IMPROVED CLASSIFICATION OF SATELLITE IMAGERY USING SPATIAL FEATURE MAPS EXTRACTED FROM SOCIAL MEDIA

Artem Leichter[1], Dennis Wittich[2], Franz Rottensteiner[2], Martin Werner[3] and Monika Sester[1]

[1] Institute of Cartography and Geoinformatics, Leibniz University Hannover, Germany
(Artem.Leichter, Monika.Sester)@ikg.uni-hannover.de
[2] Institute of Photogrammetry and GeoInformation, Leibniz University Hannover, Germany
(Wittich, Rottensteiner)@ipi.uni-hannover.de
[3] Institut für Methodik der Fernerkundung, Deutsches Zentrum für Luft- und Raumfahrt, Munich, Germany - martin.werner@dlr.de

Commission IV, WG IV/4

KEY WORDS: Deep Learning, Satellite Images, Classification, Social Media Mining, Data Fusion

ABSTRACT:

In this work, we consider the exploitation of social media data in the context of Remote Sensing and Spatial Information Sciences. To this end, we explore a way of augmenting and integrating information represented by geo-located feature vectors into a system for the classification of satellite images. For that purpose, we present a quite general data fusion framework based on Convolutional Neural Network (CNN) and an initial examination of our approach on features from geo-located social media postings on the Twitter and Sentinel images. For this examination, we selected six simple Twitter features derived from the metadata, which we believe could contain information for the spatial context. We present initial experiments using geotagged Twitter data from Washington DC and Sentinel images showing this area. The goal of classification is to determine local climate zones (LCZ). First, we test whether our selected feature maps are not correlated with the LCZ classification at the geo-tag position. We apply a simple boost tree classifier on this data. The result turns out not to be a mere random classifier. Therefore, this data can be correlated with LCZ. To show the improvement by our method, we compare classification with and without the Twitter feature maps. In our experiments, we apply a standard pixel-based CNN classification of the Sentinel data and use it as a *baseline model*. After that, we expand the input augmenting additional Twitter feature maps within the CNN and assess the contribution of these additional features to the overall F1-score of the classification, which we determine from spatial cross-validation.

1. INTRODUCTION

Humans are very powerful sensors, capable of not only perceiving the world, but interpreting it as well. Social media give us the opportunity to access information gathered by humans and thus can be seen as an interface to access this powerful sensor. The usage of this implicit, user generated data can reduce the demand for explicit data, which has to be gathered with additional overhead and costs. One of the scenarios where user generated data could help, is the development of land use and land cover classification models.

The classification of land use and land cover has been an important research topic in remote sensing for a couple of decades (Anderson, 1976), (Foody, 2002). Recently, a world-wide data set with a resolution of 30m has been created ((Chen et al., 2015)). In the last decades, fine-grained classification of urban regions has shown great potential in understanding urban dynamics. To this end, Oke et al. proposed the local climate zones classification scheme which has seen wide adoption (Stewart and Oke, 2012), (Stewart et al., 2014). In this scheme, 17 classes have been defined that jointly cover surface structure (height and density) and surface cover (pervious or impervious). The most amount of publicly available LCZ data is provided by the World Urban Database and Access Portal Tools (WUDAPT). This organization provides rastered LCZ data for a couple of cities around the world. Ground truth for local climate zone classification has been acquired by individuals for a few cities in the WUDAPT project (Mills et al., 2015). Many efforts have been taken to predict LCZ classes from various remote sensing sources (Yokoya et al., 2018) (Bechtel and Daneke, 2012)(Bechtel et al., 2015) most notably including the last edition of the IEEE GRSS Data Fusion Contest in 2017.

