	0.92	0.91	92.20	91.26	87.61	100.0	100.0	93.66
EsP	0.75	0.71	70.72	67.30	64.92	100.0	96.72	77.68
<b>Context (36,217 documents)</b>								
Random	0.81	0.80	88.95	83.03	76.32	100.0	100.0	84.79
Top-25%	0.78	0.77	83.52	78.12	72.43	100.0	99.70	82.25
Top-10%	0.76	0.74	77.76	73.21	68.78	100.0	98.33	79.84
Top-5%	0.74	0.71	74.31	69.89	66.22	100.0	96.84	77.92

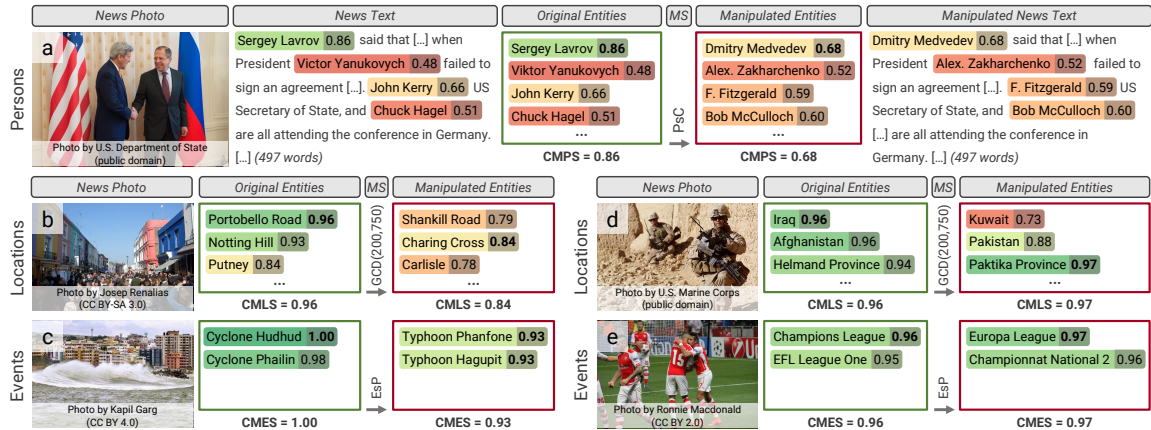


Figure 4.5: Positive (a-c, higher CMS for original entities) and negative (d-e, higher CMS for manipulated entities) verification results of some *TamperedNews* documents. Within each example, the similarities (from red to green with intervals: persons [0.45, 1], locations [0.7, 1], events [0.8, 1]) of the news photo to a set of original and manipulated entities using *one* specific manipulation strategy (MS) are shown. Photos are replaced with similar ones depicting the same entity relations due to image copyright restrictions. Links to the original documents can be found on: [https://github.com/TIBHannover/cross-modal\\_entity\\_consistency/tree/master/supplemental\\_material](https://github.com/TIBHannover/cross-modal_entity_consistency/tree/master/supplemental_material)

**Results for Person Entities:** As expected, person verification achieves the best performance since the entities and the retrieved example material are very unambiguous, and neural networks for face recognition, such as *FaceNet* [228], can achieve impressive results. Despite the more challenging manipulation techniques, our approach is still able to produce similar results. We have only experienced problems if persons were depicted in challenging conditions (e.g., extreme poses as shown in Figure 4.5a for *John Kerry*) or were relatively unknown, which can lead to confusions with other persons (e.g., with a similar name) and consequently false NED results or retrieval of irrelevant example photos.

**Results for Location Entities:** To evaluate the performance for location entities, we distinguish between images in indoor and outdoor environments using the place probabilities  $\hat{y}_S$  according to Section 4.3.2 and the hierarchy provided by the *Places365* dataset [315]. Due to the data diversity, ambiguity, and unequal distribution of photos on Earth, geolocation estimation is a complex problem, as discussed in Section 1.2 and Section 3.2.1. Therefore, the results were expected to be worse compared to the person verification. Despite the complexity, good results were achieved for news with outdoor images, whereas the assessment of modified indoor locations is more challenging given the low amount of geographical cues and their ambiguity. But even when entities are manipulated with locations of similar appearance and low *Great Circle Distance* (GCD), as in Figure 4.5b and Figure 4.5d, the system can operate on a reasonable level and shows promising results.

Table 4.5: AUC for a selection of location types on the *TamperedNews (Top-50%)* subset.  $|\mathbb{D}|$  is the number of documents containing at least one entity of this type and  $|\mathbb{D}_s|$  is the number of times this type achieves the highest cross-modal similarity within the original data. Results are reported for the documents  $\mathbb{D}_s$  of each type.

Selection of 12 / 1,063 location entity types						
Type (num. of entities)	$ \mathbb{D} $	$ \mathbb{D}_s $	AUC			
			Random	$\text{GCD}_{750}^{2500}$	$\text{GCD}_{200}^{750}$	$\text{GCD}_{25}^{200}$
continent (7)	1,510	116	0.82	0.76	0.77	0.78
country (184)	9,516	2,892	0.86	0.76	0.72	0.70
state (109)	1,969	411	0.87	0.75	0.73	0.71
city (706)	9,781	3,333	0.86	0.83	0.78	0.75
town (592)	4,774	1,655	0.80	0.88	0.71	0.69
street (24)	300	64	0.80	0.78	0.76	0.73
tourist attraction (63)	880	172	0.94	0.91	0.90	0.88
building (40)	640	77	0.91	0.87	0.90	0.88
mountain range (13)	201	42	0.94	0.86	0.70	0.62
mountain (9)	85	31	0.97	0.92	0.79	0.77
ocean (4)	344	59	0.89	0.63	0.58	0.63
river (36)	412	106	0.90	0.82	0.78	0.71

Unlike person entities, locations are an instance of various parent classes such as *countries* or *cities*. For an in-depth analysis, we calculate results for different types of locations using the documents  $\mathbb{D}_s$  where an instance of a given type achieves the highest CMLS within the original set of entities. The results for some location types are presented in Table 4.5 and show that the performance is best for more fine-grained entities such as *tourist attractions*, *buildings*, and *cities*. The performance for coarse location types such as *oceans*, *mountain ranges*, and *country states* are typically worse since they do not provide sufficient geographical cues or are too broad to retrieve suitable reference images. Although the results for *continents* or *countries* are also comparatively high, the candidates used for manipulation are easier to distinguish since locations of those types have higher geographical and cultural differences. The manipulations are much more challenging for fine-grained entities, as illustrated in Figure 4.5b and Figure 4.5d.

**Results for Event Entities:** Referring to Table 4.4, good results were achieved for event verification. As for locations, we have provided results of common event types in Table 4.6. While the results for *festivals*, *holiday*, and *disasters* are promising, event types such as *football club competitions*, *protests*, and *wars* are hard to distinguish due to the high visual similarity of events within these types. For example, many news articles on *football club cups* contain images which, unlike articles on *sport competitions* that refer to different types of sports, depict typical scenes (e.g., players on the pitch) of the same sport. Thus, reference

Table 4.6: AUC for a selection of event types on the *TamperedNews (Top-50%)* subset.  $|\mathbb{D}|$  is the number of documents containing at least one entity of this type, and  $|\mathbb{D}_s|$  is the number of times this type achieves the highest cross-modal similarity within the original data. Results are reported for the documents  $\mathbb{D}_s$  of each type.

Selection of 12 / 479 event entity types				
Type (num. of entities)	$ \mathbb{D} $	$ \mathbb{D}_s $	AUC	
			Random	EsP
football club cup (3)	1,094	801	0.93	0.49
sport competition (14)	155	111	0.95	0.76
festival (64)	516	421	0.90	0.77
award (6)	260	206	0.91	0.74
holiday (28)	285	141	0.91	0.84
television series (16)	144	123	0.87	0.70
war (39)	919	665	0.83	0.67
murder (19)	154	134	0.93	0.78
disaster (5)	70	61	0.93	0.91
scandal (10)	112	95	0.95	0.68
protest (7)	60	54	0.89	0.67
legal case (10)	37	34	0.89	0.79

images for the different competitions are very similar. Moreover, the utilized event classification approach [174] presented in Section 3.1 distinguishes between event types such as *football*, *elections*, or types of *natural disasters* rather than between sub-types or concrete event instances such as *UEFA Champions League* or *2020 U.S. elections*. Despite these limitations, the results are superior to the place classification approach used in our previous work [175], as discussed in more detail in Section 4.5.4.

**Cross-modal Context Similarity:** The results for context verification in Table 4.4 indicate that our system can reliably detect documents with randomly changed images. However, as also stated by Sabir et al. [219], this task is relatively easy as the semantic coherence is not maintained. When similar images are used for manipulation, this task becomes much more challenging. Since networks for object classification (used for manipulation) and place classification (used for verification) can produce comparable results, performance steadily decreases using more similar images for manipulation that might even show the same topic, e.g., *sports*. However, our system is still able to hint towards cross-modal consistencies.

#### 4.5.3.2 Results on News400

Since the number of documents is rather limited and the cross-modal mutual presence of entities was manually verified, results for *News400* are reported for all documents with

#### 4 Multimodal Analytics using Measures of Cross-modal Consistency

verified relations. Based on the results reported in Table 4.7, similar conclusions on the overall system performance can be drawn. However, results while retrieving manipulated documents are noticeably worse. This is mainly caused by the fact that some original entities with valid cross-modal relations can be unspecific (e.g., mention of a *country*), or the retrieved images for visual verification do not fit the document’s image content.

Table 4.7: Results for document verification (DV) and collection retrieval for the *News400* dataset. Results are reported for all verified documents for different entity test sets (notations according to Section 4.4.1).

Test set	DV		Collection Retrieval					
	VA	AUC	AP-original [%]			AP-manipulated [%]		
			@25%	@50%	@100%	@25%	@50%	@100%
<b>Persons</b> (116 verified documents)								
Random	0.95	0.91	100.0	100.0	93.70	85.19	85.96	85.95
PsC	0.92	0.90	100.0	99.49	92.14	83.07	84.52	84.14
PsG	0.91	0.90	99.10	98.40	92.34	82.40	84.36	84.64
PsCG	0.92	0.91	100.0	100.0	94.00	84.38	85.25	85.60
<b>Locations - Outdoor</b> (54 verified documents)								
Random	0.89	0.85	100.0	98.01	87.72	83.19	80.76	79.47
GCD $\frac{2500}{750}$	0.81	0.80	92.61	89.94	81.49	72.19	72.95	73.20
GCD $\frac{750}{200}$	0.80	0.74	86.70	82.42	74.76	63.03	66.73	67.33
GCD $\frac{200}{25}$	0.80	0.72	86.70	82.25	72.85	63.59	67.97	66.35
<b>Locations - Indoor</b> (15 verified documents)								
Random	0.80	0.75	91.67	80.94	74.85	88.75	86.44	77.20
GCD $\frac{2500}{750}$	0.67	0.64	62.20	58.74	60.04	80.42	82.28	69.37
GCD $\frac{750}{200}$	0.87	0.69	85.42	74.31	68.90	88.75	85.23	72.96
GCD $\frac{200}{25}$	0.80	0.62	69.17	62.64	61.40	80.42	78.06	67.12
<b>Events</b> (31 verified documents)								
Random	1.00	0.93	100.0	96.18	92.58	100.0	99.63	93.93
EsP	0.74	0.72	63.44	66.19	65.49	89.57	86.45	74.76
<b>Context</b> (91 verified documents)								
Random	0.70	0.70	87.03	87.50	73.62	61.11	63.09	63.19
Top-25%	0.70	0.68	92.19	88.43	72.96	53.60	57.77	59.69
Top-10%	0.64	0.66	70.54	74.12	65.58	56.15	59.72	59.75
Top-5%	0.66	0.63	74.48	73.09	64.18	50.77	55.99	56.98

This problem was bypassed because subsets of the Top- $d\%$  documents for *TamperedNews* were used to counteract the influence of original documents that do not show any cross-modal relations (as discussed in Section 4.5.1, paragraph "Test Document Selection for TamperedNews"). We have verified the same behavior for *News400* when experimenting on these subsets. For more details, we refer to Appendix A.3. In addition, performance for context verification is worse compared to *TamperedNews*. We assume that this is due to the less powerful word embedding for the *German* language. Overall, the system achieves promising performance for cross-modal consistency verification. Since it dynamically gathers example data from the Web, it is robust to changes in topics and entities, even when applied to news articles from another country and publication date.

#### 4.5.4 Comparison of Event Feature Descriptors

As discussed in Section 4.2.2, we use the ontology-driven event classification approach [174] presented in Section 3.1 to compute event features for our proposed system. Due to the absence of suitable methods for event classification, a more general place classification model was applied in our previous approach [175]. The visual features  $\mathbf{f}_E$  are obtained from the last pooling layer of a *ResNet-50* model<sup>28</sup> [92, 93] trained on 365 place categories covered by the *Places365* dataset [315].

To compare both approaches, we evaluate their performances on the *News400* dataset as it contains documents with verified event relations. As explained in Section 4.5.1 (paragraph "Test Document Selection for TamperedNews"), we have used the *TamperedNews* (*Top-50%*) documents as subsets for evaluation since they more likely contain cross-modal relations. However, this complicates the comparison of two approaches as those subsets can be different depending on their specified parameters (e.g., feature descriptor, function  $\Psi$ ). Thus, we report results on all documents as well as on the intersection and union of the *TamperedNews* (*Top-50%*) document sets of both approaches. In this way, the test sets contain documents that are either considered relevant from both or at least one approach, respectively. The results are presented in Table 4.8 and demonstrate that the event classification approach achieves superior performances. However, as already discussed in Section 4.5.3.1, the approach is not trained to classify concrete event instances and instead focuses on more generic event types. As a consequence, improvements for the EsP test set containing manipulated events of the same parent class are not as significant as for the randomly manipulated test set. Further limitations and dependencies are discussed in the next section.

#### 4.5.5 Limitations and Dependencies

News covered in the World Wide Web are dynamic, and new entities and topics evolve every day. We have deliberately chosen *Wikifier* for *Named Entity Disambiguation* (NED) because

#### 4 Multimodal Analytics using Measures of Cross-modal Consistency

Table 4.8: *Verification Accuracy* (VA) and AUC using a place classification network trained on *Places365* [315] and the ontology-driven event classification approach trained on VisE-D according to Section 3.1 as a feature extractor for different event manipulation strategies and datasets.

Feature Extractor	Event Manipulation Technique			
	Random		EsP	
	VA	AUC	VA	AUC
<b>News400</b> (31 verified documents)				
Place CNN trained on <i>Places365</i> [315]	0.87	0.85	0.68	0.64
Event CNN $CO_{\gamma}^{cos}$ trained on VisE-D (Section 3.1)	<b>1.00</b>	<b>0.93</b>	<b>0.74</b>	<b>0.72</b>
<b>TamperedNews</b> (all 15,467 documents)				
Place CNN trained on <i>Places365</i> [315]	0.67	0.66	0.59	0.56
Event CNN $CO_{\gamma}^{cos}$ trained on VisE-D (Section 3.1)	<b>0.70</b>	<b>0.70</b>	<b>0.59</b>	<b>0.57</b>
<b>TamperedNews</b> (Top-50% intersection - 6,080 documents)				
Place CNN trained on <i>Places365</i> [315]	0.91	0.89	0.75	0.70
Event CNN $CO_{\gamma}^{cos}$ trained on VisE-D (Section 3.1)	<b>0.94</b>	<b>0.93</b>	<b>0.76</b>	<b>0.71</b>
<b>TamperedNews</b> (Top-50% union - 9,388 documents)				
Place CNN trained on <i>Places365</i> [315]	0.83	0.82	0.70	0.65
Event CNN $CO_{\gamma}^{cos}$ trained on VisE-D (Section 3.1)	<b>0.86</b>	<b>0.86</b>	<b>0.71</b>	<b>0.67</b>

it can dynamically cover *Wikipedia* entities. However, the proposed system is restricted to entities that exist in a knowledge base at the time of inference. Besides, the system relies on the rankings and response times of image search engines. In this regard, the reference images for coarse entities such as *countries* or *continents* crawled from the Web might not match the news image. Some named entities such as "Hanover" (German or U.S. city) or "Tesla" (company or inventor) can also be ambiguous. Referring to Figure 4.1, we also noticed that querying entities such as the city "Liverpool" using Google's image search engine retrieves images that depict another (more popular) entity, in this case, the football club "Liverpool F.C." rather than the actual entity.

A potential solution to the aforementioned problems is to include knowledge graph information and relations that are already extracted by the system. For example, adding the country (*Wikidata* property *P17*) "Germany" to the query "Hanover" (*Wikidata* identifier *Q1715*) or using the entity type (*P31*) "city" in combination with the query "Liverpool" (*Q24826*) can prevent potential ambiguities. Since reference images for coarse location entities rarely match the news image, the classification output of the geolocation approach presented in Section 3.2 might be a better alternative to quantify the CMLS due to its strong



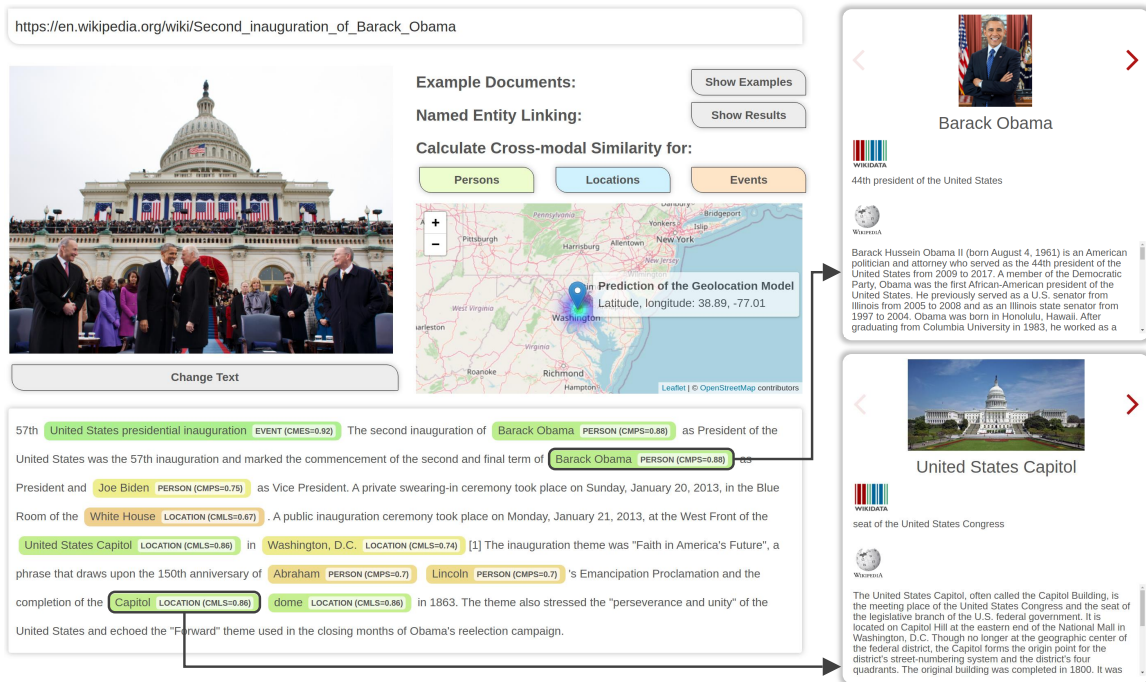


Figure 4.6: Screenshot of the demonstrator for news analytics for an exemplary multimodal article. Entity information (right) is revealed by hovering over specific entities.

performance at the country and continent level. This approach could also be combined with a large-scale reference dataset covering locations around the Earth for verification, as proposed by Vo et al. [266, 267]. Although the feasibility of this solution needs to be evaluated in the future, we have already integrated the classification outputs of the geolocation model into our demonstrator, as presented in the next section.

## 4.6 Demonstrator

We have developed a demonstrator of the proposed system for multimodal news analytics [238]. The demonstrator is publicly available as a lab service of the *Leibniz Information Centre for Science and Technology* (TIB) at <https://labs.tib.eu/newsanalytics>. A screenshot to illustrate the functionality is shown in Figure 4.6.

The demonstrator allows users to copy the web link of a news article or alternatively upload an image and text to analyze the multimodal content for cross-modal entity occurrences. Once an image-text pair has been selected, NER & NED are applied according to Section 4.2.1 to extract mentions of persons, locations, and events. By hovering over specific entities, the user is provided with entity descriptions extracted from *Wikipedia* and *Wikidata* as well as an image, if available, using the *Wikidata* property *P18*. After the named entities are extracted from the text, the user can click on the corresponding buttons to let

the system calculate the proposed *Cross-modal Similarity* (CMS) for persons, locations, and events according to Section 4.2. Subsequently, the results are presented, and the crawled web images for the named entities can be displayed by hovering over them. Furthermore, the output of the geolocation approach presented in Section 3.2 is visualized on a world map. These illustrations also allow the user to assess the cross-modal consistency of news articles manually.

## 4.7 Summary

In this chapter, we have presented a novel analytics system and benchmark datasets to measure the cross-modal consistency between photos and text in real-world news articles. Tools for *Named Entity Recognition and Disambiguation* (NER & NED) are combined to detect persons, locations, and events in the news text. Example images for these entities are automatically gathered from the Web and used in combination with novel measures of cross-modal similarity to quantify the entity consistency between photos and text. For this purpose, suitable computer vision methods for information extraction from photos are applied. Furthermore, a more general measure to evaluate the cross-modal context consistency between photo and text has been introduced. Unlike previous solutions that quantify image-text relations [96, 127, 185, 294, 306], we explicitly consider cross-modal relations of named entities to provide more differentiated measures of *Cross-modal Mutual Information* (CMI). In contrast to supervised approaches on image repurposing detection [114, 115, 219], our system is unsupervised and does not rely on labeled training or pre-defined reference data. Thus, it is applicable to real-world news since it can better cope with the growing amount and variety of entities. Experiments were conducted on two datasets that contain real-world news articles across different topics, domains, and languages and have demonstrated the feasibility of the proposed approach.

As mentioned in Section 4.5.5, the system performance for coarse (e.g., *countries* or *continents*), ambiguous, or less popular entities can suffer due to the lack of relevant reference images crawled by the automatic web image search. Thus, we aim to refine the image search queries based on knowledge graph information and entity relationships in the future. Furthermore, the event classification approach (Section 3.1) can only distinguish *event types* such as types of *sports*, *natural disasters*, or *elections*. The system can benefit from an event classification approach capable of differentiating between more fine-grained event types and concrete *event instances*, e.g., *UEFA Champions League* or *2020 U.S. elections*. Another research direction is to measure the cross-modal consistency of other types of entities such as dates, times, or organizations.

## 5 Conclusions

News articles are an essential part of our everyday lives. Articles published on news sites, social media, or in the newspaper typically use different modalities, such as photos and text, to convey information more effectively or to attract attention. The relations between the modalities, such as the number of shared entities and the semantic correlation, are an important aspect to understand the overall message and meaning of multimodal documents [31].

In this thesis, the question has been investigated whether an unsupervised approach can automatically quantify the cross-modal consistency of named entities, e.g., persons, locations, dates, and events between photos and text in real-world news articles. This functionality enables many tasks and applications. For example, it allows for efficient exploration of news and facilitates semantic search and multimedia retrieval in large (web) archives. In some use cases, cross-modal consistency measures can assist users and manual fact-checking efforts in assessing the credibility of news, which is an important task given the increasing amount of misinformation, i.e., *fake news*, published on the World Wide Web.

We have proposed a novel system that quantifies the cross-modal consistency of named entities detected by suitable methods for *Named Entity Recognition and Disambiguation* in news articles. Quantification has been realized by comparing information extracted from the news photo to example photos for these named entities crawled automatically from the Web. As this step requires the extraction of semantic, geospatial, temporal, and spatio-temporal information from photos, novel approaches for the related computer vision tasks of event classification, geolocation estimation, date estimation, and face recognition have been suggested. To the best of our knowledge, we have proposed the first unsupervised solution to quantify the cross-modal consistency of entities that is applicable to real-world news. Unlike previous approaches that quantify image-text relations without explicitly considering named entities [96, 127, 185, 294, 306], this approach allows for more differentiated measures of *Cross-modal Mutual Information* with respect to entity consistency. Compared to supervised multimodal deep learning approaches for the detection of repurposed image content based on entity verification [114, 115, 219], it does not require labeled training or pre-defined reference data and can better cope with the growing amount and variety of entities in the media.

In the following, we summarize the contributions and provide answers to the research questions of this thesis (Section 5.1). Section 5.2 discusses the limitations of the proposed solutions and directions of future work.

## 5.1 Answers to the Research Questions

This section provides answers to the three research questions addressed in this thesis.

**Research Question 1:** *Can we develop an unsupervised approach for the quantification of cross-modal entity consistency in news articles? What are the advantages, limitations, and challenges in comparison to supervised approaches?*

We have proposed and evaluated an unsupervised approach for the quantification of cross-modal entity consistency in real-world news articles (Chapter 4). As mentioned above, we applied *Named Entity Recognition and Disambiguation* to extract named entities such as events, locations, and persons from the text. For each detected entity, example images were crawled automatically from the Web and served as visual evidence. Depending on the entity type, features were extracted using the proposed solutions for event classification, geolocation estimation, and person recognition presented in Chapter 3. These features were compared to the news photo based on novel measures of cross-modal similarity to quantify the cross-modal entity consistency. Moreover, semantic features from a place (or scene) classification approach have been extracted and used with word embeddings to measure contextual relationships between image and text. Finally, a publicly available web demonstrator of the system was presented. Experimental results on novel datasets for different tasks, domains, and languages (English and German) have shown that the approach can determine multimodal relationships between photos and text when appropriate example images for the entities can be acquired. In particular, promising results have been achieved for quantifying the cross-modal consistency of persons and fine-grained geographical locations such as *tourist attractions, buildings, streets, and cities*. Thus, we have successfully presented a first solution towards the quantification of cross-modal entity consistency in real-world news.

The main advantage of our proposed approach over supervised solutions is that it does not rely on pre-defined training or reference datasets labeled for cross-modal relations that (1) are difficult and tedious to annotate and (2) restrict these learning-based approaches to the verification of entities already covered in these datasets. Our unsupervised solution can provide differentiated measures of cross-modal entity consistency for a vast amount and diversity of entities appearing in the news. However, the performance relies on the automatic retrieval of suitable example images for the named entities from the Web. As a consequence, the system depends on the rankings and response times of the used image search engines. In particular, retrieved images for coarse, less popular, or ambiguous entities can be irrelevant or depict the wrong entities. For example, querying image search engines based on the name of a certain *country* or *continent* is typically too broad. Thus, the retrieved photos usually do not reflect the content in the news photo and are therefore not valuable for quantifying cross-modal relations. Although suitable deep learning approaches were applied to extract features for different entity types from the photos, the complexity of the individual tasks

limits the applications of the proposed solution to a certain degree, as discussed in the answers to the following research question.

**Research Question 2:** *How suitable are deep learning approaches in recognizing events, locations, dates, and persons in photos specifically with respect to information extraction from news articles?*

The extraction of information from photos is a vital prerequisite to quantify cross-modal entity consistency. Thus, novel deep learning solutions for event classification, geolocation estimation, date estimation, and person recognition have been proposed in Chapter 3. These approaches (except for date estimation) were used to quantify the entity consistency between photos and text in Chapter 4. The answers to the research question for each entity type are provided below.

**Events:** In Section 3.1, an ontology-driven deep learning approach for event type classification in photos has been introduced. We presented the *Visual Event Classification Dataset* (VisE-D) comprising 570,540 images that, unlike previous datasets, covers the majority of event classes important to news. Besides, we have proposed a *Visual Event Ontology* (VisE-O) based on *Wikidata* knowledge base information that contains relations for a total of 148 event types. Several loss functions and weighting schemes have been suggested to integrate event relations from structured knowledge graph information into an ontology-driven deep learning approach. Results on several benchmarks and two novel test datasets have shown that the integration of structured information from an ontology improves event classification in photos. We noticed that the performance for expected (scheduled or regular) event types, such as elections and sport-centric events, is better than for unexpected or rare events, e.g., natural disasters. Experimental results in Chapter 4 have confirmed that the approach can generally quantify cross-modal event consistency in news. However, news articles typically mention concrete event instances like specific sports competitions (e.g., *UEFA Champions League*, *Premier League*), elections (e.g., *2016 U.S. election*, *2020 U.S. election*), or epidemics (e.g., *2014-15 Ebola epidemic*, *COVID-19 pandemic*) that are difficult to distinguish by the proposed solution as they are subordinate to the event types.

**Locations:** Novel deep learning approaches for planet-scale photo geolocation estimation have been introduced in Section 3.2. We have proposed a hierarchical approach that uses geographical information from partitionings of the Earth with varying granularity. It helps to measure the cross-modal consistency of locations at different geographical levels (e.g., *street*, *city*, *country*), which are frequently used in the news. Moreover, it addresses a critical trade-off problem where a higher number of cells leads to a more fine-grained partitioning of the Earth but results in fewer training images per cell, making a model more susceptible to overfitting. Besides, we have suggested two strategies to include contextual scene infor-

## 5 Conclusions

mation into a geolocation estimation approach to learn important features for photos taken in different environmental settings, e.g., *urban*, *natural*, or *indoor*. It has been shown that individual CNNs trained with images depicting a particular environmental setting along with partitionings of different granularity yield more accurate classifiers for geolocation estimation. The proposed solution achieves state-of-the-art performance on two benchmark datasets and outperforms strong baselines from the literature while relying on a significantly smaller amount of training images. Given the complexity of geolocation estimation, impressive results have been reported for localizing photos that depict *tourist attractions*, *buildings*, *streets*, and *cities* in Section 3.2 and Chapter 4. Geolocation estimation of photos taken in natural environments is also promising, yet it faces more challenges which lead to slightly less accurate results. On the other hand, close-ups, stock images, or photos of indoor environments are often demanding to localize since they lack unique geographical cues or could even be misleading. Moreover, the geospatial distribution of photos covered in public datasets is biased. Fewer photographs are available for continents such as Africa and South America compared to Europe and North America. As a result, it is more difficult to precisely estimate the location of photos taken in less frequently captured regions. However, recent studies [267] have already shown that deep learning approaches exceed human performance for this task. Overall, the proposed solution can reliably recognize locations from photos that depict unique and unambiguous geographical cues.

**Dates:** Related work on date estimation is restricted to historical color photographs or specific concepts such as *persons* or *cars*. For this reason, a novel large-scale dataset entitled *Date Estimation in the Wild* with more than one million photos from *Flickr* captured between 1930 and 1999 that is neither restricted to specific concepts nor to color photographs has been introduced in Section 3.3. Two deep learning approaches that treat date estimation as a regression and classification problem have been proposed. Both systems outperform untrained human annotators with an average error of less than eight years. Although the approaches exceed human performance, the relatively high average error indicates that the approach cannot reliably verify the acquisition year of photos in news articles. Furthermore, the system can only predict the acquisition year of photos taken between 1930 and 1999. Overall, this limits the current approach to the verification of dates in articles about historical events with a precision of five to ten years. Many real-world use cases require the exact date of capture, including precise date and time of day information, from contemporary photographs. For this reason, experiments on cross-modal date verification were omitted in this thesis and will be investigated in the future.

**Persons:** In Section 3.4 of this thesis, an approach for the identification of public figures in news photos extracted from the *Internet Archive* has been presented. Example images for relevant persons in a given time period and domain (e.g., politics or entertainment) were

crawled automatically from the Web. Several strategies have been investigated to exclude photos that do not depict the queried personalities, including a clustering technique that allows for a fully-automatic identification of public figures in news images. The identification has been realized by comparing facial features extracted from the news photo and the filtered example images. Results and case studies have demonstrated that the suggested approach can extract meaningful relations between public figures in news images. It also turned out that the quantification of person consistency in news articles provides the best results across all entity types (Chapter 4). We only experienced issues for relatively unknown personalities since the example images crawled from the Web might not depict them. In some cases, pose-variations, occlusions, and aging can be very challenging when quantifying the cross-modal consistency of individuals. However, in summary, deep learning approaches are capable of identifying persons in news photos.

**Research Question 3:** *Can contextual information, derived from knowledge bases or related tasks like scene classification, improve image recognition and interpretation and provide better performance for computer vision tasks?*

Since event classification and geolocation estimation are very challenging tasks that require a profound scene understanding, we suggested novel techniques to provide deep learning approaches with contextual information. It turned out that the proposed ontology-driven event classification approach (Section 3.1), which exploits event relations extracted from structured knowledge graph information, achieved the best results on several datasets for event recognition in photos. In Section 3.2, it has been demonstrated that hierarchical geographic information from partitionings of the Earth with different granularity in combination with complementary contextual scene information about the environmental setting achieves the best results for geolocation estimation. In summary, experimental results have shown that the integration of contextual information into deep learning approaches improves event classification and geolocation estimation, leading to the conclusion that contextual information indeed improves image recognition and interpretation.

## 5.2 Limitations & Future Work

In this thesis, a first unsupervised approach for the quantification of entity consistency in real-world news has been presented. Several challenges and limitations can be the subject of future work.

**Named Entity Recognition and Disambiguation:** We have used *Wikifier* [40, 41] for *Named Entity Recognition and Disambiguation* as it can dynamically detect entities in the text covered in *Wikipedia* at inference time. Thus, it is more flexible than learning-based approaches, which are limited to the entities mentioned in the training data and require

## 5 Conclusions

fine-tuning to adapt to new entities. However, despite its flexibility, it cannot handle new entities without entries in the knowledge base that may appear in the news every day, such as unexpected events. A potential solution is to additionally consider named entities from *Named Entity Recognition* (NER) tools such as *spaCy* [99] that were not linked to a knowledge base. However, this would introduce new challenges. For example, it significantly increases the computational time since each mention of an entity, even if it represents the same entity (e.g., *President Obama* and *Barack Obama*), would be considered individually.

**Example Image Retrieval:** As discussed in Section 5.1, the system relies on suitable example photos for named entities detected in the text. A potential solution to find more meaningful example images is to refine image search queries. This can be achieved in numerous ways. For example, knowledge graph information (e.g., parent classes indicating the entity type, location coordinates, event dates) or contextual information, such as the topic of the news article, can be included in the query. Rather than crawling example images for specific entities, it might be worth investigating whether images from news articles that mention the same entity (or even entities) in a similar context can be retrieved. These images more likely reflect the image content of the investigated news article. Moreover, we found that queries for broader locations, such as *continents*, *countries*, and (in some cases) even *cities*, are not specific enough to retrieve valuable example photos. Since these location entities are relatively static, i.e., there are rarely new cities or countries mentioned in the media, a pre-defined reference dataset covering most relevant locations around the world might provide more meaningful example images. This direction has already been pursued by Vo et al. [267] and proven to be efficient outside the news domain.

**Cross-modal Entity Consistency:** While the proposed solution can quantify the cross-modal consistency of events, locations, and persons, the verification of other entity types such as times (e.g., decades, dates, daytime) or organizations remains an open issue. An approach that can estimate complete image dates (day, month, and year) and daytime information without any restrictions would allow for the verification of temporal information in contemporary news articles. However, this is challenging because the model needs to incorporate geographic, economic, and cultural information that influences temporal characteristics. As an alternative, the proposed news analytics system would greatly benefit from an event classification approach that is capable of identifying concrete event instances (e.g., *FIFA World Cup Final 2014*, *2020 U.S. election*, *COVID-19 pandemic*) rather than more generic event types (e.g., *association football*, *election*, *epidemic*). Such an approach would enable the quantification of image-text relations for more fine-grained events as well as their associated dates (or time periods) and location(s). In this regard, we plan to exploit strategies such as *Graph Convolutional Neural Network* that leverage the proposed *Visual Event Ontology* (VisE-O). This ontology already provides a solid foundation since it covers more



than 500 thousand real-world events. In Section 3.2, it was demonstrated that incorporating contextual information on the environmental setting and hierarchical geospatial information into deep learning models improves geolocation estimation. It would be interesting to investigate how additional information on aspects such as culture, climate, or economy can impact the performance of CNNs for this task. Moreover, the consideration of human-interpretable aspects allows for more plausible predictions from deep learning approaches that are usually black-box systems. In general, the interpretability of results generated by the proposed deep learning approaches is another exciting research direction given the potential impact of misinformation in news and social media. Thus, we plan to explore approaches on explainable artificial intelligence to provide users with plausible and interpretable system outputs for cross-modal entity consistency in real-world news.



# References

## Own Publications

- [42] Andreas Breitbarth, Eric Müller, Peter Kühmstedt, Gunther Notni, and Joachim Denzler. “Phase unwrapping of fringe images for dynamic 3D measurements without additional pattern projection”. In: *Dimensional Optical Metrology and Inspection for Practical Applications IV*. Ed. by Kevin G. Harding and Toru Yoshizawa. Vol. 9489. International Society for Optics and Photonics. SPIE, 2015, pp. 8–17. URL: <https://doi.org/10.1117/12.2176822>.
- [49] Gullal S. Cheema, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth. “A Fair and Comprehensive Comparison of Multimodal Tweet Sentiment Analysis Methods”. In: *MMPT@ICMR2021: Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding, Taipei, Taiwan, August 21, 2021*. Ed. by Bei Liu, Jianlong Fu, Shizhe Chen, Qin Jin, Alexander G. Hauptmann, and Yong Rui. ACM, 2021, pp. 37–45. DOI: 10.1145/3463945.3469058.
- [50] Gullal S. Cheema, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth. “On the Role of Images for Analyzing Claims in Social Media”. In: *Proceedings of the 2nd International Workshop on Cross-lingual Event-centric Open Analytics co-located with the 30th The Web Conference (WWW 2021), Ljubljana, Slovenia, April 12, 2021 (online event due to COVID-19 outbreak)*. Ed. by Elena Demidova, Sherzod Hakimov, Jane Winters, and Marko Tadic. Vol. 2829. CEUR Workshop Proceedings. CEUR-WS.org, 2021, pp. 32–46. URL: <http://ceur-ws.org/Vol-2829/paper3.pdf>.
- [68] Ralph Ewerth, Christian Otto, and Eric Müller-Budack. “Computational Approaches for the Interpretation of Image-text Relations”. In: *Empirical Multimodality Research: Methods, Evaluations, Implications*. Ed. by Jana Pflaeging, Janina Wildfeuer, and John A. Bateman. De Gruyter, 2021, pp. 109–138. DOI: 10.1515/9783110725001-005. URL: <https://doi.org/10.1515/9783110725001-005>.
- [69] Ralph Ewerth, Matthias Springstein, Eric Müller, Alexander Balz, Jan Gehlhaar, Tolga Naziyok, Krzysztof Dembczynski, and Eyke Hüllermeier. “Estimating relative depth in single images via rankboost”. In: *2017 IEEE International Conference on Multimedia and Expo, ICME 2017, Hong Kong, China, July 10-14, 2017*. IEEE Computer Society, 2017, pp. 919–924. DOI: 10.1109/ICME.2017.8019434.
- [168] David Morris, Eric Müller-Budack, and Ralph Ewerth. “SlideImages: A Dataset for Educational Image Classification”. In: *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II*. Ed. by Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins. Vol. 12036. Lecture Notes in Computer Science. Springer, 2020, pp. 289–296. DOI: 10.1007/978-3-030-45442-5\_36. URL: [https://doi.org/10.1007/978-3-030-45442-5\\_36](https://doi.org/10.1007/978-3-030-45442-5_36).

- [169] Markus Mühling, Nikolaus Korfhage, Eric Müller, Christian Otto, Matthias Springstein, Thomas Langelage, Uli Veith, Ralph Ewerth, and Bernd Freisleben. “Deep learning for content-based video retrieval in film and television production”. In: *Multim. Tools Appl.* 76.21 (2017), pp. 22169–22194. DOI: 10.1007/s11042-017-4962-9.
- [171] Eric Müller-Budack, Kader Pustu-Iren, Sebastian Diering, and Ralph Ewerth. “Finding Person Relations in Image Data of News Collections in the Internet Archive”. In: *Digital Libraries for Open Knowledge, 22nd International Conference on Theory and Practice of Digital Libraries, TPDL 2018, Porto, Portugal, September 10-13, 2018, Proceedings*. Ed. by Eva Méndez, Fabio Crestani, Cristina Ribeiro, Gabriel David, and João Correia Lopes. Vol. 11057. Lecture Notes in Computer Science. Springer, 2018, pp. 229–240. DOI: 10.1007/978-3-030-00066-0\_20. URL: [https://doi.org/10.1007/978-3-030-00066-0\\_20](https://doi.org/10.1007/978-3-030-00066-0_20).
- [172] Eric Müller-Budack, Kader Pustu-Iren, Sebastian Diering, Matthias Springstein, and Ralph Ewerth. “Image Analytics in Web Archives”. In: *The Past Web: Exploring Web Archives*. Ed. by Daniel Gomes, Elena Demidova, Jane Winters, and Thomas Risse. Cham: Springer International Publishing, 2021, pp. 141–151. ISBN: 978-3-030-63291-5. DOI: 10.1007/978-3-030-63291-5\_11. URL: [https://doi.org/10.1007/978-3-030-63291-5\\_11](https://doi.org/10.1007/978-3-030-63291-5_11).
- [173] Eric Müller-Budack, Kader Pustu-Iren, and Ralph Ewerth. “Geolocation Estimation of Photos Using a Hierarchical Model and Scene Classification”. In: *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*. Ed. by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss. Vol. 11216. Lecture Notes in Computer Science. Springer, 2018, pp. 575–592. DOI: 10.1007/978-3-030-01258-8\_35. URL: [https://doi.org/10.1007/978-3-030-01258-8\\_35](https://doi.org/10.1007/978-3-030-01258-8_35).
- [174] Eric Müller-Budack, Matthias Springstein, Sherzod Hakimov, Kevin Mrutzek, and Ralph Ewerth. “Ontology-driven Event Type Classification in Images”. In: *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*. IEEE, 2021, pp. 2927–2937. DOI: 10.1109/WACV48630.2021.00297.
- [175] Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. “Multimodal Analytics for Real-world News using Measures of Cross-modal Entity Consistency”. In: *Proceedings of the 2020 on International Conference on Multimedia Retrieval, ICMR 2020, Dublin, Ireland, June 8-11, 2020*. Ed. by Cathal Gurrin, Björn Þór Jónsson, Noriko Kando, Klaus Schöffmann, Yi-Ping Phoebe Chen, and Noel E. O’Connor. ACM, 2020, pp. 16–25. DOI: 10.1145/3372278.3390670.
- [176] Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, Sherzod Hakimov, and Ralph Ewerth. “Multimodal news analytics using measures of cross-modal entity and context consistency”. In: *Int. J. Multim. Inf. Retr.* 10.2 (2021), pp. 111–125. DOI: 10.1007/s13735-021-00207-4.
- [177] Eric Müller-Budack, Jonas Theiner, Robert Rein, and Ralph Ewerth. “"Does 4-4-2 exist?" - An Analytics Approach to Understand and Classify Football Team Formations in Single Match Situations”. In: *Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports, MMSports@MM 2019, Nice, France, October 25, 2019*. Ed. by Rainer Lienhart, Thomas B. Moeslund, and Hideo Saito. ACM, 2019, pp. 25–33. DOI: 10.1145/3347318.3355527.

- [178] Eric Müller, Christian Otto, and Ralph Ewerth. “Semi-supervised Identification of Rarely Appearing Persons in Video by Correcting Weak Labels”. In: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ICMR 2016, New York, New York, USA, June 6-9, 2016*. Ed. by John R. Kender, John R. Smith, Jiebo Luo, Susanne Boll, and Winston H. Hsu. ACM, 2016, pp. 381–384. DOI: 10.1145/2911996.2912073.
- [179] Eric Müller, Matthias Springstein, and Ralph Ewerth. “"When Was This Picture Taken?" - Image Date Estimation in the Wild”. In: *Advances in Information Retrieval - 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings*. Ed. by Joemon M. Jose, Claudia Hauff, Ismail Sengör Altingövde, Dawei Song, Dyaa Albakour, Stuart N. K. Watt, and John Tait. Vol. 10193. Lecture Notes in Computer Science. 2017, pp. 619–625. DOI: 10.1007/978-3-319-56608-5\_57. URL: [https://doi.org/10.1007/978-3-319-56608-5\\_57](https://doi.org/10.1007/978-3-319-56608-5_57).
- [195] Kader Pustu-Iren, Eric Müller-Budack, Sherzod Hakimov, and Ralph Ewerth. “Visualizing Copyright-Protected Video Archive Content Through Similarity Search”. In: *Linking Theory and Practice of Digital Libraries - 25th International Conference on Theory and Practice of Digital Libraries, TPDL 2021, Virtual Event, September 13-17, 2021, Proceedings*. Ed. by Gerd Berget, Mark Michael Hall, Daniel Brenn, and Sanna Kumpulainen. Vol. 12866. Lecture Notes in Computer Science. Springer, 2021, pp. 123–127. DOI: 10.1007/978-3-030-86324-1\_15. URL: [https://doi.org/10.1007/978-3-030-86324-1\\_15](https://doi.org/10.1007/978-3-030-86324-1_15).
- [238] Matthias Springstein, Eric Müller-Budack, and Ralph Ewerth. “QuTI! Quantifying Text-Image Consistency in Multimodal Documents”. In: *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. Ed. by Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai. ACM, 2021, pp. 2575–2579. DOI: 10.1145/3404835.3462796.
- [239] Matthias Springstein, Eric Müller-Budack, and Ralph Ewerth. “Unsupervised Training Data Generation of Handwritten Formulas using Generative Adversarial Networks with Self-Attention”. In: *MMPT@ICMR2021: Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding, Taipei, Taiwan, August 21, 2021*. Ed. by Bei Liu, Jianlong Fu, Shizhe Chen, Qin Jin, Alexander G. Hauptmann, and Yong Rui. ACM, 2021, pp. 46–54. DOI: 10.1145/3463945.3469059.
- [253] Golsa Tahmasebzadeh, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth. “A Feature Analysis for Multimodal News Retrieval”. In: *Proceedings of the 1st International Workshop on Cross-lingual Event-centric Open Analytics co-located with the 17th Extended Semantic Web Conference (ESWC 2020), Heraklion, Crete, Greece, June 3, 2020 (online event due to COVID-19 outbreak)*. Ed. by Elena Demidova, Sherzod Hakimov, Jane Winters, and Marko Tadic. Vol. 2611. CEUR Workshop Proceedings. CEUR-WS.org, 2020, pp. 43–56. URL: <http://ceur-ws.org/Vol-2611/paper4.pdf>.
- [254] Golsa Tahmasebzadeh, Endri Kacupaj, Eric Müller-Budack, Sherzod Hakimov, Jens Lehmann, and Ralph Ewerth. “GeoWINE: Geolocation based Wiki, Image, News and Event Retrieval”. In: *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. Ed. by Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie

Jones, and Tetsuya Sakai. ACM, 2021, pp. 2565–2569. DOI: 10 . 1145 / 3404835 . 3462786.

- [257] Jonas Theiner, Eric Müller-Budack, and Ralph Ewerth. “Interpretable Semantic Photo Geolocation”. In: *IEEE Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 4-8, 2022*. IEEE, 2022, pp. 750–760.