Social media as novel modality is quite orthogonal to usual remote sensing data: it is very sparse, it is local, it is mostly generated by humans thereby exploiting human knowledge, but it is noisy. While the text content of tweets is difficult to relate to rather morphological LCZ classification, it can still provide hints on how to distinguish different classes that look similar from space. In this paper, however, we focus on aggregated tweet metadata including the number of tweets and the type of users. We aggregate this metadata on a coregistered grid with downsampled Sentinel imagery. In this work we first show that Twitter data contains implicit information about land use and land cover (Experiment 1) and that this data improves the LCZ classification with a baseline Convolutional Neural Network (CNN) based on satellite imagery (Experiment 2). Note that our experimental setup is considered not to provide the best possible classification performance as this always comes at the risk of high overfitting and overly targeted models. Instead, a simple baseline classification system is analyzed with respect to its behavior with and without Twitter data on the classification quality.

The remainder of this paper is structured as follows: In chapter 2 we give a brief overview on the background of this work including the LCZ system, CNNs for different image related tasks and Twitter data. We continue in chapter 3 with a more detailed description of the dataset we created and used. In chapter 4 we describe our methodological approach. In the subsequent two chap-

ters 5 and 6 we focus on the experiments including our evaluation strategy. The results of the experiments are presented in chapter 7, followed by a discussion in chapter 8. In the last chapter we conclude our work and give a brief outlook regarding possible future work.

2. BACKGROUND

Local climate zones are a land use / land cover classification system currently receiving high attention within the GI community. It has a focus on the built environment, which leads to its capability to improve climatic modeling (Stewart and Oke, 2012) and to provide a generalized and therefore comparable representation of urban architectural topographies (Taubenböck et al., 2012).

Built types	Definition
1. Compact high-rise	Dense mix of tall buildings to tens of stories. Few or no trees. Land cover mostly paved. Concrete, steel, stone, and glass construction materials.
2. Compact mid-rise	High density of massive buildings with height 10 m to 25 m.
3. Compact low-rise	Dense mix of midrise buildings (3–9 stories). Few or no trees. Land cover mostly paved. Stone, brick, tile, and concrete construction materials.
4. Open high-rise	Dense mix of low-rise buildings (1–3 stories). Few or no trees. Land cover mostly paved. Stone, brick, tile, and concrete construction materials.
5. Open mid-rise	Open arrangement of tall buildings to tens of stories. Abundance of pervious land cover (low plants, scattered trees). Concrete, steel, stone, and glass construction materials.
6. Open low-rise	Open arrangement of midrise buildings (3–9 stories). Abundance of pervious land cover (low plants, scattered trees). Concrete, steel, stone, and glass construction materials.
7. Lightweight low-r.	Dense mix of single-story buildings. Few or no trees. Land cover mostly hard-packed. Lightweight construction materials (e.g., wood, thatch, corrugated metal).
8. Large low-rise	Open arrangement of large low-rise buildings (1–3 stories). Few or no trees. Land cover mostly paved. Steel, concrete, metal, and stone construction materials.
9. Sparsely built	Sparse arrangement of small or medium-sized buildings in a natural setting. Abundance of pervious land cover (low plants, scattered trees).
10. Heavy industry	Low-rise and midrise industrial structures (towers, tanks, stacks). Few or no trees. Land cover mostly paved or hard-packed. Metal, steel, and concrete construction materials.

Table 1. LCZ classification system, built-up structures.

Land cover types	Definition
A. Dense trees	Heavily wooded landscape of deciduous and/or evergreen trees. Land cover mostly pervious (low plants). Zone function is natural forest, tree cultivation, or urban park.
B. Scattered trees	Lightly wooded landscape of deciduous and/or evergreen trees. Land cover mostly pervious (low plants). Zone function is natural forest, tree cultivation, or urban park.
C. Bush, scrub	Open arrangement of bushes, shrubs, and short, woody trees. Land cover mostly pervious (bare soil or sand). Zone function is natural scrubland or agriculture.
D. Low plants	Featureless landscape of grass or herbaceous plants/crops. Few or no trees. Zone function is natural grassland, agriculture, or urban park
E. Bare rock or paved	Featureless landscape of rock or paved cover. Few or no trees or plants. Zone function is natural desert (rock) or urban transportation.
F. Bare soil or sand	Featureless landscape of soil or sand cover. Few or no trees or plants. Zone function is natural desert or agriculture.
G. Water	Large, open water bodies such as seas and lakes, or small bodies such as rivers, reservoirs, and lagoons.