## Other Publications

- [1] Martín Abadi et al. “TensorFlow: A System for Large-Scale Machine Learning”. In: *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016*. Ed. by Kimberly Keeton and Timothy Roscoe. USENIX Association, 2016, pp. 265–283. URL: <https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi>.
- [2] Kashif Ahmad and Nicola Conci. “How Deep Features Have Improved Event Recognition in Multimedia: A Survey”. In: *ACM Trans. Multim. Comput. Commun. Appl.* 15.2 (2019). CLIP, 39:1–39:27. DOI: 10.1145/3306240.
- [3] Kashif Ahmad, Nicola Conci, Giulia Boato, and Francesco G. B. De Natale. “USED: a large-scale social event detection dataset”. In: *Proceedings of the 7th International Conference on Multimedia Systems, MMSys 2016, Klagenfurt, Austria, May 10-13, 2016*. Ed. by Christian Timmerer. ACM, 2016, 50:1–50:6. DOI: 10.1145/2910017.2910624.
- [4] Kashif Ahmad, Nicola Conci, and Francesco G. B. De Natale. “A saliency-based approach to event recognition”. In: *Signal Process. Image Commun.* 60 (2018), pp. 42–51. DOI: 10.1016/j.image.2017.09.009.
- [5] Kashif Ahmad, Mohamed Lamine Mekhalfi, Nicola Conci, Giulia Boato, Farid Melgani, and Francesco G. B. De Natale. “A pool of deep models for event recognition”. In: *2017 IEEE International Conference on Image Processing, ICIP 2017, Beijing, China, September 17-20, 2017*. IEEE, 2017, pp. 2886–2890. DOI: 10.1109/ICIP.2017.8296810.
- [6] Kashif Ahmad, Mohamed Lamine Mekhalfi, Nicola Conci, Farid Melgani, and Francesco G. B. De Natale. “Ensemble of Deep Models for Event Recognition”. In: *ACM Trans. Multim. Comput. Commun. Appl.* 14.2 (2018), 51:1–51:20. DOI: 10.1145/3199668.
- [7] Farah Ahmed, Fouad Khelifi, Ashref Lawgaly, and Ahmed Bouridane. “Temporal Image Forensic Analysis for Picture Dating with Deep Learning”. In: *2020 International Conference on Computing, Electronics Communications Engineering (iCCECE)*. 2020, pp. 109–114. DOI: 10.1109/iCCECE49321.2020.9231160.
- [8] Unaiza Ahsan, Chen Sun, James Hays, and Irfan A. Essa. “Complex Event Recognition from Images with Few Training Examples”. In: *2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017, Santa Rosa, CA, USA, March 24-31, 2017*. IEEE Computer Society, 2017, pp. 669–678. DOI: 10.1109/WACV.2017.80.
- [9] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. “FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*. Ed. by Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh. Association for Computational Linguistics, 2019, pp. 54–59. DOI: 10.18653/v1/n19-4010.

- [10] Alan Akbik, Tanja Bergmann, and Roland Vollgraf. “Pooled Contextualized Embeddings for Named Entity Recognition”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, 2019, pp. 724–728. DOI: 10.18653/v1/n19-1078.
- [11] Alan Akbik, Duncan Blythe, and Roland Vollgraf. “Contextual String Embeddings for Sequence Labeling”. In: *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. Ed. by Emily M. Bender, Leon Derczynski, and Pierre Isabelle. Association for Computational Linguistics, 2018, pp. 1638–1649. URL: <https://www.aclweb.org/anthology/C18-1139/>.
- [12] Tareq Al-Moslmi, Marc Gallofré Ocaña, Andreas L. Opdahl, and Csaba Veres. “Named Entity Extraction for Knowledge Graphs: A Literature Overview”. In: *IEEE Access* 8 (2020), pp. 32862–32881. DOI: 10.1109/ACCESS.2020.2973928.
- [13] Hunt Allcott and Matthew Gentzkow. “Social media and fake news in the 2016 election”. In: *Journal of Economic Perspectives* 31.2 (2017), pp. 211–36. DOI: 10.3386/w23089.
- [14] Hani Altwaijry, Eduard Trulls, James Hays, Pascal Fua, and Serge J. Belongie. “Learning to Match Aerial Images with Deep Attentive Architectures”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 3539–3547. DOI: 10.1109/CVPR.2016.385.
- [15] Zhanfu An, Weihong Deng, Tongtong Yuan, and Jiani Hu. “Deep Transfer Network with 3D Morphable Models for Face Recognition”. In: *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi’an, China, May 15-19, 2018*. IEEE Computer Society, 2018, pp. 416–422. DOI: 10.1109/FG.2018.00067.
- [16] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 6077–6086. DOI: 10.1109/CVPR.2018.00636. URL: [http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Anderson\\_Bottom-Up\\_and\\_Top-Down\\_CVPR\\_2018\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.html).
- [17] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. “Boosting cross-age face verification via generative age normalization”. In: *2017 IEEE International Joint Conference on Biometrics, IJCB 2017, Denver, CO, USA, October 1-4, 2017*. IEEE, 2017, pp. 191–199. DOI: 10.1109/BTAS.2017.8272698.
- [18] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. “Face aging with conditional generative adversarial networks”. In: *2017 IEEE International Conference on Image Processing, ICIP 2017, Beijing, China, September 17-20, 2017*. IEEE, 2017, pp. 2089–2093. DOI: 10.1109/ICIP.2017.8296650.



- [19] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. “VQA: Visual Question Answering”. In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 2425–2433. DOI: 10.1109/ICCV.2015.279.
- [20] Relja Arandjelovic, Petr Gronát, Akihiko Torii, Tomás Pajdla, and Josef Sivic. “NetVLAD: CNN Architecture for Weakly Supervised Place Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 5297–5307. DOI: 10.1109/CVPR.2016.572.
- [21] Shota Ashida, Adam Jatowt, Antoine Doucet, and Masatoshi Yoshikawa. “Determining image age with rank-consistent ordinal classification and object-centered ensemble”. In: *MMAsia 2020: ACM Multimedia Asia, Virtual Event / Singapore, 7-9 March, 2021*. Ed. by Tat-Seng Chua, Jingdong Wang, Qi Tian, Cathal Gurrin, Jia Jia, Hanwang Zhang, and Qianru Sun. ACM, 2020, 50:1–50:8. DOI: 10.1145/3444685.3446326.
- [22] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. “DBpedia: A Nucleus for a Web of Open Data”. In: *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*. Ed. by Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux. Vol. 4825. Lecture Notes in Computer Science. Springer, 2007, pp. 722–735. DOI: 10.1007/978-3-540-76298-0\_52. URL: [https://doi.org/10.1007/978-3-540-76298-0\\_52](https://doi.org/10.1007/978-3-540-76298-0_52).
- [23] Yannis Avrithis, Yannis Kalantidis, Giorgos Tolias, and Evaggelos Spyrou. “Retrieving landmark and non-landmark images from community photo collections”. In: *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*. Ed. by Alberto Del Bimbo, Shih-Fu Chang, and Arnold W. M. Smeulders. ACM, 2010, pp. 153–162. DOI: 10.1145/1873951.1873973.
- [24] Georges Baatz, Olivier Saurer, Kevin Köser, and Marc Pollefeys. “Large Scale Visual Geo-Localization of Images in Mountainous Terrain”. In: *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II*. Ed. by Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid. Vol. 7573. Lecture Notes in Computer Science. Springer, 2012, pp. 517–530. DOI: 10.1007/978-3-642-33709-3\_37. URL: [https://doi.org/10.1007/978-3-642-33709-3\\_37](https://doi.org/10.1007/978-3-642-33709-3_37).
- [25] Siham Bacha, Mohand Saïd Allili, and Nadjia Benblidia. “Event recognition in photo albums using probabilistic graphical models and feature relevance”. In: *J. Vis. Commun. Image Represent.* 40 (2016), pp. 546–558. DOI: 10.1016/j.jvcir.2016.07.021.
- [26] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. “Multimodal Machine Learning: A Survey and Taxonomy”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 41.2 (2019), pp. 423–443. DOI: 10.1109/TPAMI.2018.2798607.

- [27] Mayank Bansal, Kostas Daniilidis, and Harpreet S. Sawhney. “Ultrawide Baseline Facade Matching for Geo-localization”. In: *Deep Learning and Convolutional Neural Networks for Medical Image Computing - Precision Medicine, High Performance and Large-Scale Datasets*. Ed. by Le Lu, Yefeng Zheng, Gustavo Carneiro, and Lin Yang. Advances in Computer Vision and Pattern Recognition. Springer, 2016, pp. 77–98. DOI: 10.1007/978-3-319-25781-5\_5. URL: [https://doi.org/10.1007/978-3-319-25781-5\\_5](https://doi.org/10.1007/978-3-319-25781-5_5).
- [28] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. “CVAE-GAN: Fine-Grained Image Generation through Asymmetric Training”. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2764–2773. DOI: 10.1109/ICCV.2017.299.
- [29] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. “Towards Open-Set Identity Preserving Face Synthesis”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 6713–6722. DOI: 10.1109/CVPR.2018.00702. URL: [http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Bao\\_Towards\\_Open-Set\\_Identity\\_CVPR\\_2018\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Bao_Towards_Open-Set_Identity_CVPR_2018_paper.html).
- [30] Roland Barthes. “Image-Music-Text, ed. and trans”. In: *S. Heath, London: Fontana* 332 (1977).
- [31] John Bateman. *Text and image: A critical introduction to the visual/verbal divide*. Routledge, 2014. DOI: 10.4324/9781315773971.
- [32] Tim Berners-Lee. “Linked data”. In: *Int. J. on Semantic Web and Information Systems* 4.2 (2006). DOI: 10.4018/978-1-60960-593-3.ch008.
- [33] Tim Berners-Lee, James Hendler, and Ora Lassila. “The semantic web”. In: *Scientific american* 284.5 (2001), pp. 34–43. DOI: 10.1038/scientificamerican0501-34.
- [34] Joachim Bingel and Anders Søgaard. “Identifying beneficial task relations for multi-task learning in deep neural networks”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Association for Computational Linguistics, 2017, pp. 164–169. DOI: 10.18653/v1/e17-2026.
- [35] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. “Enriching Word Vectors with Subword Information”. In: *Trans. Assoc. Comput. Linguistics* 5 (2017), pp. 135–146. URL: <https://transacl.org/ojs/index.php/tacl/article/view/999>.
- [36] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. “Event Recognition in Photo Collections with a Stopwatch HMM”. In: *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. IEEE Computer Society, 2013, pp. 1193–1200. DOI: 10.1109/ICCV.2013.151.
- [37] Léon Bottou. “Large-Scale Machine Learning with Stochastic Gradient Descent”. In: *19th International Conference on Computational Statistics, COMPSTAT 2010, Paris, France, August 22-27, 2010 - Keynote, Invited and Contributed Papers*. Ed. by Yves Lechevallier and Gilbert Saporta. Physica-Verlag, 2010, pp. 177–186. DOI: 10.1007/978-3-7908-2604-3\_16. URL: [https://doi.org/10.1007/978-3-7908-2604-3\\_16](https://doi.org/10.1007/978-3-7908-2604-3_16).

- [38] Alexandre Bovet and Hernán A Makse. “Influence of fake news in Twitter during the 2016 US presidential election”. In: *Nature communications* 10.1 (2019), pp. 1–14. DOI: 10.1038/s41467-018-07761-2.
- [39] Marco Brambilla, Stefano Ceri, Emanuele Della Valle, Riccardo Volonterio, and Felix Xavier Acero Salazar. “Extracting Emerging Knowledge from Social Media”. In: *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*. Ed. by Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich. ACM, 2017, pp. 795–804. DOI: 10.1145/3038912.3052697.
- [40] Janez Brank, Gregor Leban, and Marko Grobelnik. “Annotating Documents with Relevant Wikipedia Concepts”. In: *Proceedings of the Slovenian Conference on Data Mining and Data Warehouses (SiKDD 2017), October 9, 2017, Ljubljana, Slovenia*. 2017.
- [41] Janez Brank, Gregor Leban, and Marko Grobelnik. “Semantic Annotation of Documents Based on Wikipedia Concepts”. In: *Informatica (Slovenia)* 42.1 (2018). URL: <http://www.informatica.si/index.php/informatica/article/view/2228>.
- [43] Jan Brejcha and Martin Cadik. “State-of-the-art in visual geo-localization”. In: *Pattern Anal. Appl.* 20.3 (2017), pp. 613–637. DOI: 10.1007/s10044-017-0611-1.
- [44] Marcel Broersma and Todd Graham. “Twitter as a news source: How Dutch and British newspapers used tweets in their news coverage, 2007–2011”. In: *Journalism practice* 7.4 (2013), pp. 446–464. DOI: 10.1080/17512786.2013.802481.
- [45] K. Selçuk Candan, Marco Bertini, Xiao-Yong Wei, and Lexing Xie. “Editorial for the ICMR 2019 special issue”. In: *Int. J. Multim. Inf. Retr.* 9.1 (2020), pp. 1–2. DOI: 10.1007/s13735-020-00192-0.
- [46] Liangliang Cao, John R. Smith, Zhen Wen, Zhijun Yin, Xin Jin, and Jiawei Han. “BlueFinder: estimate where a beach photo was taken”. In: *Proceedings of the 21st World Wide Web Conference, WWW 2012, Lyon, France, April 16-20, 2012 (Companion Volume)*. Ed. by Alain Mille, Fabien L. Gandon, Jacques Misselis, Michael Rabinovich, and Steffen Staab. ACM, 2012, pp. 469–470. DOI: 10.1145/2187980.2188081.
- [47] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. “VG-Face2: A Dataset for Recognising Faces across Pose and Age”. In: *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi’an, China, May 15-19, 2018*. IEEE Computer Society, 2018, pp. 67–74. DOI: 10.1109/FG.2018.00020.
- [48] Weilong Chai, Weihong Deng, and Haifeng Shen. “Cross-Generating GAN for Facial Identity Preserving”. In: *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi’an, China, May 15-19, 2018*. IEEE Computer Society, 2018, pp. 130–134. DOI: 10.1109/FG.2018.00028.
- [51] Bor-Chun Chen and Larry S Davis. “Deep Representation Learning for Metadata Verification”. In: *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. IEEE. 2019, pp. 73–82. DOI: 10.1109/wacvw.2019.00019.

- [52] Bor-Chun Chen, Pallabi Ghosh, Vlad I. Morariu, and Larry S. Davis. “Detection of Metadata Tampering Through Discrepancy Between Image Content and Metadata Using Multi-task Deep Learning”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 1872–1880. DOI: 10.1109/CVPRW.2017.234.
- [53] David M. Chen, Georges Baatz, Kevin Köser, Sam S. Tsai, Ramakrishna Vedantham, Timo Pylvänäinen, Kimmo Roimela, Xin Chen, Jeff Bach, Marc Pollefeys, Bernd Girod, and Radek Grzeszczuk. “City-scale landmark identification on mobile devices”. In: *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*. IEEE Computer Society, 2011, pp. 737–744. DOI: 10.1109/CVPR.2011.5995610.
- [54] François Chollet. “Xception: Deep Learning with Depthwise Separable Convolutions”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 1800–1807. DOI: 10.1109/CVPR.2017.195.
- [55] Silviu Cucerzan. “Large-Scale Named Entity Disambiguation Based on Wikipedia Data”. In: *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*. Ed. by Jason Eisner. ACL, 2007, pp. 708–716. URL: <https://www.aclweb.org/anthology/D07-1074/>.
- [56] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. “Improving efficiency and accuracy in multilingual entity extraction”. In: *I-SEMANTICS 2013 - 9th International Conference on Semantic Systems, ISEM '13, Graz, Austria, September 4-6, 2013*. Ed. by Marta Sabou, Eva Blomqvist, Tommaso Di Noia, Harald Sack, and Tassilo Pellegrini. ACM, 2013, pp. 121–124. DOI: 10.1145/2506182.2506198.
- [57] Navneet Dalal and Bill Triggs. “Histograms of Oriented Gradients for Human Detection”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*. IEEE Computer Society, 2005, pp. 886–893. DOI: 10.1109/CVPR.2005.177.
- [58] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.
- [59] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. “ArcFace: Additive Angular Margin Loss for Deep Face Recognition”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 4690–4699. DOI: 10.1109/CVPR.2019.00482. URL: [http://openaccess.thecvf.com/content\\_CVPR\\_2019/html/Deng\\_ArcFace\\_Additive\\_Angular\\_Margin\\_Loss\\_for\\_Deep\\_Face\\_Recognition\\_CVPR\\_2019\\_paper.html](http://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html).

- [60] Jiankang Deng, Yuxiang Zhou, and Stefanos Zafeiriou. “Marginal Loss for Deep Face Recognition”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 2006–2014. DOI: 10.1109/CVPRW.2017.251.
- [61] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/n19-1423.
- [62] Changxing Ding and Dacheng Tao. “Robust Face Recognition via Multimodal Deep Face Representation”. In: *IEEE Trans. Multimedia* 17.11 (2015), pp. 2049–2058. DOI: 10.1109/TMM.2015.2477042.
- [63] Changxing Ding and Dacheng Tao. “Trunk-Branch Ensemble Convolutional Neural Networks for Video-Based Face Recognition”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 40.4 (2018), pp. 1002–1014. DOI: 10.1109/TPAMI.2017.2700390.
- [64] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL: <https://openreview.net/forum?id=YicbFdNTTy>.
- [65] Sean R Eddy. “Hidden markov models”. In: *Current opinion in structural biology* 6.3 (1996), pp. 361–365. DOI: 10.1016/s0959-440x(96)80056-x.
- [66] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi González, Hugo Jair Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon. “ChaLearn Looking at People 2015: Apparent Age and Cultural Event Recognition Datasets and Results”. In: *2015 IEEE International Conference on Computer Vision Workshop, ICCV Workshops 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 243–251. DOI: 10.1109/ICCVW.2015.40.
- [67] Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuya Yamada, and Omer Levy. “Named Entity Disambiguation for Noisy Text”. In: *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*. Ed. by Roger Levy and Lucia Specia. Association for Computational Linguistics, 2017, pp. 58–68. DOI: 10.18653/v1/K17-1008.
- [70] Wei Fang, Jianwen Zhang, Dilin Wang, Zheng Chen, and Ming Li. “Entity Disambiguation by Knowledge and Text Jointly Embedding”. In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*. Ed. by Yoav Goldberg and Stefan Riezler. ACL, 2016, pp. 260–269. DOI: 10.18653/v1/k16-1026.
- [71] Delia Fernandez, David Varas, Joan Espadaler, Issey Masuda, Jordi Ferreira, Alejandro Woodward, David Rodriguez, Xavier Giró-i-Nieto, Juan Carlos Riveiro, and Elisenda Bou. “ViTS: Video Tagging System from Massive Web Multimedia Collections”. In: *2017 IEEE International Conference on Computer Vision Workshops*,

- ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 337–346. DOI: 10.1109/ICCVW.2017.48.
- [72] Basura Fernando, Damien Muselet, Rahat Khan, and Tinne Tuytelaars. “Color features for dating historical color images”. In: *2014 IEEE International Conference on Image Processing, ICIP 2014, Paris, France, October 27-30, 2014*. IEEE, 2014, pp. 2589–2593. DOI: 10.1109/ICIP.2014.7025524.
- [73] Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling”. In: *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*. Ed. by Kevin Knight, Hwee Tou Ng, and Kemal Oflazer. The Association for Computer Linguistics, 2005, pp. 363–370. DOI: 10.3115/1219840.1219885. URL: <https://www.aclweb.org/anthology/P05-1045/>.
- [74] Matthew Francis-Landau, Greg Durrett, and Dan Klein. “Capturing Semantic Similarity for Entity Linking with Convolutional Neural Networks”. In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. Ed. by Kevin Knight, Ani Nenkova, and Owen Rambow. The Association for Computational Linguistics, 2016, pp. 1256–1261. DOI: 10.18653/v1/n16-1150.
- [75] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. “Deep Ordinal Regression Network for Monocular Depth Estimation”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 2002–2011. DOI: 10.1109/CVPR.2018.00214. URL: [http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Fu\\_Deep\\_Ordinal\\_Regression\\_CVPR\\_2018\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Fu_Deep_Ordinal_Regression_CVPR_2018_paper.html).
- [76] Octavian-Eugen Ganea and Thomas Hofmann. “Deep Joint Entity Disambiguation with Local Neural Attention”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Association for Computational Linguistics, 2017, pp. 2619–2629. DOI: 10.18653/v1/d17-1277.
- [77] Shiry Ginosar, Kate Rakelly, Sarah Sachs, Brian Yin, and Alexei A. Efros. “A Century of Portraits: A Visual Historical Record of American High School Yearbooks”. In: *2015 IEEE International Conference on Computer Vision Workshop, ICCV Workshops 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 652–658. DOI: 10.1109/ICCVW.2015.87.
- [78] Michael Goebel, Arjuna Flenner, Lakshmanan Nataraj, and Bangalore S. Manjunath. “Deep Learning Methods for Event Verification and Image Repurposing Detection”. In: *Media Watermarking, Security, and Forensics 2019, Burlingame, CA, USA, 13-17 January 2019*. Ed. by Adnan M. Alattar, Nasir D. Memon, and Gaurav Sharma. Ingenta, 2019. DOI: 10.2352/ISSN.2470-1173.2019.5.MWSF-530.
- [79] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. <http://www.deeplearningbook.org>. MIT Press, 2016.

- [80] Albert Gordo, Jon Almazán, Jérôme Revaud, and Diane Larlus. “Deep Image Retrieval: Learning Global Representations for Image Search”. In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Vol. 9910. Lecture Notes in Computer Science. Springer, 2016, pp. 241–257. DOI: 10.1007/978-3-319-46466-4\_15. URL: [https://doi.org/10.1007/978-3-319-46466-4\\_15](https://doi.org/10.1007/978-3-319-46466-4_15).
- [81] Simon Gottschalk and Elena Demidova. “EventKG: A Multilingual Event-Centric Temporal Knowledge Graph”. In: *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*. Ed. by Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam. Vol. 10843. Lecture Notes in Computer Science. Springer, 2018, pp. 272–287. DOI: 10.1007/978-3-319-93417-4\_18. URL: [https://doi.org/10.1007/978-3-319-93417-4\\_18](https://doi.org/10.1007/978-3-319-93417-4_18).
- [82] Simon Gottschalk and Elena Demidova. “EventKG - the hub of event knowledge on the web - and biographical timeline generation”. In: *Semantic Web 10.6 (2019)*, pp. 1039–1070. DOI: 10.3233/SW-190355.
- [83] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. “Learning Word Vectors for 157 Languages”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. Ed. by Nicoletta Calzolari et al. European Language Resources Association (ELRA), 2018. URL: <http://www.lrec-conf.org/proceedings/lrec2018/summaries/627.html>.
- [84] Alex Graves and Jürgen Schmidhuber. “Framewise phoneme classification with bidirectional LSTM and other neural network architectures”. In: *Neural Networks 18.5-6 (2005)*, pp. 602–610. DOI: 10.1016/j.neunet.2005.06.042.
- [85] Wenzhong Guo, Jianwen Wang, and Shiping Wang. “Deep Multimodal Representation Learning: A Survey”. In: *IEEE Access 7 (2019)*, pp. 63373–63394. DOI: 10.1109/ACCESS.2019.2916887.
- [86] Xin Guo, Luisa F. Polanía, Bin Zhu, Charles Boncelet, and Kenneth E. Barner. “Graph Neural Networks for Image Understanding Based on Multiple Cues: Group Emotion Recognition and Event Recognition as Use Cases”. In: *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*. IEEE, 2020, pp. 2910–2919. DOI: 10.1109/WACV45572.2020.9093547.
- [87] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. “MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition”. In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Vol. 9907. Lecture Notes in Computer Science. Springer, 2016, pp. 87–102. DOI: 10.1007/978-3-319-46487-9\_6. URL: [https://doi.org/10.1007/978-3-319-46487-9\\_6](https://doi.org/10.1007/978-3-319-46487-9_6).
- [88] Michael Alexander Kirkwood Halliday and Christian MIM Matthiessen. *Halliday’s introduction to functional grammar*. Routledge, 2013. DOI: 10.4324/9780203431269.

- [89] James Hays and Alexei A. Efros. “IM2GPS: estimating geographic information from a single image”. In: *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. IEEE Computer Society, 2008. DOI: 10.1109/CVPR.2008.4587784.
- [90] James Hays and Alexei A. Efros. “Large-Scale Image Geolocalization”. In: *Multi-modal Location Estimation of Videos and Images*. Ed. by Jaeyoung Choi and Gerald Friedland. Springer, 2015, pp. 41–62. DOI: 10.1007/978-3-319-09861-6\_3. URL: [https://doi.org/10.1007/978-3-319-09861-6\\_3](https://doi.org/10.1007/978-3-319-09861-6_3).
- [91] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 1026–1034. DOI: 10.1109/ICCV.2015.123.
- [92] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.
- [93] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Identity Mappings in Deep Residual Networks”. In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Vol. 9908. Lecture Notes in Computer Science. Springer, 2016, pp. 630–645. DOI: 10.1007/978-3-319-46493-0\_38. URL: [https://doi.org/10.1007/978-3-319-46493-0\\_38](https://doi.org/10.1007/978-3-319-46493-0_38).
- [94] Sheng He, Petros Samara, Jan Burgers, and Lambert Schomaker. “Image-based historical manuscript dating using contour and stroke fragments”. In: *Pattern Recognit.* 58 (2016), pp. 159–171. DOI: 10.1016/j.patcog.2016.03.032.
- [95] Marti A. Hearst. “Trends & Controversies: Support Vector Machines”. In: *IEEE Intell. Syst.* 13.4 (1998), pp. 18–28. DOI: 10.1109/5254.708428.
- [96] Christian Andreas Henning and Ralph Ewerth. “Estimating the Information Gap between Textual and Visual Representations”. In: *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR 2017, Bucharest, Romania, June 6-9, 2017*. Ed. by Bogdan Ionescu, Nicu Sebe, Jiashi Feng, Martha A. Larson, Rainer Lienhart, and Cees Snoek. ACM, 2017, pp. 14–22. DOI: 10.1145/3078971.3078991.
- [97] Christian Andreas Henning and Ralph Ewerth. “Estimating the information gap between textual and visual representations”. In: *Int. J. Multim. Inf. Retr.* 7.1 (2018), pp. 43–56. DOI: 10.1007/s13735-017-0142-y.
- [98] Johannes Hoffart, Mohamed Amir Yosef, Iliaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. “Robust Disambiguation of Named Entities in Text”. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, 2011, pp. 782–792. URL: <https://www.aclweb.org/anthology/D11-1072/>.



- [99] Matthew Honnibal and Ines Montani. “spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing”. In: *To appear* (2017).
- [100] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”. In: *CoRR* abs/1704.04861 (2017). arXiv: 1704.04861. URL: <http://arxiv.org/abs/1704.04861>.
- [101] Andrew Howard, Ruoming Pang, Hartwig Adam, Quoc V. Le, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, and Yukun Zhu. “Searching for MobileNetV3”. In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 1314–1324. DOI: 10.1109/ICCV.2019.00140.
- [102] Jeremy Howard and Sebastian Ruder. “Universal Language Model Fine-tuning for Text Classification”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. Ed. by Iryna Gurevych and Yusuke Miyao. Association for Computational Linguistics, 2018, pp. 328–339. DOI: 10.18653/v1/P18-1031. URL: <https://www.aclweb.org/anthology/P18-1031/>.
- [103] Jie Hu, Li Shen, and Gang Sun. “Squeeze-and-Excitation Networks”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 7132–7141. DOI: 10.1109/CVPR.2018.00745. URL: [http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Hu\\_Squeeze-and-Excitation\\_Networks\\_CVPR\\_2018\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html).
- [104] Lanqing Hu, Meina Kan, Shiguang Shan, Xingguang Song, and Xilin Chen. “LDF-Net: Learning a Displacement Field Network for Face Recognition across Pose”. In: *12th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2017, Washington, DC, USA, May 30 - June 3, 2017*. IEEE Computer Society, 2017, pp. 9–16. DOI: 10.1109/FG.2017.12.
- [105] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. “Learning to Reason: End-to-End Module Networks for Visual Question Answering”. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 804–813. DOI: 10.1109/ICCV.2017.93.
- [106] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. “Densely Connected Convolutional Networks”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 2261–2269. DOI: 10.1109/CVPR.2017.243.
- [107] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Tech. rep. 07-49. University of Massachusetts, Amherst, Oct. 2007.
- [108] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. “Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis”. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2458–2467. DOI: 10.1109/ICCV.2017.267.

- [109] David H Hubel and Torsten N Wiesel. “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex”. In: *The Journal of physiology* 160.1 (1962), p. 106. DOI: 10.1113/jphysiol.1962.sp006837.
- [110] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. Ed. by Francis R. Bach and David M. Blei. Vol. 37. JMLR Workshop and Conference Proceedings. JMLR.org, 2015, pp. 448–456. URL: <http://proceedings.mlr.press/v37/ioffe15.html>.
- [111] Mike Izbicki, Evangelos E. Papalexakis, and Vassilis J. Tsotras. “Exploiting the Earth’s Spherical Geometry to Geolocate Images”. In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2019, Würzburg, Germany, September 16-20, 2019, Proceedings, Part II*. Ed. by Ulf Brefeld, Élisabeth Fromont, Andreas Hotho, Arno J. Knobbe, Marloes H. Maathuis, and Céline Robardet. Vol. 11907. Lecture Notes in Computer Science. Springer, 2019, pp. 3–19. DOI: 10.1007/978-3-030-46147-8\_1. URL: [https://doi.org/10.1007/978-3-030-46147-8\\_1](https://doi.org/10.1007/978-3-030-46147-8_1).
- [112] Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z. Leibo, David Silver, and Koray Kavukcuoglu. “Reinforcement Learning with Unsupervised Auxiliary Tasks”. In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL: <https://openreview.net/forum?id=SJ6yPD5xg>.
- [113] Vidit Jain, Amit Singhal, and Jiebo Luo. “Selective hidden random fields: Exploiting domain-specific saliency for event classification”. In: *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. IEEE Computer Society, 2008. DOI: 10.1109/CVPR.2008.4587431.
- [114] Ayush Jaiswal, Ekraam Sabir, Wael Abd-Almageed, and Premkumar Natarajan. “Multimedia Semantic Integrity Assessment Using Joint Embedding Of Images And Text”. In: *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*. Ed. by Qiong Liu, Rainer Lienhart, Hao-hong Wang, Sheng-Wei "Kuan-Ta" Chen, Susanne Boll, Yi-Ping Phoebe Chen, Gerald Friedland, Jia Li, and Shuicheng Yan. ACM, 2017, pp. 1465–1471. DOI: 10.1145/3123266.3123385.
- [115] Ayush Jaiswal, Yue Wu, Wael AbdAlmageed, Iacopo Masi, and Premkumar Natarajan. “AIRD: Adversarial Learning Framework for Image Repurposing Detection”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 11330–11339. DOI: 10.1109/CVPR.2019.01159. URL: [http://openaccess.thecvf.com/content\\_CVPR\\_2019/html/Jaiswal\\_AIRD\\_Adversarial\\_Learning\\_Framework\\_for\\_Image\\_Repurposing\\_Detection\\_CVPR\\_2019\\_paper.html](http://openaccess.thecvf.com/content_CVPR_2019/html/Jaiswal_AIRD_Adversarial_Learning_Framework_for_Image_Repurposing_Detection_CVPR_2019_paper.html).
- [116] Xin Jin and Xiaoyang Tan. “Face alignment in-the-wild: A Survey”. In: *Comput. Vis. Image Underst.* 162 (2017), pp. 1–22. DOI: 10.1016/j.cviu.2017.08.008.

- [117] Pravin Kakar and N. Sudha. “Verifying Temporal Data in Geotagged Images Via Sun Azimuth Estimation”. In: *IEEE Trans. Inf. Forensics Secur.* 7.3 (2012), pp. 1029–1039. DOI: 10.1109/TIFS.2012.2188796.
- [118] Andrej Karpathy and Fei-Fei Li. “Deep visual-semantic alignments for generating image descriptions”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 3128–3137. DOI: 10.1109/CVPR.2015.7298932.
- [119] Alex Kendall, Matthew Grimes, and Roberto Cipolla. “PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization”. In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 2938–2946. DOI: 10.1109/ICCV.2015.336.
- [120] J. Kiefer and J. Wolfowitz. “Stochastic Estimation of the Maximum of a Regression Function”. In: *The Annals of Mathematical Statistics* 23.3 (1952), pp. 462–466. ISSN: 0003-4851. DOI: 10.1214/aoms/1177729392.
- [121] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. “Predicting Good Features for Image Geo-Localization Using Per-Bundle VLAD”. In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 1170–1178. DOI: 10.1109/ICCV.2015.139.
- [122] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. “Learned Contextual Feature Reweighting for Image Geo-Localization”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 3251–3260. DOI: 10.1109/CVPR.2017.346.
- [123] Davis E. King. “Dlib-ml: A Machine Learning Toolkit”. In: *Journal of Machine Learning Research* 10 (2009), pp. 1755–1758.
- [124] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: <http://arxiv.org/abs/1412.6980>.
- [125] Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. “End-to-End Neural Entity Linking”. In: *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*. Ed. by Anna Korhonen and Ivan Titov. Association for Computational Linguistics, 2018, pp. 519–529. DOI: 10.18653/v1/k18-1050.
- [126] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*. Ed. by Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger. 2012, pp. 1106–1114. URL: <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks>.

- [127] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. “Integrating Text and Image: Determining Multimodal Document Intent in Instagram Posts”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Association for Computational Linguistics, 2019, pp. 4621–4631. DOI: 10.18653/v1/D19-1469.
- [128] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”. In: *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*. Ed. by Carla E. Brodley and Andrea Pohoreckyj Danyluk. Morgan Kaufmann, 2001, pp. 282–289.
- [129] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. “Neural Architectures for Named Entity Recognition”. In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. Ed. by Kevin Knight, Ani Nenkova, and Owen Rambow. The Association for Computational Linguistics, 2016, pp. 260–270. DOI: 10.18653/v1/n16-1030.
- [130] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations”. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL: <https://openreview.net/forum?id=H1eA7AetvS>.
- [131] Martha A. Larson, Mohammad Soleymani, Guillaume Gravier, Bogdan Ionescu, and Gareth J. F. Jones. “The Benchmarking Initiative for Multimedia Evaluation: MediaEval 2016”. In: *IEEE Multim.* 24.1 (2017), pp. 93–96. DOI: 10.1109/MMUL.2017.9.
- [132] Phong Le and Ivan Titov. “Improving Entity Linking by Modeling Latent Relations between Mentions”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. Ed. by Iryna Gurevych and Yusuke Miyao. Association for Computational Linguistics, 2018, pp. 1595–1604. DOI: 10.18653/v1/P18-1148. URL: <https://www.aclweb.org/anthology/P18-1148/>.
- [133] Yong Jae Lee, Alexei A. Efros, and Martial Hebert. “Style-Aware Mid-level Representation for Discovering Visual Connections in Space and Time”. In: *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. IEEE Computer Society, 2013, pp. 1857–1864. DOI: 10.1109/ICCV.2013.233.
- [134] Jay Lemke. “Multiplying meaning: Visual and verbal semiotics in scientific text”. In: *Reading science: Critical and functional perspectives on discourses of science* (1998), pp. 87–113.
- [135] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. “A Survey on Deep Learning for Named Entity Recognition”. In: *IEEE Trans. Knowl. Data Eng.* 34.1 (2022), pp. 50–70. DOI: 10.1109/TKDE.2020.2981314.

- [136] Li-Jia Li and Fei-Fei Li. “What, where and who? Classifying events by scene and object recognition”. In: *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*. IEEE Computer Society, 2007, pp. 1–8. DOI: 10.1109/ICCV.2007.4408872.
- [137] Xiaopeng Li, Wenyuan Xu, Song Wang, and Xianshan Qu. “Are You Lying: Validating the Time-Location of Outdoor Images”. In: *Applied Cryptography and Network Security - 15th International Conference, ACNS 2017, Kanazawa, Japan, July 10-12, 2017, Proceedings*. Ed. by Dieter Gollmann, Atsuko Miyaji, and Hiroaki Kikuchi. Vol. 10355. Lecture Notes in Computer Science. Springer, 2017, pp. 103–123. DOI: 10.1007/978-3-319-61204-1\_6. URL: [https://doi.org/10.1007/978-3-319-61204-1\\_6](https://doi.org/10.1007/978-3-319-61204-1_6).
- [138] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. “Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks”. In: *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm. Vol. 12375. Lecture Notes in Computer Science. Springer, 2020, pp. 121–137. DOI: 10.1007/978-3-030-58577-8\_8. URL: [https://doi.org/10.1007/978-3-030-58577-8\\_8](https://doi.org/10.1007/978-3-030-58577-8_8).
- [139] Yuanpeng Li, Dmitriy Genzel, Yasuhisa Fujii, and Ashok C. Popat. “Publication Date Estimation for Printed Historical Documents using Convolutional Neural Networks”. In: *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing, HIP@ICDAR 2015, Nancy, France, August 22, 2015*. ACM, 2015, pp. 99–106. DOI: 10.1145/2809544.2809550.
- [140] Yunpeng Li, David J. Crandall, and Daniel P. Huttenlocher. “Landmark classification in large-scale image collections”. In: *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*. IEEE Computer Society, 2009, pp. 1957–1964. DOI: 10.1109/ICCV.2009.5459432.
- [141] Yunpeng Li, Noah Snavely, Daniel P. Huttenlocher, and Pascal Fua. “Worldwide Pose Estimation Using 3D Point Clouds”. In: *Deep Learning and Convolutional Neural Networks for Medical Image Computing - Precision Medicine, High Performance and Large-Scale Datasets*. Ed. by Le Lu, Yefeng Zheng, Gustavo Carneiro, and Lin Yang. Advances in Computer Vision and Pattern Recognition. Springer, 2016, pp. 147–163. DOI: 10.1007/978-3-319-25781-5\_8. URL: [https://doi.org/10.1007/978-3-319-25781-5\\_8](https://doi.org/10.1007/978-3-319-25781-5_8).
- [142] Tsung-Yi Lin, Serge J. Belongie, and James Hays. “Cross-View Image Geolocalization”. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. IEEE Computer Society, 2013, pp. 891–898. DOI: 10.1109/CVPR.2013.120.
- [143] Tsung-Yi Lin, Yin Cui, Serge J. Belongie, and James Hays. “Learning deep representations for ground-to-aerial geolocalization”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 5007–5015. DOI: 10.1109/CVPR.2015.7299135.

- [144] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. “Isolation Forest”. In: *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*. IEEE Computer Society, 2008, pp. 413–422. DOI: 10.1109/ICDM.2008.17.
- [145] Liu Liu, Hongdong Li, and Yuchao Dai. “Efficient Global 2D-3D Matching for Camera Localization in a Large-Scale 3D Map”. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2391–2400. DOI: 10.1109/ICCV.2017.260.
- [146] Liu Liu, Hongdong Li, and Yuchao Dai. “Stochastic Attraction-Repulsion Embedding for Large Scale Image Localization”. In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 2570–2579. DOI: 10.1109/ICCV.2019.00266.
- [147] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. “SphereFace: Deep Hypersphere Embedding for Face Recognition”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 6738–6746. DOI: 10.1109/CVPR.2017.713.
- [148] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. “Large-Margin Softmax Loss for Convolutional Neural Networks”. In: *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*. Ed. by Maria-Florina Balcan and Kilian Q. Weinberger. Vol. 48. JMLR Workshop and Conference Proceedings. JMLR.org, 2016, pp. 507–516. URL: <http://proceedings.mlr.press/v48/liud16.html>.
- [149] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *CoRR* abs/1907.11692 (2019). arXiv: 1907.11692. URL: <http://arxiv.org/abs/1907.11692>.
- [150] Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. “Zero-Shot Entity Linking by Reading Entity Descriptions”. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum, and Lluís Màrquez. Association for Computational Linguistics, 2019, pp. 3449–3460. DOI: 10.18653/v1/p19-1335.
- [151] Ilya Loshchilov and Frank Hutter. “SGDR: Stochastic Gradient Descent with Warm Restarts”. In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL: <https://openreview.net/forum?id=Skq89Scxx>.
- [152] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. “The Stanford CoreNLP Natural Language Processing Toolkit”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*. The Association for Computer Linguistics, 2014, pp. 55–60. DOI: 10.3115/v1/p14-5010.

- [153] Emily E. Marsh and Marilyn Domas White. “A taxonomy of relationships between images and text”. In: *J. Documentation* 59.6 (2003), pp. 647–672. DOI: 10.1108/00220410310506303.
- [154] Paul Martin, Antoine Doucet, and Frédéric Jurie. “Dating Color Images with Ordinal Classification”. In: *International Conference on Multimedia Retrieval, ICMR '14, Glasgow, United Kingdom - April 01 - 04, 2014*. Ed. by Mohan S. Kankanhalli, Stefan Rueger, R. Manmatha, Joemon M. Jose, and Keith van Rijsbergen. ACM, 2014, p. 447. DOI: 10.1145/2578726.2578790.
- [155] Radan Martinec and Andrew Salway. “A system for image–text relations in new (and old) media”. In: *Visual Communication* 4.3 (2005), pp. 337–371. DOI: 10.1177/1470357205055928.
- [156] José-Lázaro Martínez-Rodríguez, Aidan Hogan, and Ivan López-Arévalo. “Information extraction meets the Semantic Web: A survey”. In: *Semantic Web* 11.2 (2020), pp. 255–335. DOI: 10.3233/SW-180333.
- [157] Iacopo Masi, Tal Hassner, Anh Tuan Tran, and Gérard G. Medioni. “Rapid Synthesis of Massive Face Sets for Improved Face Recognition”. In: *12th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2017, Washington, DC, USA, May 30 - June 3, 2017*. IEEE Computer Society, 2017, pp. 604–611. DOI: 10.1109/FG.2017.76.
- [158] Iacopo Masi, Stephen Rawls, Gérard G. Medioni, and Prem Natarajan. “Pose-Aware Face Recognition in the Wild”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 4838–4846. DOI: 10.1109/CVPR.2016.523.
- [159] Iacopo Masi, Anh Tuan Tran, Tal Hassner, Jatuporn Toy Leksut, and Gérard G. Medioni. “Do We Really Need to Collect Millions of Faces for Effective Face Recognition?” In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Vol. 9909. Lecture Notes in Computer Science. Springer, 2016, pp. 579–596. DOI: 10.1007/978-3-319-46454-1\_35. URL: [https://doi.org/10.1007/978-3-319-46454-1\\_35](https://doi.org/10.1007/978-3-319-46454-1_35).
- [160] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. “Deep Face Recognition: A Survey”. In: *31st SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI 2018, Paraná, Brazil, October 29 - Nov. 1, 2018*. IEEE Computer Society, 2018, pp. 471–478. DOI: 10.1109/SIBGRAPI.2018.00067.
- [161] Iacopo Masi et al. “Learning Pose-Aware Models for Pose-Invariant Face Recognition in the Wild”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 41.2 (2019), pp. 379–393. DOI: 10.1109/TPAMI.2018.2792452.
- [162] Scott McCloud and Mark Martin. *Understanding comics: The invisible art*. Vol. 106. Kitchen sink press Northampton, MA, 1993.
- [163] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. “DBpedia spotlight: shedding light on the web of documents”. In: *Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011*. Ed. by Chiara Ghidini, Axel-Cyrille Ngonga Ngomo, Stefanie N. Lindstaedt, and Tassilo Pellegrini. ACM International Conference Proceeding Series. ACM, 2011, pp. 1–8. DOI: 10.1145/2063518.2063519.

- [164] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. “Efficient Estimation of Word Representations in Vector Space”. In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2013. URL: <http://arxiv.org/abs/1301.3781>.
- [165] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. “Distributed Representations of Words and Phrases and their Compositionality”. In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*. Ed. by Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger. 2013, pp. 3111–3119. URL: <http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality>.
- [166] Frederic Morin and Yoshua Bengio. “Hierarchical Probabilistic Neural Network Language Model”. In: *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, AISTATS 2005, Bridgetown, Barbados, January 6-8, 2005*. Ed. by Robert G. Cowell and Zoubin Ghahramani. Society for Artificial Intelligence and Statistics, 2005. URL: <http://www.gatsby.ucl.ac.uk/aistats/fullpapers/208.pdf>.
- [167] Andrea Moro, Alessandro Raganato, and Roberto Navigli. “Entity Linking meets Word Sense Disambiguation: a Unified Approach”. In: *Trans. Assoc. Comput. Linguistics* 2 (2014), pp. 231–244. URL: <https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/291>.
- [170] Markus Mühling, Manja Meister, Nikolaus Korfhage, Jörg Wehling, Angelika Hörth, Ralph Ewerth, and Bernd Freisleben. “Content-based video retrieval in historical collections of the German Broadcasting Archive”. In: *Int. J. on Digital Libraries* 20.2 (2019), pp. 167–183. DOI: 10.1007/s00799-018-0236-z.
- [180] Vinod Nair and Geoffrey E. Hinton. “Rectified Linear Units Improve Restricted Boltzmann Machines”. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*. Ed. by Johannes Fürnkranz and Thorsten Joachims. Omnipress, 2010, pp. 807–814.
- [181] Roberto Navigli and Simone Paolo Ponzetto. “BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network”. In: *Artif. Intell.* 193 (2012), pp. 217–250. DOI: 10.1016/j.artint.2012.07.001.
- [182] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. “Multimodal Deep Learning”. In: *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*. Ed. by Lise Getoor and Tobias Scheffer. Omnipress, 2011, pp. 689–696. URL: [https://icml.cc/2011/papers/399\\_icmlpaper.pdf](https://icml.cc/2011/papers/399_icmlpaper.pdf).
- [183] Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. “J-NERD: Joint Named Entity Recognition and Disambiguation with Rich Linguistic Features”. In: *Trans. Assoc. Comput. Linguistics* 4 (2016), pp. 215–229. URL: <https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/698>.



- [184] Christian Otto, Sebastian Holzki, and Ralph Ewerth. "Is This an Example Image?" - Predicting the Relative Abstractness Level of Image and Text". In: *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part I*. Ed. by Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra. Vol. 11437. Lecture Notes in Computer Science. Springer, 2019, pp. 711–725. DOI: 10.1007/978-3-030-15712-8\_46. URL: [https://doi.org/10.1007/978-3-030-15712-8\\_46](https://doi.org/10.1007/978-3-030-15712-8_46).
- [185] Christian Otto, Matthias Springstein, Avishek Anand, and Ralph Ewerth. "Understanding, Categorizing and Predicting Semantic Image-Text Relations". In: *Proceedings of the 2019 on International Conference on Multimedia Retrieval, ICMR 2019, Ottawa, ON, Canada, June 10-13, 2019*. Ed. by Abdulmotaleb El-Saddik, Alberto Del Bimbo, Zhongfei Zhang, Alexander G. Hauptmann, K. Selçuk Candan, Marco Bertini, Lexing Xie, and Xiao-Yong Wei. ACM, 2019, pp. 168–176. DOI: 10.1145/3323873.3325049.
- [186] Christian Otto, Matthias Springstein, Avishek Anand, and Ralph Ewerth. "Characterization and classification of semantic image-text relations". In: *Int. J. Multim. Inf. Retr.* 9.1 (2020), pp. 31–45. DOI: 10.1007/s13735-019-00187-6.
- [187] Rafael Padilha, Tawfiq Salem, Scott Workman, Fernanda A. Andaló, Anderson Rocha, and Nathan Jacobs. "Content-Based Detection of Temporal Metadata Manipulation". In: *CoRR* abs/2103.04736 (2021). arXiv: 2103.04736. URL: <https://arxiv.org/abs/2103.04736>.
- [188] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. *The PageRank citation ranking: Bringing order to the web*. Tech. rep. Stanford InfoLab, 1999.
- [189] Frank Palermo, James Hays, and Alexei A. Efros. "Dating Historical Color Images". In: *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI*. Ed. by Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid. Vol. 7577. Lecture Notes in Computer Science. Springer, 2012, pp. 499–512. DOI: 10.1007/978-3-642-33783-3\_36. URL: [https://doi.org/10.1007/978-3-642-33783-3\\_36](https://doi.org/10.1007/978-3-642-33783-3_36).
- [190] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. "Deep Face Recognition". In: *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015*. Ed. by Xianghua Xie, Mark W. Jones, and Gary K. L. Tam. BMVA Press, 2015, pp. 41.1–41.12. DOI: 10.5244/C.29.41.
- [191] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "Glove: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. ACL, 2014, pp. 1532–1543. DOI: 10.3115/v1/d14-1162.
- [192] Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. "Semi-supervised sequence tagging with bidirectional language models". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. Ed. by Regina Barzilay and Min-Yen Kan. Association for Computational Linguistics, 2017, pp. 1756–1765. DOI: 10.18653/v1/P17-1161.