Table 2. LCZ classification system, land cover structures

LCZ is a raster data set, where each pixel is assigned a single class from the 17 classes, which are equally defined all over the world. The Tables 1 and 2 show the definitions of the classes from (Stewart and Oke, 2012). The classes 1 to 10 describe areas with buildings, they are derived from parameters like height of the buildings, ratio of width and height and materials. The classes A to F describe land cover and derived from parameters like height of trees, density of trees and soil type. And finally class G represents water. This classification uses resolutions between 100 m and 200 m, because it is not constructive to describe topographic layout and morphology in a finer resolution. Large amount of work in context of LCZ is directed toward classifying single cities (Danylo et al., 2016) (Verdonck et al., 2017).

Two organisation are responsible for collection of LCZ data and their securing their quality, those are GeoWiki and World Urban Database and Portal (WUDAPT). The data contribution is driven by croudsourcing campaigns (See et al., 2013) (Foody et al., 2013) and games (Laso Bayas et al., 2016).

Convolutional Neural Networks (CNNs) are deep learning architectures, inspired by the visual perception mechanism of animals based on receptive fields in the visual cortex (Gu et al., 2018). CNNs have been known for a couple of decades, e.g. LeChun *et al.* (LeCun et al., 1990) showed in 1989 that CNNs can be used to classify images of hand-written digits. Since the mid 2000s CNNs started to get more popular because the previously occurring problems like the lack of training data and computational resources became less relevant (Gu et al., 2018). This was on the one hand due to the improvement of the methods on the other

hand due to the usage of more efficient hardware. Nowadays CNN architectures reach state-of-the-art performance in many image related tasks like image classification, which is the task of assigning the correct class to an image (or image patch) or semantic segmentation, where the goal is to assign the correct class to each pixel of the input image.

Twitter data is present in a large amount, but it is unequally distributed in space. There are areas where only sparse information is present, as opposed to other, mostly urban areas with high amount of Twitter data. Among all the Tweets, the amount of geotagged messages is rather low, and also differs from country to country. Geotagged tweets have a very unclear relation between their content and their location: usually people use the it as communication medium and not as means to convey information about the environment. However, in emergency scenarios (such as earthquakes, flooding), people also talk about it and thus, the content can be beneficially exploited (see e.g. (Dittrich et al., 2015)). Independent of this problems there are several research topics on employing spatial information like automated derivation of features with spatial relevance (Sengstock and Gertz, 2012) or inferring home locations (Lin and Cromley, 2018) from geotagged tweets. Tweets contain several types of information, like text image/video and location, which can be fused in order to identify spatial events such as floods (Feng and Sester, 2018). Nevertheless simple meta information like user mentions, count, and tag count can be used to distinguish between types of users (Guo and Chen, 2014).

3. DATASET

When we created the dataset to proof our hypothesis, the biggest limitation was the availability of LCZ data. This examination focuses on the local area around Washington DC, defined by the corresponding LCZ label raster provided by WUDAPT, visualized in Figure1. We decided to investigate this area because it seemed suitable for our purpose due to high amount of Twitter data as well as the availability of open satellite imagery. The LCZ label raster is used as ground truth in our experiments. Although the provided labels are not annotated by humans, but instead created by a classification algorithm, we assume the data to be error free. The LCZ classes are distributed in this area as shown in Table 3. The classes 7 (Lightweight low-rise), 10 (Heavy industry), C (Bush scrub) and F (Bare soil or sand) are not present in this data set.