- [193] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. “Deep Contextualized Word Representations”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*. Ed. by Marilyn A. Walker, Heng Ji, and Amanda Stent. Association for Computational Linguistics, 2018, pp. 2227–2237. DOI: 10.18653/v1/n18-1202.
- [194] Francesco Piccinno and Paolo Ferragina. “From TagME to WAT: a new entity annotator”. In: *ERD’14, Proceedings of the First ACM International Workshop on Entity Recognition & Disambiguation, July 11, 2014, Gold Coast, Queensland, Australia*. Ed. by David Carmel, Ming-Wei Chang, Evgeniy Gabrilovich, Bo-June Paul Hsu, and Kuansan Wang. ACM, 2014, pp. 55–62. DOI: 10.1145/2633211.2634350.
- [196] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*. Ed. by Asli Çelikyilmaz and Tsung-Hsien Wen. Association for Computational Linguistics, 2020, pp. 101–108. URL: <https://www.aclweb.org/anthology/2020.acl-demos.14/>.
- [197] Xianbiao Qi and Lei Zhang. “Face Recognition via Centralized Coordinate Learning”. In: *CoRR* abs/1801.05678 (2018). arXiv: 1801.05678. URL: <http://arxiv.org/abs/1801.05678>.
- [198] Ning Qian. “On the momentum term in gradient descent learning algorithms”. In: *Neural Networks* 12.1 (1999), pp. 145–151. DOI: 10.1016/S0893-6080(98)00116-6.
- [199] Yichen Qian, Weihong Deng, and Jiani Hu. “Task Specific Networks for Identity and Face Variation”. In: *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi’an, China, May 15-19, 2018*. IEEE Computer Society, 2018, pp. 271–277. DOI: 10.1109/FG.2018.00047.
- [200] Till Quack, Bastian Leibe, and Luc Van Gool. “World-scale mining of objects and events from community photo collections”. In: *Proceedings of the 7th ACM International Conference on Image and Video Retrieval, CIVR 2008, Niagara Falls, Canada, July 7-9, 2008*. Ed. by Jiebo Luo, Ling Guan, Alan Hanjalic, Mohan S. Kankanhalli, and Ivan Lee. ACM, 2008, pp. 47–56. DOI: 10.1145/1386352.1386363.
- [201] J. Ross Quinlan. “Induction of Decision Trees”. In: *Mach. Learn.* 1.1 (1986), pp. 81–106. DOI: 10.1023/A:1022643204877.
- [202] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. “CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples”. In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Vol. 9905. Lecture Notes in Computer Science. Springer, 2016, pp. 3–20. DOI: 10.1007/978-3-319-46448-0\_1. URL: [https://doi.org/10.1007/978-3-319-46448-0\\_1](https://doi.org/10.1007/978-3-319-46448-0_1).

- [203] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. “Improving language understanding by generative pre-training”. In: *URL [https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf)* (2018).
- [204] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. “Language models are unsupervised multitask learners”. In: *OpenAI Blog* 1.8 (2019), p. 9.
- [205] Jonathan Raiman and Olivier Raiman. “DeepType: Multilingual Entity Linking by Neural Type System Evolution”. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, 2018, pp. 5406–5413. URL: <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17148>.
- [206] Srikumar Ramalingam, Sofien Bouaziz, Peter F. Sturm, and Matthew Brand. “SKYLINE2GPS: Localization in urban canyons using omni-skylines”. In: *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 18-22, 2010, Taipei, Taiwan*. IEEE, 2010, pp. 3816–3823. DOI: 10.1109/IRoS.2010.5649105.
- [207] Arnau Ramisa, Fei Yan, Francesc Moreno-Noguer, and Krystian Mikolajczyk. “BreakingNews: Article Annotation by Image and Text Processing”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 40.5 (2018), pp. 1072–1085. DOI: 10.1109/TPAMI.2017.2721945.
- [208] Rajeev Ranjan, Swami Sankaranarayanan, Ankan Bansal, Navaneeth Bodla, Jun-Cheng Chen, Vishal M. Patel, Carlos Domingo Castillo, and Rama Chellappa. “Deep Learning for Understanding Faces: Machines May Be Just as Good, or Better, than Humans”. In: *IEEE Signal Process. Mag.* 35.1 (2018), pp. 66–83. DOI: 10.1109/MSP.2017.2764116.
- [209] Lev-Arie Ratinov and Dan Roth. “Design Challenges and Misconceptions in Named Entity Recognition”. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning, CoNLL 2009, Boulder, Colorado, USA, June 4-5, 2009*. Ed. by Suzanne Stevenson and Xavier Carreras. ACL, 2009, pp. 147–155. URL: <https://www.aclweb.org/anthology/W09-1119/>.
- [210] Joseph Redmon and Ali Farhadi. “YOLO9000: Better, Faster, Stronger”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 6517–6525. DOI: 10.1109/CVPR.2017.690.
- [211] Timo Reuter, Symeon Papadopoulos, Georgios Petkos, Vasileios Mezaris, Yiannis Kompatsiaris, Philipp Cimiano, Christopher M. De Vries, and Shlomo Geva. “Social Event Detection at MediaEval 2013: Challenges, Datasets, and Evaluation”. In: *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, October 18-19, 2013*. Ed. by Martha A. Larson, Xavier Anguera, Timo Reuter, Gareth J. F. Jones, Bogdan Ionescu, Markus Schedl, Tomas Piatrik, Claudia Hauff, and Mohammad Soleymani. Vol. 1043. CEUR Workshop Proceedings. CEUR-WS.org, 2013. URL: [http://ceur-ws.org/Vol-1043/mediaeval2013\\_submission\\_9.pdf](http://ceur-ws.org/Vol-1043/mediaeval2013_submission_9.pdf).

- [212] Giuseppe Rizzo and Raphaël Troncy. “NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools”. In: *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012*. Ed. by Walter Daelemans, Mirella Lapata, and Lluís Màrquez. The Association for Computer Linguistics, 2012, pp. 73–76. URL: <https://www.aclweb.org/anthology/E12-2015/>.
- [213] Herbert Robbins and Sutton Monro. “A Stochastic Approximation Method”. In: *The annals of mathematical statistics* 22 (1951), pp. 400–407. ISSN: 0003-4851. DOI: 10.1214/aoms/1177729586.
- [214] Richard Rogers. “Debanalizing Twitter: the transformation of an object of study”. In: *Web Science 2013 (co-located with ECRC), WebSci '13, Paris, France, May 2-4, 2013*. Ed. by Hugh C. Davis, Harry Halpin, Alex Pentland, Mark Bernstein, and Lada A. Adamic. ACM, 2013, pp. 356–365. DOI: 10.1145/2464464.2464511.
- [215] Frank Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological review* 65.6 (1958), p. 386. DOI: 10.1037/h0042519.
- [216] Sebastian Ruder. “An Overview of Multi-Task Learning in Deep Neural Networks”. In: *CoRR* abs/1706.05098 (2017). arXiv: 1706.05098. URL: <http://arxiv.org/abs/1706.05098>.
- [217] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. “Learning representations by back-propagating errors”. In: *nature* 323.6088 (1986), pp. 533–536. ISSN: 0028-0836. DOI: 10.1038/323533a0.
- [218] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. “ImageNet Large Scale Visual Recognition Challenge”. In: *Int. J. Comput. Vis.* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.
- [219] Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan. “Deep Multimodal Image-Repurposing Detection”. In: *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*. Ed. by Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei. ACM, 2018, pp. 1337–1345. DOI: 10.1145/3240508.3240707.
- [220] Tawfiq Salem, Scott Workman, and Nathan Jacobs. “Learning a Dynamic Map of Visual Appearance”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020, pp. 12432–12441. DOI: 10.1109/CVPR42600.2020.01245.
- [221] Tawfiq Salem, Scott Workman, Menghua Zhai, and Nathan Jacobs. “Analyzing human appearance as a cue for dating images”. In: *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016*. IEEE Computer Society, 2016, pp. 1–8. DOI: 10.1109/WACV.2016.7477678.
- [222] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. “MobileNetV2: Inverted Residuals and Linear Bottlenecks”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 4510–4520. DOI: 10.1109/CVPR.2018.00474. URL: <http://openaccess.thecvf.com/>

- content\_cvpr\_2018/html/Sandler\_MobileNetV2\_Inverted\_Residuals\_CVPR\_2018\_paper.html.
- [223] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”. In: *CoRR* abs/1910.01108 (2019). arXiv: 1910.01108. URL: <http://arxiv.org/abs/1910.01108>.
- [224] Swami Sankaranarayanan, Azadeh Alavi, Carlos Domingo Castillo, and Rama Chelappa. “Triplet probabilistic embedding for face verification and clustering”. In: *8th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2016, Niagara Falls, NY, USA, September 6-9, 2016*. IEEE, 2016, pp. 1–8. DOI: 10.1109/BTAS.2016.7791205.
- [225] Olivier Saurer, Georges Baatz, Kevin Köser, Lubor Ladicky, and Marc Pollefeys. “Image Based Geo-localization in the Alps”. In: *Int. J. Comput. Vis.* 116.3 (2016), pp. 213–225. DOI: 10.1007/s11263-015-0830-0.
- [226] Grant Schindler, Matthew A. Brown, and Richard Szeliski. “City-Scale Location Recognition”. In: *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*. IEEE Computer Society, 2007. DOI: 10.1109/CVPR.2007.383150.
- [227] Grant Schindler, Frank Dellaert, and Sing Bing Kang. “Inferring Temporal Order of Images From 3D Structure”. In: *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*. IEEE Computer Society, 2007. DOI: 10.1109/CVPR.2007.383088.
- [228] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “FaceNet: A unified embedding for face recognition and clustering”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682.
- [229] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. “CPLaNet: Enhancing Image Geolocalization by Combinatorial Partitioning of Maps”. In: *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*. Ed. by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss. Vol. 11214. Lecture Notes in Computer Science. Springer, 2018, pp. 544–560. DOI: 10.1007/978-3-030-01249-6\_33. URL: [https://doi.org/10.1007/978-3-030-01249-6\\_33](https://doi.org/10.1007/978-3-030-01249-6_33).
- [230] Hatem Mousselly Sergieh, Daniel Watzinger, Bastian Huber, Mario Döllner, Elöd Egyed-Zsigmond, and Harald Kosch. “World-wide scale geotagged image dataset for automatic image annotation and reverse geotagging”. In: *Multimedia Systems Conference 2014, MMSys '14, Singapore, March 19-21, 2014*. Ed. by Roger Zimmermann. ACM, 2014, pp. 47–52. DOI: 10.1145/2557642.2563673.
- [231] Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. “Neural Entity Linking: A Survey of Models based on Deep Learning”. In: *CoRR* abs/2006.00575 (2020). arXiv: 2006.00575. URL: <https://arxiv.org/abs/2006.00575>.

- [232] Qi Shan, Changchang Wu, Brian Curless, Yasutaka Furukawa, Carlos Hernández, and Steven M. Seitz. “Accurate Geo-Registration by Ground-to-Aerial Image Matching”. In: *2nd International Conference on 3D Vision, 3DV 2014, Tokyo, Japan, December 8-11, 2014, Volume 1*. IEEE Computer Society, 2014, pp. 525–532. DOI: 10.1109/3DV.2014.69.
- [233] Wei Shen, Jianyong Wang, and Jiawei Han. “Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions”. In: *IEEE Trans. Knowl. Data Eng.* 27.2 (2015), pp. 443–460. DOI: 10.1109/TKDE.2014.2327028.
- [234] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. “FaceID-GAN: Learning a Symmetry Three-Player GAN for Identity-Preserving Face Synthesis”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 821–830. DOI: 10.1109/CVPR.2018.00092. URL: [http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Shen\\_FaceID-GAN\\_Learning\\_a\\_CVPR\\_2018\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Shen_FaceID-GAN_Learning_a_CVPR_2018_paper.html).
- [235] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: <http://arxiv.org/abs/1409.1556>.
- [236] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. “Towards VQA Models That Can Read”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 8317–8326. DOI: 10.1109/CVPR.2019.00851. URL: [http://openaccess.thecvf.com/content\\_CVPR\\_2019/html/Singh\\_Towards\\_VQA\\_Models\\_That\\_Can\\_Read\\_CVPR\\_2019\\_paper.html](http://openaccess.thecvf.com/content_CVPR_2019/html/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.html).
- [237] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh C. Jain. “Content-Based Image Retrieval at the End of the Early Years”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 22.12 (2000), pp. 1349–1380. DOI: 10.1109/34.895972.
- [240] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. “Dropout: a simple way to prevent neural networks from overfitting”. In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 1929–1958. URL: <http://dl.acm.org/citation.cfm?id=2670313>.
- [241] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. “Yago: a core of semantic knowledge”. In: *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*. Ed. by Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy. ACM, 2007, pp. 697–706. DOI: 10.1145/1242572.1242667.
- [242] Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. “Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation”. In: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*. Ed. by Qiang Yang and Michael J. Wooldridge. AAAI Press, 2015, pp. 1333–1339. URL: <http://ijcai.org/Abstract/15/192>.

- [243] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. “Deep Learning Face Representation by Joint Identification-Verification”. In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Ed. by Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger. 2014, pp. 1988–1996. URL: <http://papers.nips.cc/paper/5416-deep-learning-face-representation-by-joint-identification-verification>.
- [244] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. “DeepID3: Face Recognition with Very Deep Neural Networks”. In: *CoRR* abs/1502.00873 (2015). arXiv: 1502.00873. URL: <http://arxiv.org/abs/1502.00873>.
- [245] Yi Sun, Xiaogang Wang, and Xiaoou Tang. “Deep Learning Face Representation from Predicting 10, 000 Classes”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, 2014, pp. 1891–1898. DOI: 10.1109/CVPR.2014.244.
- [246] Yi Sun, Xiaogang Wang, and Xiaoou Tang. “Deeply learned face representations are sparse, selective, and robust”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 2892–2900. DOI: 10.1109/CVPR.2015.7298907.
- [247] Yi Sun, Xiaogang Wang, and Xiaoou Tang. “Sparsifying Neural Network Connections for Face Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 4856–4864. DOI: 10.1109/CVPR.2016.525.
- [248] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. “ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding”. In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 8968–8975. URL: <https://aaai.org/ojs/index.php/AAAI/article/view/6428>.
- [249] Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton. “On the importance of initialization and momentum in deep learning”. In: *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*. Vol. 28. JMLR Workshop and Conference Proceedings. JMLR.org, 2013, pp. 1139–1147. URL: <http://proceedings.mlr.press/v28/sutskever13.html>.
- [250] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning”. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. Ed. by Satinder P. Singh and Shaul Markovitch. AAAI Press, 2017, pp. 4278–4284. URL: <http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14806>.
- [251] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. “Going deeper with convolutions”. In: *IEEE Conference on Computer Vision and Pattern*

- Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594.
- [252] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. “Rethinking the Inception Architecture for Computer Vision”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308.
- [255] Mingxing Tan and Quoc V. Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 6105–6114. URL: <http://proceedings.mlr.press/v97/tan19a.html>.
- [256] Edson C Tandoc Jr and Erika Johnson. “Most students get breaking news first from Twitter”. In: *Newspaper research journal* 37.2 (2016), pp. 153–166. DOI: 10.1177/0739532916648961.
- [258] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. “YFCC100M: the new data in multimedia research”. In: *Commun. ACM* 59.2 (2016), pp. 64–73. DOI: 10.1145/2812802.
- [259] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. “A Closer Look at Spatiotemporal Convolutions for Action Recognition”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 6450–6459. DOI: 10.1109/CVPR.2018.00675. URL: [http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Tran\\_A\\_Closer\\_Look\\_CVPR\\_2018\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Tran_A_Closer_Look_CVPR_2018_paper.html).
- [260] Luan Tran, Xi Yin, and Xiaoming Liu. “Disentangled Representation Learning GAN for Pose-Invariant Face Recognition”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 1283–1292. DOI: 10.1109/CVPR.2017.141.
- [261] Eric Tzeng, Andrew Zhai, Matthew Clements, Raphael Townshend, and Avideh Zakhor. “User-Driven Geolocation of Untagged Desert Imagery Using Digital Elevation Models”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2013, Portland, OR, USA, June 23-28, 2013*. IEEE Computer Society, 2013, pp. 237–244. DOI: 10.1109/CVPRW.2013.42.
- [262] Len Unsworth. “Image/text relations and intersemiosis: Towards multimodal text description for multiliteracies education”. In: *Proceedings of the 33rd International Systemic Functional Congress*. 2007, pp. 1165–1205.
- [263] Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Michael Röder, Daniel Gerber, Sandro Athaide Coelho, Sören Auer, and Andreas Both. “AGDISTIS - Graph-Based Disambiguation of Named Entities Using Linked Data”. In: *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*. Ed. by Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig A. Knoblock, Denny Vrandečić, Paul Groth, Natasha F. Noy, Krzysztof Janowicz, and Carole A. Goble. Vol. 8796. Lecture Notes in Computer Sci-



- ence. Springer, 2014, pp. 457–471. DOI: 10.1007/978-3-319-11964-9\_29. URL: [https://doi.org/10.1007/978-3-319-11964-9\\_29](https://doi.org/10.1007/978-3-319-11964-9_29).
- [264] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett. 2017, pp. 5998–6008. URL: <http://papers.neurips.cc/paper/7181-attention-is-all-you-need.pdf>.
- [265] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. “Show and tell: A neural image caption generator”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 3156–3164. DOI: 10.1109/CVPR.2015.7298935.
- [266] Nam N. Vo and James Hays. “Localizing and Orienting Street Views Using Overhead Imagery”. In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Vol. 9905. Lecture Notes in Computer Science. Springer, 2016, pp. 494–509. DOI: 10.1007/978-3-319-46448-0\_30. URL: [https://doi.org/10.1007/978-3-319-46448-0\\_30](https://doi.org/10.1007/978-3-319-46448-0_30).
- [267] Nam N. Vo, Nathan Jacobs, and James Hays. “Revisiting IM2GPS in the Deep Learning Era”. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2640–2649. DOI: 10.1109/ICCV.2017.286.
- [268] Denny Vrandečić and Markus Krötzsch. *Wikidata: a free collaborative knowledge base*. 2014. DOI: 10.1145/2629489.
- [269] Ji Wan, Dayong Wang, Steven Chu-Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. “Deep Learning for Content-Based Image Retrieval: A Comprehensive Study”. In: *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*. Ed. by Kien A. Hua, Yong Rui, Ralf Steinmetz, Alan Hanjalic, Apostol Natsev, and Wenwu Zhu. ACM, 2014, pp. 157–166. DOI: 10.1145/2647868.2654948.
- [270] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. “The Devil of Face Recognition Is in the Noise”. In: *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX*. Ed. by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss. Vol. 11213. Lecture Notes in Computer Science. Springer, 2018, pp. 780–795. DOI: 10.1007/978-3-030-01240-3\_47. URL: [https://doi.org/10.1007/978-3-030-01240-3\\_47](https://doi.org/10.1007/978-3-030-01240-3_47).
- [271] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. “Additive Margin Softmax for Face Verification”. In: *IEEE Signal Process. Lett.* 25.7 (2018), pp. 926–930. DOI: 10.1109/LSP.2018.2822810.
- [272] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. “NormFace: L<sub>2</sub> Hypersphere Embedding for Face Verification”. In: *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*. Ed. by Qiong Liu, Rainer Lienhart, Haohong Wang, Sheng-Wei "Kuan-Ta" Chen,

- Susanne Boll, Yi-Ping Phoebe Chen, Gerald Friedland, Jia Li, and Shuicheng Yan. ACM, 2017, pp. 1041–1049. DOI: 10.1145/3123266.3123359.
- [273] Limin Wang, Zhe Wang, Yu Qiao, and Luc Van Gool. “Transferring Deep Object and Scene Representations for Event Recognition in Still Images”. In: *Int. J. Comput. Vis.* 126.2-4 (2018), pp. 390–409. DOI: 10.1007/s11263-017-1043-5.
- [274] Yang Wang and Liangliang Cao. “Discovering Latent Clusters from Geotagged Beach Images”. In: *Advances in Multimedia Modeling, 19th International Conference, MMM 2013, Huangshan, China, January 7-9, 2013, Proceedings, Part II*. Ed. by Shipeng Li, Abdulmotaleb El-Saddik, Meng Wang, Tao Mei, Nicu Sebe, Shuicheng Yan, Richang Hong, and Cathal Gurrin. Vol. 7733. Lecture Notes in Computer Science. Springer, 2013, pp. 133–142. DOI: 10.1007/978-3-642-35728-2\_13. URL: [https://doi.org/10.1007/978-3-642-35728-2\\_13](https://doi.org/10.1007/978-3-642-35728-2_13).
- [275] Zongwei Wang, Xu Tang, Weixin Luo, and Shenghua Gao. “Face Aging With Identity-Preserved Conditional Generative Adversarial Networks”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 7939–7947. DOI: 10.1109/CVPR.2018.00828. URL: [http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Wang\\_Face\\_Aging\\_With\\_CVPR\\_2018\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Face_Aging_With_CVPR_2018_paper.html).
- [276] Joe H Ward Jr. “Hierarchical grouping to optimize an objective function”. In: *Journal of the American statistical association* 58.301 (1963), pp. 236–244. DOI: 10.1080/01621459.1963.10500845.
- [277] Yandong Wen, Zhifeng Li, and Yu Qiao. “Latent Factor Guided Convolutional Neural Networks for Age-Invariant Face Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 4893–4901. DOI: 10.1109/CVPR.2016.529.
- [278] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. “A Discriminative Feature Learning Approach for Deep Face Recognition”. In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Vol. 9911. Lecture Notes in Computer Science. Springer, 2016, pp. 499–515. DOI: 10.1007/978-3-319-46478-7\_31. URL: [https://doi.org/10.1007/978-3-319-46478-7\\_31](https://doi.org/10.1007/978-3-319-46478-7_31).
- [279] Tobias Weyand, Ilya Kostrikov, and James Philbin. “PlaNet - Photo Geolocation with Convolutional Neural Networks”. In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Vol. 9912. Lecture Notes in Computer Science. Springer, 2016, pp. 37–55. DOI: 10.1007/978-3-319-46484-8\_3. URL: [https://doi.org/10.1007/978-3-319-46484-8\\_3](https://doi.org/10.1007/978-3-319-46484-8_3).
- [280] Scott Workman, Richard Souvenir, and Nathan Jacobs. “Wide-Area Image Geolocalization with Aerial Reference Imagery”. In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 3961–3969. DOI: 10.1109/ICCV.2015.451.

- [281] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. “Scalable Zero-shot Entity Linking with Dense Entity Retrieval”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Association for Computational Linguistics, 2020, pp. 6397–6407. DOI: 10.18653/v1/2020.emnlp-main.519.
- [282] Yue Wu and Qiang Ji. “Facial Landmark Detection: A Literature Survey”. In: *Int. J. Comput. Vis.* 127.2 (2019), pp. 115–142. DOI: 10.1007/s11263-018-1097-z.
- [283] Yue Wu, Hongfu Liu, Jun Li, and Yun Fu. “Deep Face Recognition with Center Invariant Loss”. In: *Proceedings of the on Thematic Workshops of ACM Multimedia 2017, Mountain View, CA, USA, October 23 - 27, 2017*. Ed. by Wanmin Wu, Jianchao Yang, Qi Tian, and Roger Zimmermann. ACM, 2017, pp. 408–414. DOI: 10.1145/3126686.3126693.
- [284] Zifeng Wu, Yongzhen Huang, and Liang Wang. “Learning Representative Deep Features for Image Set Analysis”. In: *IEEE Trans. Multimedia* 17.11 (2015), pp. 1960–1968. DOI: 10.1109/TMM.2015.2477681.
- [285] Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. “Self-Training With Noisy Student Improves ImageNet Classification”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020, pp. 10684–10695. DOI: 10.1109/CVPR42600.2020.01070.
- [286] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. “Aggregated Residual Transformations for Deep Neural Networks”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 5987–5995. DOI: 10.1109/CVPR.2017.634.
- [287] Yuanjun Xiong, Kai Zhu, Dahua Lin, and Xiaoou Tang. “Recognize complex events from static images by fusing deep channels”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 1600–1609. DOI: 10.1109/CVPR.2015.7298768.
- [288] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. “Empirical Evaluation of Rectified Activations in Convolutional Network”. In: *CoRR* abs/1505.00853 (2015). arXiv: 1505.00853. URL: <http://arxiv.org/abs/1505.00853>.
- [289] Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. “A discriminative CNN video representation for event detection”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 1798–1807. DOI: 10.1109/CVPR.2015.7298789.
- [290] Hongyu Yang, Di Huang, Yunhong Wang, and Anil K. Jain. “Learning Face Age Progression: A Pyramid Architecture of GANs”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 31–39. DOI: 10.1109/CVPR.2018.00011. URL: [http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Yang\\_Learning\\_Face\\_Age\\_CVPR\\_2018\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Yang_Learning_Face_Age_CVPR_2018_paper.html).

- [291] Jimei Yang, Scott E. Reed, Ming-Hsuan Yang, and Honglak Lee. “Weakly-supervised Disentangling with Recurrent Transformations for 3D View Synthesis”. In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. Ed. by Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett. 2015, pp. 1099–1107. URL: <http://papers.nips.cc/paper/5639-weakly-supervised-disentangling-with-recurrent-transformations-for-3d-view-synthesis>.
- [292] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. “From Facial Parts Responses to Face Detection: A Deep Learning Approach”. In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 3676–3684. DOI: 10.1109/ICCV.2015.419.
- [293] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. “XLNet: Generalized Autoregressive Pretraining for Language Understanding”. In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett. 2019, pp. 5754–5764. URL: <http://papers.nips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding>.
- [294] K. Ye, N. Honarvar Nazari, J. Hahn, Z. Hussain, M. Zhang, and A. Kovashka. “Interpreting the Rhetoric of Visual Advertisements”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), pp. 1–1. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2019.2947440.
- [295] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. “Learning Face Representation from Scratch”. In: *CoRR* abs/1411.7923 (2014). arXiv: 1411.7923. URL: <http://arxiv.org/abs/1411.7923>.
- [296] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Du-Sik Park, and Junmo Kim. “Rotating your face using multi-task deep neural network”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 676–684. DOI: 10.1109/CVPR.2015.7298667.
- [297] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. “Towards Large-Pose Face Frontalization in the Wild”. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 4010–4019. DOI: 10.1109/ICCV.2017.430.
- [298] Xiaodong Yu, Stephen Mayhew, Mark Sammons, and Dan Roth. “On the Strength of Character Language Models for Multilingual Named Entity Recognition”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. Ed. by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii. Association for Computational Linguistics, 2018, pp. 3073–3077. DOI: 10.18653/v1/d18-1345.

- [299] Sangdoon Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. “CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features”. In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 6022–6031. DOI: 10.1109/ICCV.2019.00612.
- [300] Sergey Zagoruyko and Nikos Komodakis. “Wide Residual Networks”. In: *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*. Ed. by Richard C. Wilson, Edwin R. Hancock, and William A. P. Smith. BMVA Press, 2016. URL: <http://www.bmva.org/bmvc/2016/papers/paper087/index.html>.
- [301] Amir Roshan Zamir and Mubarak Shah. “Accurate Image Localization Based on Google Maps Street View”. In: *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*. Ed. by Kostas Daniilidis, Petros Maragos, and Nikos Paragios. Vol. 6314. Lecture Notes in Computer Science. Springer, 2010, pp. 255–268. DOI: 10.1007/978-3-642-15561-1\_19. URL: [https://doi.org/10.1007/978-3-642-15561-1\\_19](https://doi.org/10.1007/978-3-642-15561-1_19).
- [302] Amir Roshan Zamir and Mubarak Shah. “Image Geo-Localization Based on Multiple Nearest Neighbor Feature Matching Using Generalized Graphs”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 36.8 (2014), pp. 1546–1558. DOI: 10.1109/TPAMI.2014.2299799.
- [303] Matthew D. Zeiler and Rob Fergus. “Visualizing and Understanding Convolutional Networks”. In: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*. Ed. by David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars. Vol. 8689. Lecture Notes in Computer Science. Springer, 2014, pp. 818–833. DOI: 10.1007/978-3-319-10590-1\_53. URL: [https://doi.org/10.1007/978-3-319-10590-1\\_53](https://doi.org/10.1007/978-3-319-10590-1_53).
- [304] Eyasu Zemene, Yonatan Tariku Tesfaye, Haroon Idrees, Andrea Prati, Marcello Pelillo, and Mubarak Shah. “Large-Scale Image Geo-Localization Using Dominant Sets”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 41.1 (2019), pp. 148–161. DOI: 10.1109/TPAMI.2017.2787132.
- [305] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. “Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks”. In: *IEEE Signal Process. Lett.* 23.10 (2016), pp. 1499–1503. DOI: 10.1109/LSP.2016.2603342.
- [306] Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. “Equal But Not The Same: Understanding the Implicit Relationship Between Persuasive Images and Text”. In: *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. BMVA Press, 2018, p. 8. URL: <http://bmvc2018.org/contents/papers/0228.pdf>.
- [307] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. “BERTScore: Evaluating Text Generation with BERT”. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL: <https://openreview.net/forum?id=SkeHuCVFDr>.

- [308] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. “Range Loss for Deep Face Recognition with Long-Tailed Training Data”. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 5419–5428. DOI: 10.1109/ICCV.2017.578.
- [309] Yizhe Zhang, Ming Shao, Edward K. Wong, and Yun Fu. “Random Faces Guided Sparse Many-to-One Encoder for Pose-Invariant Face Recognition”. In: *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. IEEE Computer Society, 2013, pp. 2416–2423. DOI: 10.1109/ICCV.2013.300.
- [310] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. “Learning Deep Representation for Face Alignment with Auxiliary Attributes”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 38.5 (2016), pp. 918–930. DOI: 10.1109/TPAMI.2015.2469286.
- [311] Jian Zhao, Lin Xiong, Jayashree Karlekar, Jianshu Li, Fang Zhao, Zhecan Wang, Sugiri Pranata, Shengmei Shen, Shuicheng Yan, and Jiashi Feng. “Dual-Agent GANs for Photorealistic and Identity Preserving Profile Face Synthesis”. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett. 2017, pp. 66–76. URL: <http://papers.nips.cc/paper/6612-dual-agent-gans-for-photorealistic-and-identity-preserving-profile-face-synthesis>.
- [312] Tianyue Zheng, Weihong Deng, and Jiani Hu. “Age Estimation Guided Convolutional Neural Network for Age-Invariant Face Recognition”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 503–511. DOI: 10.1109/CVPRW.2017.77.
- [313] Yantao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddemeier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, and Hartmut Neven. “Tour the world: Building a web-scale landmark recognition engine”. In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 2009, pp. 1085–1092. DOI: 10.1109/CVPR.2009.5206749.
- [314] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. “Learning Deep Features for Discriminative Localization”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 2921–2929. DOI: 10.1109/CVPR.2016.319.
- [315] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. “Places: A 10 Million Image Database for Scene Recognition”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 40.6 (2018), pp. 1452–1464. DOI: 10.1109/TPAMI.2017.2723009.
- [316] Erjin Zhou, Zhimin Cao, and Jian Sun. “GridFace: Face Rectification via Learning Local Homography Transformations”. In: *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*. Ed. by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss.

- Vol. 11220. Lecture Notes in Computer Science. Springer, 2018, pp. 3–20. DOI: 10.1007/978-3-030-01270-0\_1. URL: [https://doi.org/10.1007/978-3-030-01270-0\\_1](https://doi.org/10.1007/978-3-030-01270-0_1).
- [317] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. “Face Alignment Across Large Poses: A 3D Solution”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 146–155. DOI: 10.1109/CVPR.2016.23.
- [318] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. “Deep Learning Identity-Preserving Face Space”. In: *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. IEEE Computer Society, 2013, pp. 113–120. DOI: 10.1109/ICCV.2013.21.
- [319] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. “Multi-View Perceptron: a Deep Model for Learning Face Identity and View Representations”. In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Ed. by Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger. 2014, pp. 217–225. URL: <http://papers.nips.cc/paper/5546-multi-view-perceptron-a-deep-model-for-learning-face-identity-and-view-representations>.
- [320] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. “Recover Canonical-View Faces in the Wild with Deep Neural Networks”. In: *CoRR* abs/1404.3543 (2014). arXiv: 1404.3543. URL: <http://arxiv.org/abs/1404.3543>.
- [321] Barret Zoph and Quoc V. Le. “Neural Architecture Search with Reinforcement Learning”. In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL: <https://openreview.net/forum?id=r1Ue8Hcxg>.
- [322] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. “Learning Transferable Architectures for Scalable Image Recognition”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 8697–8710. DOI: 10.1109/CVPR.2018.00907. URL: [http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Zoph\\_Learning\\_Transferable\\_Architectures\\_CVPR\\_2018\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Zoph_Learning_Transferable_Architectures_CVPR_2018_paper.html).





# A Appendix

## A.1 Event Classification of Photos

This section contains detailed dataset statistics for the train and test datasets (Appendix A.1.1) as well as the results using different inference strategies (Appendix A.1.2).

### A.1.1 Detailed Dataset Statistics

We presented the *Visual Event Classification Dataset* (VisE-D) including two test datasets in Section 3.1.2.3 of this thesis. In this section, detailed statistics on the image distribution for the *Leaf Event Nodes* are provided in Figure A.1 – A.3. The illustrations also provide the complete list of *Leaf Event Nodes*.

### A.1.2 Results using other Inference Strategies

We have evaluated the ontology-driven approaches using an inference strategy that combines two different probabilities  $\hat{\mathbf{y}}_L = \hat{\mathbf{y}}_L^o \odot \hat{\mathbf{y}}_L^{cos}$  (Section 3.1.3.3). The results using the individual probabilities  $\hat{\mathbf{y}}_L^o$  or  $\hat{\mathbf{y}}_L^{cos}$  are provided in Table A.1 and Table A.2.

In general, the probabilities  $\hat{\mathbf{y}}_L^o$  provide slightly better results, in particular for the top-3 and top-5 accuracy. We argue that *Leaf Event Nodes* with shorter paths in the *Ontology* tend to achieve higher probabilities  $\hat{\mathbf{y}}_L^{cos}$ , as the overall (accumulated) weight of *Branch Event Nodes* is lower for the respective *Subgraph*. However, similar results are achieved in comparison to the reported numbers of the combined strategy presented in Table 3.2. Thus, the results allow the same conclusion with respect to the overall performance of the ontology-driven loss functions and weighting schemes.

## A Appendix

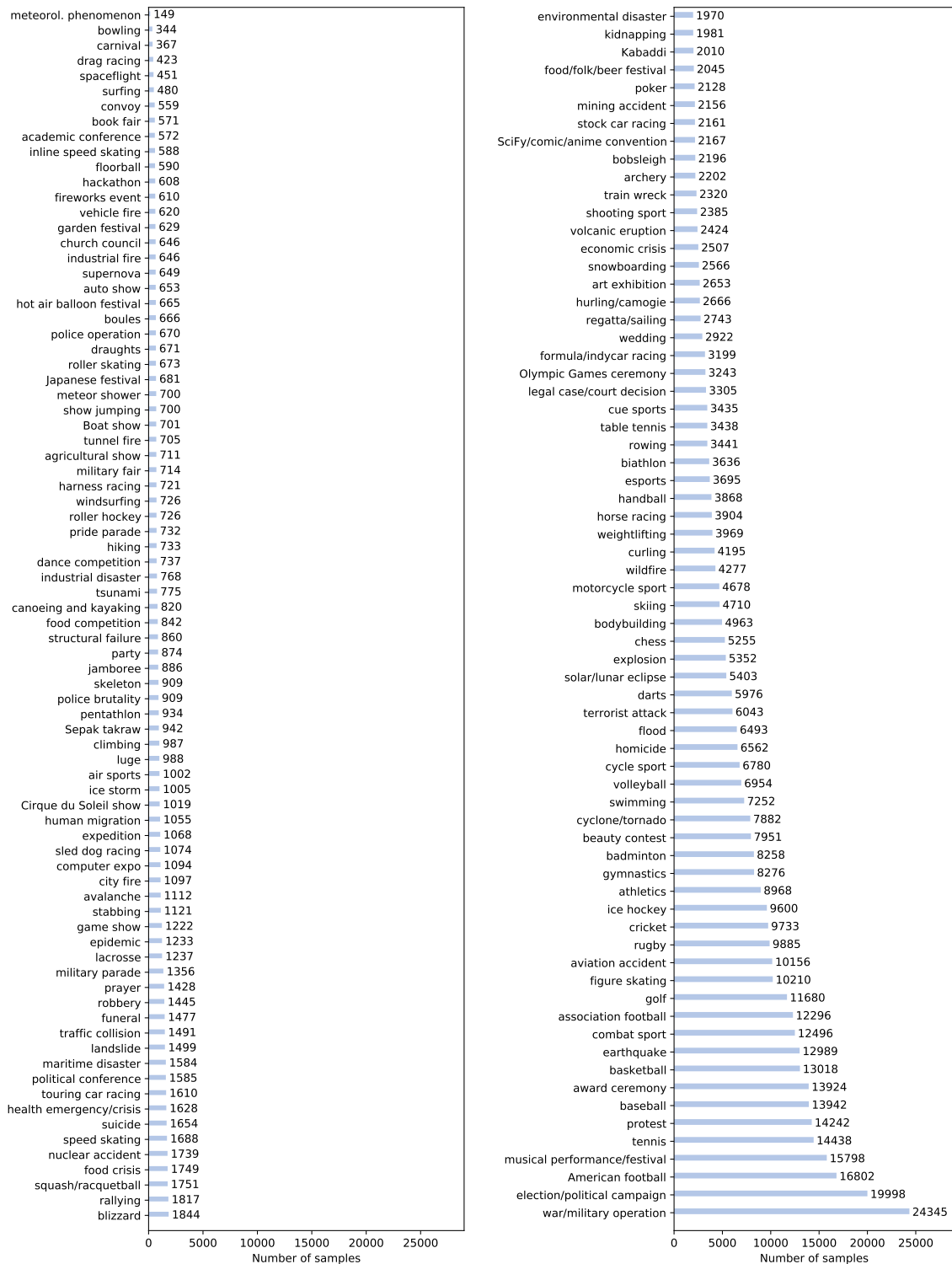


Figure A.1: Number of training images for all *Leaf Event Nodes* in the *Visual Event Classification Dataset* (VisE-D)

## A.1 Event Classification of Photos

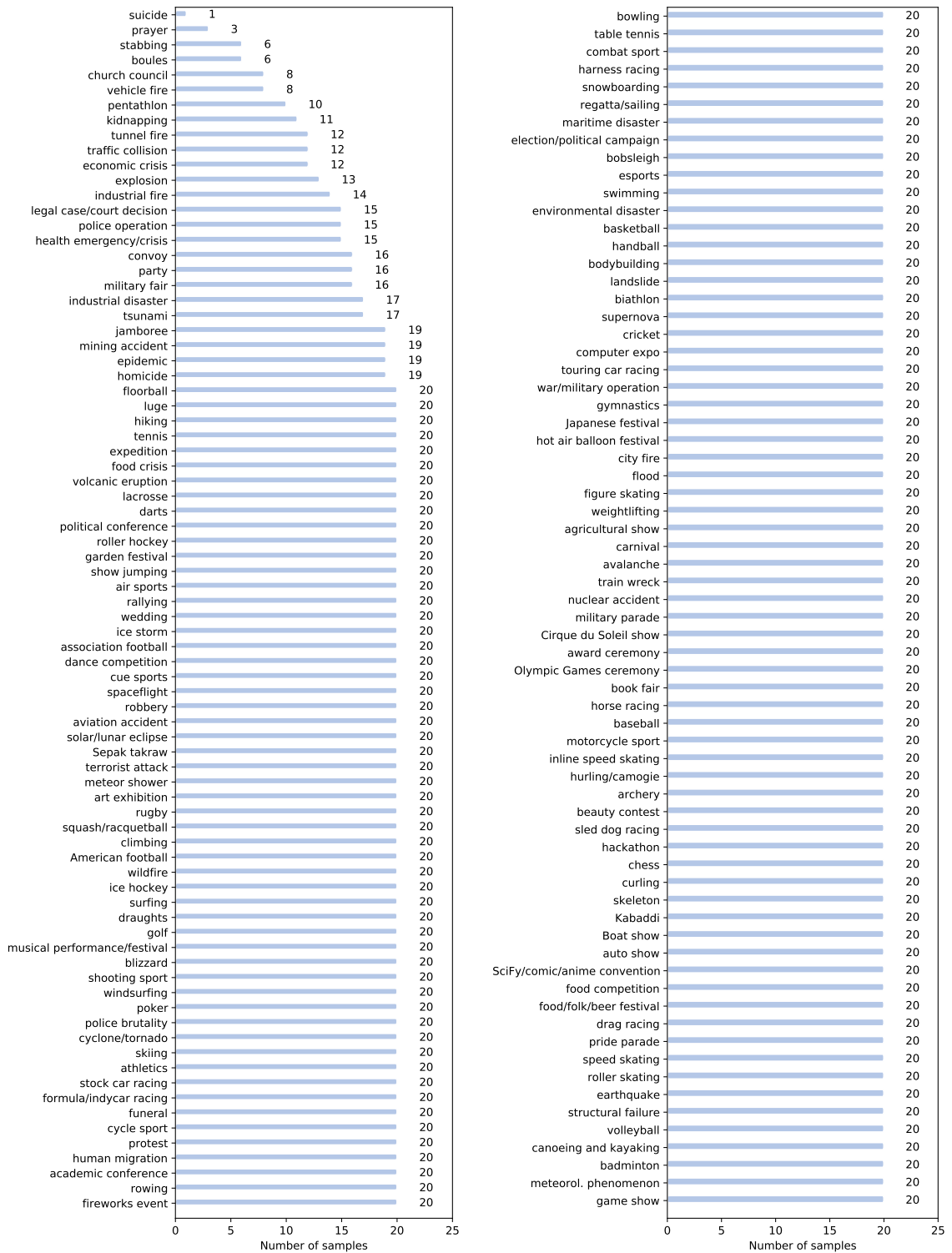


Figure A.2: Number of images for all *Leaf Event Nodes* in the manually annotated *VisE-Bing* test dataset.

## A Appendix



Figure A.3: Number of images for all *Leaf Event Nodes* in the *VisE-Wiki* test dataset.

Table A.1: Results (numbers are multiplied by 100) on the manually annotated *VisE-Bing* test dataset using the probabilities  $\hat{\mathbf{y}}_L^o$  for classification in combination with different loss functions, weighting schemes (WS), and ontology redundancy removal (RR).

Model Notation	Loss	WS	RR	Accuracy			<i>JSC</i>	<i>CS</i>
				Top1	Top3	Top5		
$C$	$\mathcal{L}_c$			77.4	89.8	93.6	84.7	87.7
$O^{cel}$	$\mathcal{L}_o^{cel}$			66.9	83.2	88.6	80.2	84.5
$O_\omega^{cel}$	$\mathcal{L}_o^{cel}$	$\omega, \omega_L = 1$		67.7	83.1	88.9	80.3	84.5
$O_{6\omega}^{cel}$	$\mathcal{L}_o^{cel}$	$\omega, \omega_L = 6$		79.8	91.0	94.3	86.5	89.2
$O_{6\omega}^{cel} + RR$	$\mathcal{L}_o^{cel}$	$\omega, \omega_L = 6$	✓	<b>81.9</b>	<b>91.7</b>	<b>94.6</b>	<b>87.9</b>	<b>90.4</b>
$O_\gamma^{cel}$	$\mathcal{L}_o^{cel}$	$\gamma$		66.7	83.6	89.9	78.3	82.7
$O_\gamma^{cel} + RR$	$\mathcal{L}_o^{cel}$	$\gamma$	✓	73.2	87.2	91.8	82.4	86.0
$O^{cos}$	$\mathcal{L}_o^{cos}$			67.5	77.8	81.5	82.2	86.2
$O_\omega^{cos}$	$\mathcal{L}_o^{cos}$	$\omega, \omega_L = 1$		72.5	83.8	87.7	84.1	87.6
$O_{6\omega}^{cos}$	$\mathcal{L}_o^{cos}$	$\omega, \omega_L = 6$		80.4	90.7	93.6	86.4	89.0
$O_{6\omega}^{cos} + RR$	$\mathcal{L}_o^{cos}$	$\omega, \omega_L = 6$	✓	80.9	90.1	93.3	86.9	89.5
$O_\gamma^{cos}$	$\mathcal{L}_o^{cos}$	$\gamma$		81.3	90.1	93.6	87.3	89.7
$O_\gamma^{cos} + RR$	$\mathcal{L}_o^{cos}$	$\gamma$	✓	80.9	90.4	93.1	87.0	89.5
$CO_{6\omega}^{cel} + RR$	$\mathcal{L}_c + \mathcal{L}_o^{cel}$	$\omega, \omega_L = 6$	✓	81.6	<b>91.7</b>	94.5	87.5	90.0
$CO_\gamma^{cos}$	$\mathcal{L}_c + \mathcal{L}_o^{cos}$	$\gamma$		<b>81.9</b>	90.8	93.5	<b>87.9</b>	<b>90.4</b>

A Appendix

Table A.2: Results (numbers are multiplied by 100) on the manually annotated *VisE-Bing* test dataset using the probabilities  $\hat{\mathbf{y}}_L^{cos}$  for classification in combination with different loss functions, weighting schemes (WS), and ontology redundancy removal (RR).

Model Notation	Loss	WS	RR	Accuracy			<i>JSC</i>	<i>CS</i>
				Top1	Top3	Top5		
$C$	$\mathcal{L}_c$			77.4	89.8	93.6	84.7	87.7
$O^{cel}$	$\mathcal{L}_o$			68.0	77.5	81.0	82.1	86.4
$O_\omega^{cel}$	$\mathcal{L}_o^{cel}$	$\omega, \omega_L = 1$		68.0	78.2	82.4	81.6	85.8
$O_{6\omega}^{cel}$	$\mathcal{L}_o^{cel}$	$\omega, \omega_L = 6$		79.7	89.9	92.0	86.5	89.2
$O_{6\omega}^{cel} + RR$	$\mathcal{L}_o^{cel}$	$\omega, \omega_L = 6$	✓	<b>81.5</b>	90.8	92.9	<b>87.8</b>	<b>90.3</b>
$O_\gamma^{cel}$	$\mathcal{L}_o^{cel}$	$\gamma$		66.3	80.5	85.8	78.3	82.9
$O_\gamma^{cel} + RR$	$\mathcal{L}_o^{cel}$	$\gamma$	✓	72.7	84.7	88.2	82.3	86.0
$O^{cos}$	$\mathcal{L}_o^{cos}$			68.8	78.8	82.1	83.9	87.7
$O_\omega^{cos}$	$\mathcal{L}_o^{cos}$	$\omega, \omega_L = 1$		72.5	82.9	85.2	84.7	88.1
$O_{6\omega}^{cos}$	$\mathcal{L}_o^{cos}$	$\omega, \omega_L = 6$		80.1	89.5	92.0	86.3	89.0
$O_{6\omega}^{cos} + RR$	$\mathcal{L}_o^{cos}$	$\omega, \omega_L = 6$	✓	80.8	89.3	91.9	86.9	89.5
$O_\gamma^{cos}$	$\mathcal{L}_o^{cos}$	$\gamma$		79.8	87.5	89.6	86.6	89.4
$O_\gamma^{cos} + RR$	$\mathcal{L}_o^{cos}$	$\gamma$	✓	78.3	86.6	88.3	86.1	89.0
$CO_{6\omega}^{cel} + RR$	$\mathcal{L}_c + \mathcal{L}_o^{cel}$	$\omega, \omega_L = 6$	✓	81.4	<b>91.0</b>	<b>93.1</b>	87.4	89.9
$CO_\gamma^{cos}$	$\mathcal{L}_c + \mathcal{L}_o^{cos}$	$\gamma$		81.4	90.5	92.5	87.2	89.8

## A.2 Results for Geolocation Estimation

The tables on the following pages contain the results for the geolocation approaches presented in Section 3.2 for the *Im2GPS* [89] as well as *Im2GPS3k* [267] benchmark datasets. Furthermore, results for the benchmark subsets containing images of a specific scenery or environmental setting (*indoor*, *natural*, and *urban*) are reported.

A Appendix

Table A.3: Results on the *Im2GPS* test dataset of all images (ovr). Percentage is the fraction of images localized within the given radius using the *Great Circle Distance* (GCD).