The Twitter data set is generated by filtering 680.982.894 Tweets, collected in the time between 09 February 2018 and 19 June 2018, by their location with respect to our target area. The result is 392559 tweets with geotags in Washington DC. The spatial distribution of the Tweets is shown in Figure 2. It is obvious, that most of the Tweets are present in the urban agglomerations.

For the baseline of the CNN we take advantage of the publicly available satellite imagery, gathered during the SENTINEL-2 earth observation mission developed by the ESA. In particular we used rectified and georeferenced images of different spectral bands (Processing Level-1C). We decided to use the infrared- red- green- and blue channels with a spatial resolution of 10 m as well as the bands 11 and 12 with a spectral resolution around 1610 nm and 2190 nm respective. These two bands have a spatial resolution of 20 m.

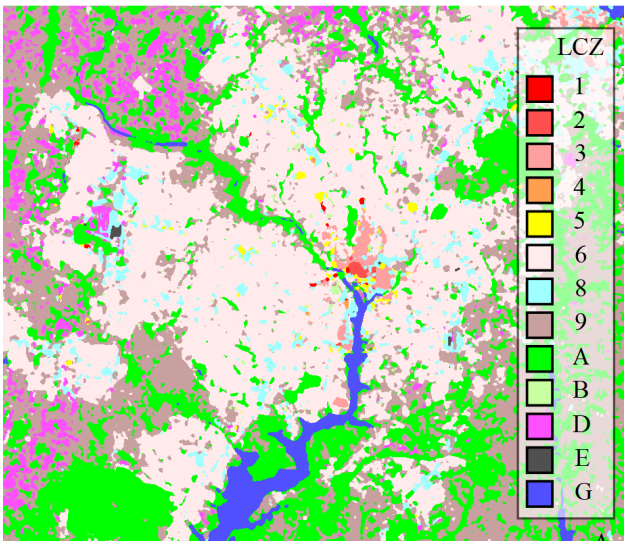

Figure 1. LCZ reference data, colored by LCZ class.

4. METHODOLOGY

The Washington DC tweets are rasterized according to the area and cell size of the ground truth. The Tweet count per cell is shown in the Figure 2. The resulting rasters contain large areas without Twitter data. Tweet activity is mostly concentrated in urban areas, especially in city centers. The distribution of the LCZ classes within the areas with Twitter data is shown in third column in table 3. Twitter data is present only for 2.36 % of the target area. As expected, the representation of rural classes like D (Low plants) and A (Dense trees) is lower then in the overall distribution, which can be seen in second column of Table 3 We generate six feature maps which are used in two experiments. Those features were selected due to their potential relevance for LCZ classes. The feature maps are generated as follows:

1. **Tweet count (TC)** - summed tweets with geotag in a particular cell. This information could help distinguish between urban and rural area. TC contains integer values from zero up to 19 thousand.
2. **Mean text length (MTL)** - mean count of symbols per tweet. Text length is a simple feature which could help distinguish between private, casual tweets and bussiness tweets of companies. More general, this feature can contain information whether or not this is a spontaneous tweet or it is a well planed tweet with optimized content dense. Areas, where companies place their offices tend to have specific structure, which can be related to LCZ classes. This feature is a mean over tweet length, which is limited by 280 symbols constraint.
3. **Mean friends count (MFC)** - mean count of friends of the author of the tweet. Similar as MTL this feature could be useful to identify business usage. Companies gather all possible followers in order to reach most potential customers.The values of this feature vary up to mean 193,066.0 friends per cell.
4. **Mean time (MT)** - Hour of time, when the tweet was posted. Tweet time could indicate information whether it is residential area or recreation area. This raster contains the mean over values in range of 0 to 23, representing round down local hour of time.