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2,500 km
$base-vgg_c(L, m)$	7.6 %	22.8 %	35.0 %	50.6 %	66.7 %
$base-vgg(L, m)$	8.9 %	26.6 %	36.7 %	50.6 %	65.8 %
$vgg-ISNs(L, m, \mathbb{S}_3)$	11.8 %	32.1 %	44.3 %	56.5 %	71.7 %
$base(L, c)$	8.0 %	32.9 %	51.1 %	67.1 %	81.4 %
$base(L, m)$	13.5 %	36.3 %	50.6 %	64.1 %	79.7 %
$base(L, f)$	14.3 %	41.4 %	51.9 %	64.1 %	78.9 %
$base(M, c)$	8.9 %	33.3 %	47.3 %	63.7 %	78.1 %
$base(M, m)$	13.5 %	35.0 %	49.8 %	64.1 %	79.7 %
$base(M, f)$	14.3 %	40.1 %	49.8 %	64.6 %	79.3 %
$base(M, f^*)$	15.2 %	40.9 %	51.5 %	65.4 %	78.5 %
$ISNs(L, c, \mathbb{S}_3)$	8.9 %	30.4 %	48.5 %	66.2 %	<b>83.5 %</b>
$ISNs(L, m, \mathbb{S}_3)$	13.1 %	32.5 %	46.8 %	63.7 %	79.7 %
$ISNs(L, f, \mathbb{S}_3)$	15.2 %	40.9 %	50.6 %	62.0 %	77.6 %
$ISNs(M, c, \mathbb{S}_3)$	9.7 %	33.8 %	48.5 %	65.0 %	80.2 %
$ISNs(M, m, \mathbb{S}_3)$	15.6 %	38.8 %	<b>52.3 %</b>	<b>67.9 %</b>	82.3 %
$ISNs(M, f, \mathbb{S}_3)$	16.5 %	42.2 %	51.9 %	66.2 %	81.0 %
$ISNs(M, f^*, \mathbb{S}_3)$	<b>16.9 %</b>	<b>43.0 %</b>	51.9 %	66.7 %	80.2 %
$MTN(L, f, \mathbb{S}_3)$	13.9 %	38.4 %	48.9 %	62.9 %	79.3 %
$MTN(L, f, \mathbb{S}_{16})$	12.7 %	37.1 %	44.7 %	59.1 %	75.5 %
$MTN(L, f, \mathbb{S}_{365})$	13.9 %	37.1 %	46.0 %	60.8 %	74.3 %
$MTN(M, c, \mathbb{S}_3)$	8.4 %	32.1 %	48.1 %	62.4 %	78.1 %
$MTN(M, m, \mathbb{S}_3)$	13.1 %	34.2 %	44.3 %	63.3 %	78.9 %
$MTN(M, f, \mathbb{S}_3)$	14.3 %	38.4 %	47.3 %	63.7 %	76.4 %
$MTN(M, f^*, \mathbb{S}_3)$	13.5 %	37.6 %	46.8 %	63.3 %	78.1 %
$MTN(M, c, \mathbb{S}_{16})$	8.9 %	33.3 %	48.1 %	62.4 %	75.1 %
$MTN(M, m, \mathbb{S}_{16})$	12.7 %	35.4 %	46.8 %	61.2 %	74.7 %
$MTN(M, f, \mathbb{S}_{16})$	13.5 %	36.7 %	46.8 %	63.3 %	75.9 %
$MTN(M, f^*, \mathbb{S}_{16})$	13.5 %	38.4 %	45.1 %	59.9 %	74.7 %
$MTN(M, c, \mathbb{S}_{365})$	8.0 %	29.5 %	43.5 %	59.9 %	75.5 %
$MTN(M, m, \mathbb{S}_{365})$	13.5 %	34.2 %	44.7 %	59.9 %	77.2 %
$MTN(M, f, \mathbb{S}_{365})$	13.1 %	35.4 %	43.5 %	59.9 %	75.5 %
$MTN(M, f^*, \mathbb{S}_{365})$	13.9 %	36.7 %	44.3 %	61.6 %	77.6 %



Table A.4: Results on the *Im2GPS* test dataset of all images classified as *indoor*. Percentage is the fraction of images localized within the given radius using the *Great Circle Distance* (GCD).

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2,500 km
<i>base-vgg<sub>c</sub></i> ( $L, m$ )	5.3 %	10.5 %	15.8 %	31.6 %	47.4 %
<i>base-vgg</i> ( $L, m$ )	5.3 %	10.5 %	15.8 %	31.6 %	47.4 %
<i>vgg-ISNs</i> ( $L, m, \mathbb{S}_3$ )	5.3 %	10.5 %	10.5 %	26.3 %	36.8 %
<i>base</i> ( $L, c$ )	0.0 %	26.3 %	31.6 %	42.1 %	52.6 %
<i>base</i> ( $L, m$ )	5.3 %	21.1 %	21.1 %	31.6 %	52.6 %
<i>base</i> ( $L, f$ )	10.5 %	31.6 %	31.6 %	42.1 %	57.9 %
<i>base</i> ( $M, c$ )	0.0 %	31.6 %	36.8 %	52.6 %	68.4 %
<i>base</i> ( $M, m$ )	10.5 %	<b>36.8 %</b>	<b>42.1 %</b>	<b>57.9 %</b>	<b>78.9 %</b>
<i>base</i> ( $M, f$ )	10.5 %	<b>36.8 %</b>	<b>42.1 %</b>	<b>57.9 %</b>	<b>78.9 %</b>
<i>base</i> ( $M, f^*$ )	10.5 %	<b>36.8 %</b>	<b>42.1 %</b>	<b>57.9 %</b>	<b>78.9 %</b>
ISNs ( $L, c, \mathbb{S}_3$ )	0.0 %	26.3 %	31.6 %	47.4 %	63.2 %
ISNs ( $L, m, \mathbb{S}_3$ )	5.3 %	21.1 %	21.1 %	26.3 %	42.1 %
ISNs ( $L, f, \mathbb{S}_3$ )	10.5 %	31.6 %	31.6 %	36.8 %	47.4 %
ISNs ( $M, c, \mathbb{S}_3$ )	5.3 %	31.6 %	36.8 %	42.1 %	52.6 %
ISNs ( $M, m, \mathbb{S}_3$ )	<b>15.8 %</b>	<b>36.8 %</b>	<b>42.1 %</b>	47.4 %	68.4 %
ISNs ( $M, f, \mathbb{S}_3$ )	<b>15.8 %</b>	26.3 %	31.6 %	42.1 %	57.9 %
ISNs ( $M, f^*, \mathbb{S}_3$ )	<b>15.8 %</b>	31.6 %	36.8 %	42.1 %	57.9 %
MTN ( $L, f, \mathbb{S}_3$ )	10.5 %	31.6 %	31.6 %	52.6 %	73.7 %
MTN ( $L, f, \mathbb{S}_{16}$ )	5.3 %	15.8 %	26.3 %	26.3 %	47.4 %
MTN ( $L, f, \mathbb{S}_{365}$ )	10.5 %	21.1 %	26.3 %	42.1 %	47.4 %
MTN ( $M, c, \mathbb{S}_3$ )	0.0 %	26.3 %	36.8 %	52.6 %	57.9 %
MTN ( $M, m, \mathbb{S}_3$ )	5.3 %	21.1 %	21.1 %	42.1 %	63.2 %
MTN ( $M, f, \mathbb{S}_3$ )	0.0 %	21.1 %	21.1 %	42.1 %	63.2 %
MTN ( $M, f^*, \mathbb{S}_3$ )	0.0 %	21.1 %	21.1 %	42.1 %	57.9 %
MTN ( $M, c, \mathbb{S}_{16}$ )	5.3 %	31.6 %	31.6 %	47.4 %	63.2 %
MTN ( $M, m, \mathbb{S}_{16}$ )	5.3 %	26.3 %	26.3 %	47.4 %	63.2 %
MTN ( $M, f, \mathbb{S}_{16}$ )	5.3 %	21.1 %	26.3 %	47.4 %	63.2 %
MTN ( $M, f^*, \mathbb{S}_{16}$ )	5.3 %	26.3 %	26.3 %	52.6 %	68.4 %
MTN ( $M, c, \mathbb{S}_{365}$ )	5.3 %	26.3 %	31.6 %	47.4 %	57.9 %
MTN ( $M, m, \mathbb{S}_{365}$ )	5.3 %	15.8 %	21.1 %	36.8 %	57.9 %
MTN ( $M, f, \mathbb{S}_{365}$ )	5.3 %	21.1 %	26.3 %	42.1 %	57.9 %
MTN ( $M, f^*, \mathbb{S}_{365}$ )	0.0 %	21.1 %	26.3 %	47.4 %	63.2 %

A Appendix

Table A.5: Results on the *Im2GPS* test dataset of all images classified as *natural*. Percentage is the fraction of images localized within the given radius using the *Great Circle Distance* (GCD).

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2,500 km
$base-vgg_c(L, m)$	1.3 %	13.8 %	36.3 %	48.8 %	62.5 %
$base-vgg(L, m)$	2.5 %	22.5 %	38.8 %	51.3 %	60.0 %
$vgg-ISNs(L, m, \mathbb{S}_3)$	3.8 %	27.5 %	42.5 %	48.8 %	61.3 %
$base(L, c)$	1.3 %	18.8 %	46.3 %	58.8 %	<b>76.3 %</b>
$base(L, m)$	3.8 %	30.0 %	50.0 %	60.0 %	71.3 %
$base(L, f)$	3.8 %	<b>37.5 %</b>	<b>52.5 %</b>	<b>61.3 %</b>	73.8 %
$base(M, c)$	1.3 %	22.5 %	46.3 %	57.5 %	68.8 %
$base(M, m)$	3.8 %	26.3 %	48.8 %	57.5 %	68.8 %
$base(M, f)$	3.8 %	35.0 %	48.8 %	56.3 %	65.0 %
$base(M, f^*)$	3.8 %	33.8 %	48.8 %	57.5 %	66.3 %
$ISNs(L, c, \mathbb{S}_3)$	1.3 %	16.3 %	41.3 %	56.3 %	75.0 %
$ISNs(L, m, \mathbb{S}_3)$	3.8 %	26.3 %	45.0 %	57.5 %	68.8 %
$ISNs(L, f, \mathbb{S}_3)$	3.8 %	<b>37.5 %</b>	46.3 %	56.3 %	<b>76.3 %</b>
$ISNs(M, c, \mathbb{S}_3)$	1.3 %	22.5 %	45.0 %	58.8 %	72.5 %
$ISNs(M, m, \mathbb{S}_3)$	3.8 %	26.3 %	46.3 %	57.5 %	73.8 %
$ISNs(M, f, \mathbb{S}_3)$	2.5 %	36.3 %	48.8 %	56.3 %	71.3 %
$ISNs(M, f^*, \mathbb{S}_3)$	2.5 %	36.3 %	46.3 %	56.3 %	72.5 %
$MTN(L, f, \mathbb{S}_3)$	<b>5.0 %</b>	<b>37.5 %</b>	48.8 %	55.0 %	71.3 %
$MTN(L, f, \mathbb{S}_{16})$	2.5 %	35.0 %	43.8 %	53.8 %	67.5 %
$MTN(L, f, \mathbb{S}_{365})$	2.5 %	<b>37.5 %</b>	50.0 %	60.0 %	71.3 %
$MTN(M, c, \mathbb{S}_3)$	1.3 %	18.8 %	38.8 %	52.5 %	68.8 %
$MTN(M, m, \mathbb{S}_3)$	2.5 %	28.8 %	41.3 %	56.3 %	70.0 %
$MTN(M, f, \mathbb{S}_3)$	3.8 %	35.0 %	45.0 %	55.0 %	68.8 %
$MTN(M, f^*, \mathbb{S}_3)$	2.5 %	32.5 %	41.3 %	53.8 %	70.0 %
$MTN(M, c, \mathbb{S}_{16})$	1.3 %	20.0 %	43.8 %	55.0 %	67.5 %
$MTN(M, m, \mathbb{S}_{16})$	3.8 %	32.5 %	45.0 %	53.8 %	67.5 %
$MTN(M, f, \mathbb{S}_{16})$	3.8 %	33.8 %	47.5 %	<b>61.3 %</b>	70.0 %
$MTN(M, f^*, \mathbb{S}_{16})$	3.8 %	35.0 %	42.5 %	52.5 %	66.3 %
$MTN(M, c, \mathbb{S}_{365})$	1.3 %	18.8 %	38.8 %	53.8 %	67.5 %
$MTN(M, m, \mathbb{S}_{365})$	<b>5.0 %</b>	28.8 %	41.3 %	51.3 %	72.5 %
$MTN(M, f, \mathbb{S}_{365})$	3.8 %	33.8 %	42.5 %	52.5 %	68.8 %
$MTN(M, f^*, \mathbb{S}_{365})$	3.8 %	35.0 %	41.3 %	55.0 %	70.0 %

Table A.6: Results on the *Im2GPS* test dataset of all images classified as *urban*. Percentage is the fraction of images localized within the given radius using the *Great Circle Distance* (GCD).

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2,500 km
<i>base-vgg<sub>c</sub></i> ( $L, m$ )	11.6 %	29.7 %	37.0 %	54.3 %	71.7 %
<i>base-vgg</i> ( $L, m$ )	13.0 %	31.2 %	38.4 %	52.9 %	71.7 %
<i>vgg-ISNs</i> ( $L, m, \mathbb{S}_3$ )	17.4 %	37.7 %	50.0 %	65.2 %	82.6 %
<i>base</i> ( $L, c$ )	13.0 %	42.0 %	56.5 %	75.4 %	88.4 %
<i>base</i> ( $L, m$ )	20.3 %	42.0 %	55.1 %	71.0 %	88.4 %
<i>base</i> ( $L, f$ )	21.0 %	44.9 %	54.3 %	68.8 %	84.8 %
<i>base</i> ( $M, c$ )	14.5 %	39.9 %	49.3 %	68.8 %	84.8 %
<i>base</i> ( $M, m$ )	19.6 %	39.9 %	51.4 %	68.8 %	86.2 %
<i>base</i> ( $M, f$ )	21.0 %	43.5 %	51.4 %	70.3 %	87.7 %
<i>base</i> ( $M, f^*$ )	22.5 %	45.7 %	54.3 %	71.0 %	85.5 %
ISNs ( $L, c, \mathbb{S}_3$ )	14.5 %	39.1 %	55.1 %	74.6 %	<b>91.3 %</b>
ISNs ( $L, m, \mathbb{S}_3$ )	19.6 %	37.7 %	51.4 %	72.5 %	<b>91.3 %</b>
ISNs ( $L, f, \mathbb{S}_3$ )	22.5 %	44.2 %	55.8 %	68.8 %	82.6 %
ISNs ( $M, c, \mathbb{S}_3$ )	15.2 %	40.6 %	52.2 %	71.7 %	88.4 %
ISNs ( $M, m, \mathbb{S}_3$ )	22.5 %	46.4 %	<b>57.2 %</b>	<b>76.8 %</b>	89.1 %
ISNs ( $M, f, \mathbb{S}_3$ )	24.6 %	47.8 %	56.5 %	75.4 %	89.9 %
ISNs ( $M, f^*, \mathbb{S}_3$ )	<b>25.4 %</b>	<b>48.6 %</b>	<b>57.2 %</b>	76.1 %	87.7 %
MTN ( $L, f, \mathbb{S}_3$ )	19.6 %	39.9 %	51.4 %	68.8 %	84.8 %
MTN ( $L, f, \mathbb{S}_{16}$ )	19.6 %	41.3 %	47.8 %	66.7 %	84.1 %
MTN ( $L, f, \mathbb{S}_{365}$ )	21.0 %	39.1 %	46.4 %	63.8 %	79.7 %
MTN ( $M, c, \mathbb{S}_3$ )	13.8 %	40.6 %	55.1 %	69.6 %	86.2 %
MTN ( $M, m, \mathbb{S}_3$ )	20.3 %	39.1 %	49.3 %	70.3 %	86.2 %
MTN ( $M, f, \mathbb{S}_3$ )	22.5 %	42.8 %	52.2 %	71.7 %	82.6 %
MTN ( $M, f^*, \mathbb{S}_3$ )	21.7 %	42.8 %	53.6 %	71.7 %	85.5 %
MTN ( $M, c, \mathbb{S}_{16}$ )	13.8 %	41.3 %	52.9 %	68.8 %	81.2 %
MTN ( $M, m, \mathbb{S}_{16}$ )	18.8 %	38.4 %	50.7 %	67.4 %	80.4 %
MTN ( $M, f, \mathbb{S}_{16}$ )	20.3 %	40.6 %	49.3 %	66.7 %	81.2 %
MTN ( $M, f^*, \mathbb{S}_{16}$ )	20.3 %	42.0 %	49.3 %	65.2 %	80.4 %
MTN ( $M, c, \mathbb{S}_{365}$ )	12.3 %	36.2 %	47.8 %	65.2 %	82.6 %
MTN ( $M, m, \mathbb{S}_{365}$ )	19.6 %	39.9 %	50.0 %	68.1 %	82.6 %
MTN ( $M, f, \mathbb{S}_{365}$ )	19.6 %	38.4 %	46.4 %	66.7 %	81.9 %
MTN ( $M, f^*, \mathbb{S}_{365}$ )	21.7 %	39.9 %	48.6 %	67.4 %	84.1 %

A Appendix

Table A.7: Results on the *Im2GPS3k* test dataset of all images (ovr). Percentage is the fraction of images localized within the given radius using the *Great Circle Distance* (GCD).

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2,500 km
<i>base-vgg<sub>c</sub></i> ( $L, m$ )	4.2 %	14.6 %	22.2 %	34.4 %	54.2 %
<i>base-vgg</i> ( $L, m$ )	4.8 %	16.5 %	22.6 %	34.5 %	54.4 %
<i>vgg-ISNs</i> ( $L, m, \mathbb{S}_3$ )	5.8 %	19.3 %	27.1 %	40.5 %	59.0 %
<i>base</i> ( $L, c$ )	6.1 %	23.3 %	34.0 %	48.4 %	64.9 %
<i>base</i> ( $L, m$ )	8.3 %	24.9 %	34.0 %	48.8 %	65.8 %
<i>base</i> ( $L, f$ )	9.7 %	25.8 %	33.8 %	46.7 %	63.8 %
<i>base</i> ( $M, c$ )	6.2 %	23.1 %	34.3 %	48.6 %	65.9 %
<i>base</i> ( $M, m$ )	8.2 %	25.5 %	35.1 %	48.7 %	65.2 %
<i>base</i> ( $M, f$ )	9.6 %	26.3 %	34.8 %	48.1 %	65.3 %
<i>base</i> ( $M, f^*$ )	9.7 %	27.0 %	35.6 %	49.1 %	66.0 %
ISNs ( $L, c, \mathbb{S}_3$ )	6.2 %	23.5 %	34.5 %	48.6 %	65.0 %
ISNs ( $L, m, \mathbb{S}_3$ )	8.0 %	24.8 %	34.5 %	48.7 %	65.2 %
ISNs ( $L, f, \mathbb{S}_3$ )	9.7 %	26.1 %	34.4 %	47.6 %	64.0 %
ISNs ( $M, c, \mathbb{S}_3$ )	6.4 %	23.7 %	35.2 %	<b>50.1 %</b>	66.5 %
ISNs ( $M, m, \mathbb{S}_3$ )	8.8 %	26.4 %	36.5 %	<b>50.1 %</b>	<b>66.6 %</b>
ISNs ( $M, f, \mathbb{S}_3$ )	10.1 %	27.2 %	36.2 %	49.3 %	65.6 %
ISNs ( $M, f^*, \mathbb{S}_3$ )	<b>10.5 %</b>	<b>28.0 %</b>	<b>36.6 %</b>	49.7 %	66.0 %
MTN ( $L, f, \mathbb{S}_3$ )	9.4 %	25.3 %	32.8 %	45.8 %	62.9 %
MTN ( $L, f, \mathbb{S}_{16}$ )	9.1 %	24.9 %	32.8 %	45.5 %	62.7 %
MTN ( $L, f, \mathbb{S}_{365}$ )	8.5 %	23.2 %	30.9 %	44.3 %	61.3 %
MTN ( $M, c, \mathbb{S}_3$ )	6.0 %	22.9 %	33.4 %	49.0 %	65.4 %
MTN ( $M, m, \mathbb{S}_3$ )	8.0 %	24.6 %	33.5 %	47.1 %	64.2 %
MTN ( $M, f, \mathbb{S}_3$ )	9.0 %	25.1 %	33.7 %	46.5 %	63.6 %
MTN ( $M, f^*, \mathbb{S}_3$ )	9.4 %	26.1 %	34.5 %	47.7 %	64.6 %
MTN ( $M, c, \mathbb{S}_{16}$ )	5.6 %	21.8 %	32.3 %	46.2 %	64.1 %
MTN ( $M, m, \mathbb{S}_{16}$ )	7.7 %	23.3 %	31.8 %	45.0 %	62.7 %
MTN ( $M, f, \mathbb{S}_{16}$ )	9.2 %	24.3 %	32.2 %	45.3 %	63.5 %
MTN ( $M, f^*, \mathbb{S}_{16}$ )	9.3 %	25.1 %	33.3 %	45.9 %	63.5 %
MTN ( $M, c, \mathbb{S}_{365}$ )	5.6 %	20.3 %	29.9 %	44.1 %	61.9 %
MTN ( $M, m, \mathbb{S}_{365}$ )	7.3 %	22.1 %	30.2 %	44.2 %	61.7 %
MTN ( $M, f, \mathbb{S}_{365}$ )	8.6 %	23.1 %	30.2 %	43.5 %	61.7 %
MTN ( $M, f^*, \mathbb{S}_{365}$ )	8.9 %	23.6 %	31.0 %	44.4 %	62.0 %

Table A.8: Results on the *Im2GPS3k* test dataset of all images classified as *indoor*. Percentage is the fraction of images localized within the given radius using the *Great Circle Distance* (GCD).

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2,500 km
<i>base-vgg<sub>c</sub></i> ( $L, m$ )	2.4 %	6.6 %	7.7 %	15.4 %	39.6 %
<i>base-vgg</i> ( $L, m$ )	2.8 %	7.2 %	7.7 %	15.6 %	39.8 %
<i>vgg-ISNs</i> ( $L, m, \mathbb{S}_3$ )	4.8 %	9.9 %	11.2 %	20.6 %	41.5 %
<i>base</i> ( $L, c$ )	4.6 %	13.0 %	17.4 %	28.3 %	50.1 %
<i>base</i> ( $L, m$ )	8.4 %	14.1 %	16.5 %	29.5 %	50.5 %
<i>base</i> ( $L, f$ )	7.7 %	13.4 %	16.1 %	23.7 %	48.1 %
<i>base</i> ( $M, c$ )	5.0 %	13.0 %	16.0 %	25.0 %	50.6 %
<i>base</i> ( $M, m$ )	7.3 %	13.2 %	16.3 %	26.1 %	49.2 %
<i>base</i> ( $M, f$ )	8.1 %	13.9 %	16.0 %	25.1 %	48.8 %
<i>base</i> ( $M, f^*$ )	7.9 %	14.3 %	16.9 %	26.2 %	50.3 %
ISNs ( $L, c, \mathbb{S}_3$ )	4.4 %	13.0 %	16.9 %	27.3 %	49.7 %
ISNs ( $L, m, \mathbb{S}_3$ )	7.3 %	12.5 %	15.2 %	25.7 %	47.7 %
ISNs ( $L, f, \mathbb{S}_3$ )	8.1 %	14.3 %	16.5 %	25.9 %	48.1 %
ISNs ( $M, c, \mathbb{S}_3$ )	5.3 %	14.5 %	17.6 %	28.4 %	49.4 %
ISNs ( $M, m, \mathbb{S}_3$ )	8.6 %	15.0 %	<b>17.8 %</b>	29.4 %	<b>51.7 %</b>
ISNs ( $M, f, \mathbb{S}_3$ )	<b>9.2 %</b>	15.4 %	<b>17.8 %</b>	<b>29.9 %</b>	48.4 %
ISNs ( $M, f^*, \mathbb{S}_3$ )	<b>9.2 %</b>	<b>15.8 %</b>	17.2 %	28.3 %	49.5 %
MTN ( $L, f, \mathbb{S}_3$ )	8.6 %	13.4 %	15.0 %	24.8 %	47.2 %
MTN ( $L, f, \mathbb{S}_{16}$ )	7.2 %	13.0 %	15.4 %	26.6 %	46.8 %
MTN ( $L, f, \mathbb{S}_{365}$ )	6.6 %	12.5 %	14.9 %	27.3 %	48.8 %
MTN ( $M, c, \mathbb{S}_3$ )	4.6 %	12.1 %	15.0 %	27.7 %	49.4 %
MTN ( $M, m, \mathbb{S}_3$ )	7.3 %	13.2 %	15.4 %	27.5 %	47.7 %
MTN ( $M, f, \mathbb{S}_3$ )	7.3 %	13.4 %	15.6 %	26.2 %	44.8 %
MTN ( $M, f^*, \mathbb{S}_3$ )	7.5 %	13.4 %	15.2 %	26.4 %	46.8 %
MTN ( $M, c, \mathbb{S}_{16}$ )	4.2 %	12.3 %	15.2 %	26.8 %	49.9 %
MTN ( $M, m, \mathbb{S}_{16}$ )	7.7 %	13.0 %	15.2 %	26.4 %	49.4 %
MTN ( $M, f, \mathbb{S}_{16}$ )	7.5 %	12.8 %	15.8 %	26.1 %	50.3 %
MTN ( $M, f^*, \mathbb{S}_{16}$ )	8.1 %	14.1 %	16.9 %	28.4 %	52.3 %
MTN ( $M, c, \mathbb{S}_{365}$ )	4.0 %	11.4 %	14.9 %	26.1 %	47.7 %
MTN ( $M, m, \mathbb{S}_{365}$ )	5.7 %	12.1 %	15.2 %	27.2 %	47.9 %
MTN ( $M, f, \mathbb{S}_{365}$ )	6.1 %	12.1 %	14.5 %	26.4 %	49.4 %
MTN ( $M, f^*, \mathbb{S}_{365}$ )	6.2 %	12.7 %	15.0 %	26.1 %	47.9 %

A Appendix

Table A.9: Results on the *Im2GPS3k* test dataset of all images classified as *natural*. Percentage is the fraction of images localized within the given radius using the *Great Circle Distance* (GCD).