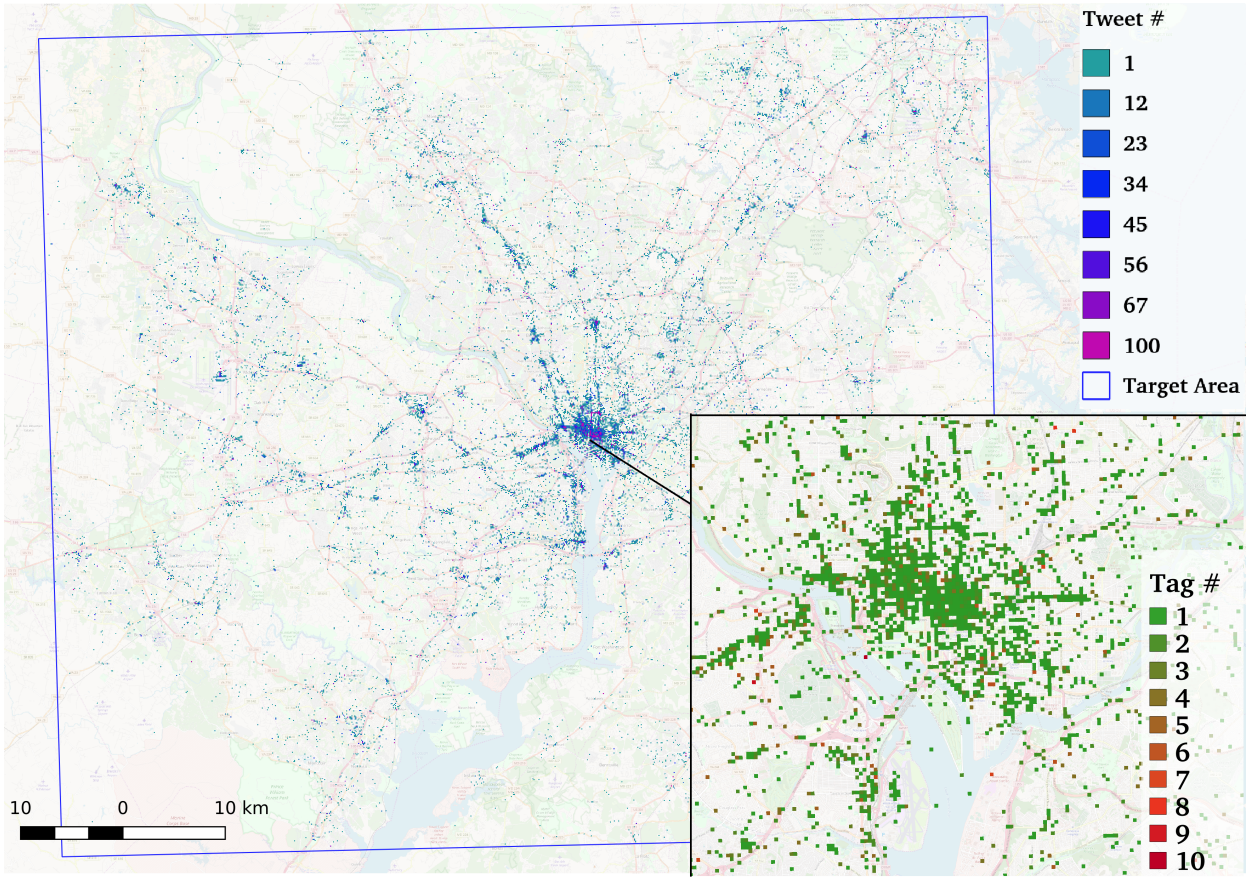

Figure 2. Spatial distribution of the Twitter data. Colors represent number of tweets in a single cell TC.The single cell hot spots are not visible in this view. Right lower corner zoom in: Tag count MTC value distribution.

5. **Mean tag count (MTC)** - mean count of tags used in a particular tweet. This could contain information about different user types (Diansheng and Chao, n.d.). Different user types could be an indicator for different LCZ areas. Values of this feature vary in the range of 0 to 11
6. **Mean user mentions count (MMC)** - mean count of Twitter users mentioned in a particular tweet. Same relevance as MTC. MTC

These Twitter features are used in the two experiments. The first experiment uses only the six Twitter feature maps to predict LCZ classes using boost trees. No additional information is used.

For the second experiment we prepare the Twitter data feature maps for the training, by means of linear normalization to the interval from 0 to 1. The feature maps TC and MFC contain hot spots with few single cells containing very high values. Linear normalization of such a feature map would contain in most cell very small normalized values, therefore we crop the values above the 95th percentile.

The second experiment uses a Convolutional Neural Network for a pixel wise (cell wise) prediction of LCZ classes based on different bands of satellite imagery and the six Twitter feature maps. In order to investigate the influence of the Twitter data we train and evaluate two classification models. The first one, referred to as *baseline model*, infers the LCZ classes based on the satellite images only. The second version of the model additionally uses the Twitter data and is referred to as *augmented model*.

5. FIRST EXPERIMENT

In the first experiment we use the six earlier described feature maps as input. The input data has the following structure for a single entry:

$$TC|MTL|MFC|MT|MTC|MMC||LCZClass$$

The distribution of the classes within the Twitter data can be seen in third column of the table 3.

In order to combat overrepresentation we apply random undersampling. All classes have at least 540 samples, except class A with 475 samples. Therefore only the following classes are trained in this experiment: $\{2, 3, 5, 6, 8, 9, A\}$ Total number of used samples is 3715. All the data is divided into test set (67%) and test set (33%).

The classification model is implemented using XGBClassifier from the eXtreme Gradient Boosting Package provided by the Distributed (Deep) Machine Learning Community (DMLC). The metaparameters are optimized by means of a grid search. The number of estimators is selected using cross validation with 5 folds.

In order to avoid the influence of the outliers, the experiment is repeated 10 times. For each repetition the following performance metrics are calculated for the test set and finally averaged over all 10 repetitions: F1 score per class, overall accuracy, recall and precision.

LCZ Class	Support			F1 [%]	
	total [%]	Twitter [%]	Twitter [#]	baseline	augmented
1 Compact high-rise	0.06	0.03	234	24.7	34.8
2 Compact mid-rise	0.11	0.07	540	34.8	55.0
3 Compact low-rise	0.97	0.17	1308	48.3	53.3
4 Open high-rise	0.02	0.01	58	3.2	0.45
5 Open mid-rise	0.45	0.09	641	27.7	30.8
6 Open low-rise	37.07	1.19	8687	86.6	87.1
7 Lightweight low-r.	0.00	-	-	-	-
8 Large low-rise	3.86	0.57	4129	72.4	72.9
9 Sparsely built	27.34	0.12	854	70.1	71.1
10 Heavy industry	0.00	-	-	-	-
A Dense trees	23.22	0.07	475	79.6	81.1
B Scattered trees	0.24	0.004	29	33.1	37.3
C Bush, scrub	0.00	-	-	-	-
D Low plants	4.18	0.01	64	56.0	62.2
E Bare rock or paved	0.06	0.01	61	15.7	11.2
F Bare soil or sand	0.00	-	-	-	-
G Water	2.43	0.01	69	92.3	92.5
Sum	100	2.36	17149	-	-
Mean	-	-	-	49.9	53.1

Table 3. LCZ support and F1 score per class.

6. SECOND EXPERIMENT

In the second experiment we investigate, if the LCZ classification results of a CNN based on satellite images can be improved by additionally feeding the generated Twitter feature maps to it. Therefore we set up a fully Convolutional Neural Network. In comparison to a standard image to class label network where input and output are the same spatial resolution, we have to deal with different spatial resolutions of the input and reference data. The highest ground sampling distance (GSD) of 10 m comes with the infrared-, red-, green- and blue channels of the satellite images. The other two used bands 11 and 12 have a GSD of 20 m and the Twitter feature maps as well as the reference label maps have a GSD of 100 m.