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2,500 km
<i>base-vgg<sub>c</sub></i> ( $L, m$ )	1.1 %	9.0 %	21.1 %	35.0 %	51.4 %
<i>base-vgg</i> ( $L, m$ )	1.3 %	11.2 %	22.2 %	35.4 %	51.6 %
<i>vgg-ISNs</i> ( $L, m, \mathbb{S}_3$ )	1.2 %	11.1 %	24.7 %	39.4 %	56.6 %
<i>base</i> ( $L, c$ )	1.3 %	12.1 %	27.5 %	43.4 %	60.1 %
<i>base</i> ( $L, m$ )	2.0 %	15.0 %	28.6 %	44.3 %	60.9 %
<i>base</i> ( $L, f$ )	3.3 %	17.2 %	28.0 %	42.7 %	60.4 %
<i>base</i> ( $M, c$ )	1.5 %	12.8 %	30.1 %	46.6 %	63.1 %
<i>base</i> ( $M, m$ )	2.1 %	15.7 %	31.7 %	46.6 %	61.3 %
<i>base</i> ( $M, f$ )	3.4 %	17.2 %	30.8 %	44.9 %	60.6 %
<i>base</i> ( $M, f^*$ )	3.3 %	17.5 %	32.0 %	46.3 %	61.7 %
ISNs ( $L, c, \mathbb{S}_3$ )	1.1 %	12.2 %	28.6 %	45.4 %	61.9 %
ISNs ( $L, m, \mathbb{S}_3$ )	2.1 %	15.1 %	30.9 %	46.7 %	62.0 %
ISNs ( $L, f, \mathbb{S}_3$ )	2.8 %	17.2 %	30.2 %	44.3 %	60.9 %
ISNs ( $M, c, \mathbb{S}_3$ )	1.9 %	12.5 %	29.9 %	<b>47.2 %</b>	<b>64.4 %</b>
ISNs ( $M, m, \mathbb{S}_3$ )	3.0 %	15.4 %	31.6 %	46.5 %	62.8 %
ISNs ( $M, f, \mathbb{S}_3$ )	3.2 %	17.0 %	31.8 %	46.6 %	63.1 %
ISNs ( $M, f^*, \mathbb{S}_3$ )	<b>3.9 %</b>	<b>18.1 %</b>	<b>32.8 %</b>	47.0 %	62.8 %
MTN ( $L, f, \mathbb{S}_3$ )	2.7 %	16.2 %	27.8 %	42.5 %	59.2 %
MTN ( $L, f, \mathbb{S}_{16}$ )	3.0 %	16.8 %	30.4 %	43.3 %	59.6 %
MTN ( $L, f, \mathbb{S}_{365}$ )	2.5 %	14.8 %	26.5 %	41.2 %	55.7 %
MTN ( $M, c, \mathbb{S}_3$ )	1.1 %	12.2 %	27.0 %	45.6 %	62.2 %
MTN ( $M, m, \mathbb{S}_3$ )	2.4 %	14.8 %	28.3 %	42.4 %	59.8 %
MTN ( $M, f, \mathbb{S}_3$ )	3.4 %	16.3 %	28.8 %	41.3 %	59.1 %
MTN ( $M, f^*, \mathbb{S}_3$ )	3.4 %	17.0 %	29.6 %	43.4 %	61.1 %
MTN ( $M, c, \mathbb{S}_{16}$ )	1.3 %	11.4 %	27.1 %	41.8 %	59.1 %
MTN ( $M, m, \mathbb{S}_{16}$ )	2.1 %	13.4 %	27.0 %	40.5 %	57.9 %
MTN ( $M, f, \mathbb{S}_{16}$ )	2.8 %	15.7 %	27.5 %	41.1 %	57.9 %
MTN ( $M, f^*, \mathbb{S}_{16}$ )	2.8 %	15.5 %	28.0 %	40.1 %	56.2 %
MTN ( $M, c, \mathbb{S}_{365}$ )	1.3 %	10.2 %	24.5 %	40.0 %	58.3 %
MTN ( $M, m, \mathbb{S}_{365}$ )	2.0 %	13.1 %	26.3 %	40.5 %	57.6 %
MTN ( $M, f, \mathbb{S}_{365}$ )	2.1 %	14.0 %	25.8 %	39.1 %	56.4 %
MTN ( $M, f^*, \mathbb{S}_{365}$ )	3.0 %	14.8 %	26.5 %	39.8 %	57.2 %

Table A.10: Results on the *Im2GPS3k* test dataset of all images classified as *urban*. Percentage is the fraction of images localized within the given radius using the *Great Circle Distance* (GCD).

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2,500 km
<i>base-vgg<sub>c</sub></i> ( $L, m$ )	6.5 %	20.3 %	27.8 %	40.5 %	60.7 %
<i>base-vgg</i> ( $L, m$ )	7.3 %	22.5 %	27.9 %	40.4 %	60.7 %
<i>vgg-ISNs</i> ( $L, m, \mathbb{S}_3$ )	8.5 %	26.8 %	33.8 %	47.9 %	66.2 %
<i>base</i> ( $L, c$ )	9.2 %	32.6 %	43.0 %	57.8 %	72.5 %
<i>base</i> ( $L, m$ )	11.6 %	33.8 %	42.8 %	57.7 %	73.5 %
<i>base</i> ( $L, f$ )	13.8 %	34.5 %	42.8 %	56.6 %	70.9 %
<i>base</i> ( $M, c$ )	9.1 %	32.0 %	42.8 %	57.7 %	72.6 %
<i>base</i> ( $M, m$ )	11.6 %	34.8 %	43.2 %	57.4 %	72.7 %
<i>base</i> ( $M, f$ )	13.4 %	35.2 %	43.3 %	57.6 %	73.3 %
<i>base</i> ( $M, f^*$ )	13.8 %	36.3 %	43.9 %	58.4 %	73.6 %
ISNs ( $L, c, \mathbb{S}_3$ )	9.6 %	33.0 %	43.6 %	57.6 %	71.8 %
ISNs ( $L, m, \mathbb{S}_3$ )	11.3 %	34.0 %	43.0 %	57.6 %	72.7 %
ISNs ( $L, f, \mathbb{S}_3$ )	13.9 %	34.8 %	42.8 %	56.8 %	71.1 %
ISNs ( $M, c, \mathbb{S}_3$ )	9.1 %	32.6 %	44.0 %	58.9 %	73.5 %
ISNs ( $M, m, \mathbb{S}_3$ )	11.9 %	36.1 %	<b>45.5 %</b>	<b>59.0 %</b>	<b>73.7 %</b>
ISNs ( $M, f, \mathbb{S}_3$ )	14.1 %	36.6 %	44.8 %	57.3 %	72.7 %
ISNs ( $M, f^*, \mathbb{S}_3$ )	<b>14.5 %</b>	<b>37.3 %</b>	45.2 %	58.4 %	73.2 %
MTN ( $L, f, \mathbb{S}_3$ )	13.3 %	34.1 %	41.4 %	54.8 %	70.1 %
MTN ( $L, f, \mathbb{S}_{16}$ )	12.9 %	33.1 %	40.0 %	53.1 %	69.8 %
MTN ( $L, f, \mathbb{S}_{365}$ )	12.3 %	31.3 %	38.6 %	51.7 %	68.5 %
MTN ( $M, c, \mathbb{S}_3$ )	9.1 %	32.2 %	43.0 %	58.1 %	72.5 %
MTN ( $M, m, \mathbb{S}_3$ )	11.1 %	33.5 %	42.3 %	56.2 %	72.2 %
MTN ( $M, f, \mathbb{S}_3$ )	12.6 %	33.6 %	42.4 %	56.2 %	72.4 %
MTN ( $M, f^*, \mathbb{S}_3$ )	13.2 %	35.2 %	43.6 %	57.2 %	72.5 %
MTN ( $M, c, \mathbb{S}_{16}$ )	8.3 %	30.4 %	40.8 %	55.1 %	71.6 %
MTN ( $M, m, \mathbb{S}_{16}$ )	10.6 %	32.0 %	40.0 %	53.6 %	69.8 %
MTN ( $M, f, \mathbb{S}_{16}$ )	13.2 %	32.7 %	40.3 %	54.0 %	71.0 %
MTN ( $M, f^*, \mathbb{S}_{16}$ )	13.1 %	33.9 %	41.6 %	54.9 %	71.1 %
MTN ( $M, c, \mathbb{S}_{365}$ )	8.3 %	28.7 %	37.9 %	52.3 %	68.6 %
MTN ( $M, m, \mathbb{S}_{365}$ )	10.7 %	30.2 %	37.3 %	51.9 %	68.6 %
MTN ( $M, f, \mathbb{S}_{365}$ )	12.9 %	31.7 %	37.9 %	51.6 %	68.6 %
MTN ( $M, f^*, \mathbb{S}_{365}$ )	12.9 %	31.9 %	38.7 %	53.1 %	69.4 %

### A.3 Results on other Subsets of TamperedNews and News400

Results for the Top-25% documents according to the cross-modal similarity values obtained for the original entities as well as for all documents of *TamperedNews* are presented in Table A.11 and Table A.12. As discussed in Section 4.5.1 (paragraph "Test Document Selection for TamperedNews") of this thesis, we have used subsets of *TamperedNews* in order to counteract the influence of original documents that do not contain a single cross-modal entity relation. Thus, results for all documents are worse compared to the proposed subsets since many documents without cross-modal relations are considered. On the other hand, results for *TamperedNews (Top-25%)* and *TamperedNews (Top-50%)* allow for similar conclusions. However, in particular, retrieval of original documents is noticeable better when using (smaller) subsets. As discussed in Section 4.5.3.2, this is mainly caused by the fact that some original entities in the documents depicted in both image and text can be either unspecific (e.g., mentioning of a country) or the retrieved images for visual verification do not fit the document's image content. When using these subsets, we have bypassed this problem. We have verified the same behavior for *News400* when experimenting on a subset with the Top-50% documents according to the cross-modal similarity values of original entities (Top-25% subset is omitted since it contains too few documents). The respective results are shown in Table A.13.



A.3 Results on other Subsets of *TamperedNews* and *News400*

Table A.11: Results for document verification (DV) and collection retrieval for the *TamperedNews* (Top-25%) dataset for different entity test sets (notations according to Section 4.4.1).

Test set	DV		Collection Retrieval					
	VA	AUC	AP-original [%]			AP-manipulated [%]		
			@25%	@50%	@100%	@25%	@50%	@100%
<b>Persons</b> (8,424 documents)								
Random	0.98	0.98	96.74	96.63	96.41	100.0	100.0	98.76
PsC	0.98	0.98	96.49	96.40	96.05	100.0	100.0	98.61
PsG	0.98	0.98	96.45	96.27	96.15	100.0	100.0	98.71
PsCG	0.98	0.98	95.21	95.71	95.79	100.0	100.0	98.65
<b>Locations - Outdoor</b> (7,057 documents)								
Random	0.94	0.93	96.34	94.06	90.66	100.0	100.0	95.38
GCD $\frac{2500}{750}$	0.93	0.90	91.62	89.27	85.98	100.0	100.0	93.21
GCD $\frac{750}{200}$	0.88	0.85	88.62	84.66	80.12	100.0	100.0	89.23
GCD $\frac{200}{25}$	0.85	0.82	86.59	82.19	77.36	100.0	100.0	87.11
<b>Locations - Indoor</b> (9,565 documents)								
Random	0.83	0.81	75.08	73.58	73.09	100.0	100.0	87.15
GCD $\frac{2500}{750}$	0.80	0.78	68.25	67.94	68.65	100.0	100.0	84.64
GCD $\frac{750}{200}$	0.82	0.79	72.63	71.50	71.00	100.0	100.0	85.61
GCD $\frac{200}{25}$	0.76	0.74	57.67	60.50	63.75	100.0	100.0	82.46
<b>Events</b> (3,867 documents)								
Random	0.97	0.96	92.75	92.85	92.36	100.0	100.0	97.20
EsP	0.78	0.75	74.81	71.28	68.69	100.0	100.0	82.02
<b>Context</b> (18,108 documents)								
Random	0.92	0.92	94.24	92.41	88.83	100.0	100.0	94.48
Top-25%	0.90	0.89	89.48	87.55	84.27	100.0	100.0	92.38
Top-10%	0.87	0.85	84.32	82.08	79.35	100.0	100.0	89.88
Top-5%	0.84	0.82	80.36	78.09	75.66	100.0	100.0	87.68

Table A.12: Results for document verification (DV) and collection retrieval for all documents of the *TamperedNews* dataset for different entity test sets (notations according to Section 4.4.1).

Test set	DV		Collection Retrieval					
	VA	AUC	AP-original [%]			AP-manipulated [%]		
			@25%	@50%	@100%	@25%	@50%	@100%
<b>Persons</b> (33,695 documents)								
Random	0.72	0.70	94.11	89.98	74.74	62.38	63.10	62.91
PsC	0.71	0.69	93.59	89.38	73.75	60.89	61.54	61.65
PsG	0.72	0.69	93.82	89.76	74.31	61.76	62.32	62.31
PsCG	0.71	0.68	93.52	89.28	73.46	60.06	60.66	61.08
<b>Locations - Outdoor</b> (28,226 documents)								
Random	0.68	0.64	85.53	77.76	66.80	60.93	61.08	59.48
GCD $\frac{2500}{750}$	0.66	0.61	81.39	73.65	63.82	56.04	57.01	56.61
GCD $\frac{750}{200}$	0.62	0.58	74.85	67.35	59.85	54.53	54.89	54.43
GCD $\frac{200}{25}$	0.59	0.56	71.89	64.71	58.19	52.73	53.34	53.22
<b>Locations - Indoor</b> (38,258 documents)								
Random	0.57	0.56	61.07	58.48	55.34	58.51	56.78	54.64
GCD $\frac{2500}{750}$	0.55	0.54	57.98	56.37	53.93	53.12	53.33	52.66
GCD $\frac{750}{200}$	0.57	0.55	60.47	57.98	54.93	55.87	55.10	53.71
GCD $\frac{200}{25}$	0.54	0.54	53.66	53.84	52.61	53.97	53.64	52.62
<b>Events</b> (15,467 documents)								
Random	0.70	0.70	89.52	83.63	71.87	74.70	71.98	66.70
EsP	0.59	0.57	66.43	62.68	57.63	57.53	56.60	55.10
<b>Context</b> (72,433 documents)								
Random	0.57	0.57	76.63	67.39	59.74	56.45	55.62	54.52
Top-25%	0.57	0.57	73.16	65.31	58.69	55.81	55.34	54.26
Top-10%	0.56	0.55	68.85	62.40	57.04	54.88	54.43	53.50
Top-5%	0.55	0.54	65.89	60.31	55.84	53.82	53.57	52.83

A.3 Results on other Subsets of TamperedNews and News400

Table A.13: Results for document verification (DV) and collection retrieval for the *News400* dataset. Results are reported for the Top-50% verified documents (sorted by the cross-modal similarity values obtained for the original entities) for different entity test sets (notations according to Section 4.4.1).

Test set	DV		Collection Retrieval					
	VA	AUC	AP-original [%]			AP-manipulated [%]		
			@25%	@50%	@100%	@25%	@50%	@100%
<b>Persons</b> (58 verified documents)								
Random	1.00	1.00	100.0	100.0	100.0	100.0	100.0	100.0
PsC	1.00	1.00	100.0	100.0	99.49	100.0	100.0	99.60
PsG	1.00	0.99	100.0	99.10	98.40	100.0	100.0	99.34
PsCG	1.00	1.00	100.0	100.0	100.0	100.0	100.0	100.0
<b>Locations - Outdoor</b> (27 verified documents)								
Random	1.00	0.99	100.0	100.0	99.29	100.0	100.0	99.42
GCD $\frac{2500}{750}$	0.89	0.93	93.79	92.61	89.94	100.0	100.0	95.36
GCD $\frac{750}{200}$	0.89	0.87	89.17	86.70	82.42	100.0	100.0	90.77
GCD $\frac{200}{25}$	0.89	0.86	89.17	86.70	82.25	100.0	100.0	89.96
<b>Locations - Indoor</b> (8 verified documents)								
Random	1.00	0.88	100.0	95.00	87.01	100.0	100.0	91.44
GCD $\frac{2500}{750}$	0.75	0.67	58.33	62.20	62.70	100.0	100.0	78.17
GCD $\frac{750}{200}$	1.00	0.78	100.0	85.42	77.81	100.0	100.0	85.27
GCD $\frac{200}{25}$	0.88	0.77	100.0	85.42	76.70	100.0	100.0	84.43
<b>Events</b> (16 verified documents)								
Random	1.00	0.97	100.0	100.0	96.18	100.0	100.0	97.92
EsP	0.81	0.82	56.67	66.36	71.42	100.0	100.0	88.76
<b>Context</b> (46 verified documents)								
Random	0.93	0.94	81.83	87.03	88.18	100.0	100.0	95.81
Top-25%	0.91	0.91	97.53	92.19	88.72	100.0	100.0	93.88
Top-10%	0.78	0.86	64.94	71.45	75.82	100.0	100.0	91.14
Top-5%	0.85	0.82	76.59	77.25	75.64	100.0	99.63	86.93





## Eric Müller-Budack

✉ eric.mueller@tib.eu

🏠 <https://eric-mb.github.io>

### EDUCATION

---

- 2014 **Master of Engineering in System Design**  
University of Applied Sciences Jena  
Master Thesis: *Dreidimensionale Objektvermessung aus Streifenbildern eines Zeitpunktes basierend auf Verstetigung mithilfe von Segmentvergleichen*
- 2012 **Bachelor of Engineering in Communication and Media Technology**  
University of Applied Sciences Jena  
Bachelor thesis: *Charakterisierung eines Multiapertur-Projektionssystems für die 3D-Vermessung mittels aktiver Streifenprojektion*
- 2007 **Abitur**  
Leuchtenburg Gymnasium Kahla

### WORK EXPERIENCE

---

- Feb 2016 – present **TIB - Leibniz Information Centre for Science and Technology**  
**Research assistant in the Research Group Visual Analytics**  
Information extraction from photos and videos using deep learning  
Cross-modal entity consistency in news articles  
Projects:  
• *TIB AV Analytics – Development of a software platform for systematic film and video analysis* (Jan 2021 – present, DFG, project number: 442397862)  
• Consulting project for *content garden technologies GmbH* (<https://content-garden.com>) on analysis of image content in advertisements (Jul 2019 – May 2020)  
• *VIVA – Visual Information Search in Video Archives* (Nov 2018 – Sep 2020, DFG, project number: 388420599)  
PhD thesis (2021): *Unsupervised Quantification of Entity Consistency between Photos and Text in Real-World News*
- Dec 2021 – present **Leibniz University Hannover**  
Sep 2019 – Dec 2019 **Research assistant at the L3S Research Center**  
Mar 2017 – Oct 2018 Automatic alpha matting using deep learning  
Multimodal deep learning approaches for information extraction from news videos  
Project: *FaAM – Fully-automated Alpha Matting for portrait photography* (Mar 2017 – Oct 2018, BMWi, ZIM-KOOP, project number: ZF4210002BZ6)
- Oct 2014 – Aug 2016 **University of Applied Sciences Jena**  
**Research assistant in the department Electrical Engineering and Information Technology**  
Person identification in videos  
Project: *Automatic methods for cost-efficient annotation of documentary film and video content* (Oct 2014 – Aug 2016, BMWi, ZIM-KOOP, project number: KF2135608KM3)
- Jun 2011 – Sep 2014 **Fraunhofer IOF Jena**  
**Student assistant in the department Imaging and Sensing**  
Photogrammetric 3D measurements

## RESEARCH ACTIVITIES

---

- Committee** Program committee member of the 1st International Workshop on Cross-lingual Event-centric Open Analytics, co-located with the 17th Extended Semantic Web Conference (ESWC) 2020
- Program committee member of the 2nd International Workshop on Cross-lingual Event-centric Open Analytics, co-located with the 30th The Web Conference (WWW) 2021
- Reviewer activities** Reviews for many peer-reviewed conferences (e.g., ACM MM, WACV, BMVC, ICMR, ICME), journals (e.g., MTAP, EURASIP), and workshops (e.g., MMSports, CLEOPATRA)
- Lectureships** Practical course on *Digital Image Processing II* at the University of Applied Sciences Jena (winter semester 2015/2016, 32 hours)
- Guest Lectures** Two guest lectures (2020, 2021) on *Pattern Recognition – Deep Learning* in the lecture *Multimedia Retrieval* from Prof. Ralph Ewerth at the Leibniz University Hannover
- One guest lecture (2021) on *Deepfake, Digital Resurrection, and Digital Forensics* in the lecture series *Critique. Alternative. Utopia. Documentary work on the social present and future* from Dr. Roman Mauer at the Johannes Gutenberg-Universität Mainz
- Tutorials** Tutorial on *Machine Learning – Deep Learning* for the CLEOPATRA Marie Skłodowska-Curie Innovative Training Network during the learning week in September 2019
- Supervision** Supervision of 14 theses, thereof six bachelor theses, and eight master theses

## HONORS & AWARDS

---

- 2020 **Best Paper Award** at the *ACM International Conference on Multimedia Retrieval (ICMR)* [175]
- 2018 **Honorable Mention Award** at the *International Conference on Theory and Practice of Digital Libraries (TPDL)* [171]
- 2015 Master thesis awarded with the **Thüringer VDI-Preis für die beste Studienabschlussarbeit**

## TECHNICAL SKILLS

---

- Operating systems** Experienced with Linux, Windows, and Mac OS X
- Office** Experienced in  $\text{\LaTeX}$ , Word, Excel, and Powerpoint
- Programming** Experienced in Python, Git, and Docker. Competent in bash, C, C++, Java, Javascript, MatLab, and Singularity
- Other** Experienced in HTML, SPARQL, Inkscape, photo and video editing (e.g., Gimp, Photoshop, Lightroom, Darktable, DaVinci Resolve, OBS Studio)

## LANGUAGES

---

- Mother tongue** German
- Other Languages** English (advanced, level C1), French (beginner A1)