Instead of sampling the satellite images down by a factor of 10 and 5 respectively, we decided to create a network with a downsampling architecture, which is capable to receive the input data within the highest available resolution. This is realized by using the input data with the highest spatial resolution as the first input and then performing strided convolutions to reduce the size of the intermediate feature maps. The remaining input images with lower resolutions as well as the Twitter feature maps are then concatenated to the intermediate feature maps with the according size. Besides the benefit of using the available input data in maximum resolution, the developed architecture has two additional advantages. Firstly, since the network is fully convolutional, it is possible to feed different sized images to the network (as long as the ratio between the different input data matches). This is used in terms of training the network on patches, where the size of the first input images is 250 x 250 px and evaluating the network on bigger parts of the image. The second advantage refers our investigation regarding the Twitter data. In order to set up the *baseline model* we simply skip the concatenation of the Twitter data, without modifying the rest of the network. The network architecture of the *augmented model* is shown in Figure 3. The filter sizes were chosen to create an overlap during the strided convolution and additionally infer an appropriate perceptive field of 1.6 x 1.6 km for the classification of each target tile. All convolutions are followed by adding a bias and applying the leaky rectified linear unit as activation function.

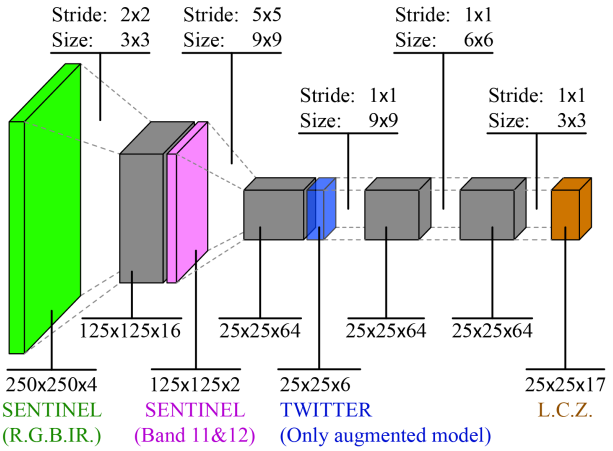

Figure 3. Architecture of the Convolutional Neural Network. On top of the network the details of the convolution operation are shown. Below we displayed the size of input and feature maps. The blue box represents the Twitter feature maps, which are only used in the *augmented model*.

In order to evaluate the models in a four-fold cross validation, the input and reference images are divided into four quarters of same size. In the second step, 5000 randomly rotated patches are extracted out of each quarter. During the extraction of a rotated patch we use bi-cubic interpolation for the satellite images and nearest neighbour interpolation for the Twitter feature maps as well as the reference LCZ data. Each resulting patch contains the satellite images, the Twitter feature maps as well as the reference labels of the same 2.5 x 2.5 km square. Both networks are trained four times on the patches of three quarters in a round robin fashion. The respectively last, unseen quarter of each run is used for the evaluation of the network. Each run involves a fixed number of training iterations. In each iteration the network predicts 4800 patches and the soft-max loss of the predicted classes w.r.t. the reference data is minimized using gradient descent. We weight the losses of each LCZ class according to the overall class distribution in the investigated area.

We decided for a fixed number of iterations as a stopping criterion, since we do not use an additional validation set due to the low amount of data. We choose iteration 500 by analyzing the mean F1 score for the test set over 1000 epochs for both models and all quarters as test set. In Figure 4 the mean F1 score for the upper right quarter over 1000 training epochs is shown, where the network uses the patches of the other three quarters as training data. The shown graphs are slightly smoothed using the mean over a window of five epochs for visualization purpose. It can be seen, that the mean F1 score reaches a plateau at epoch 500 for both models.

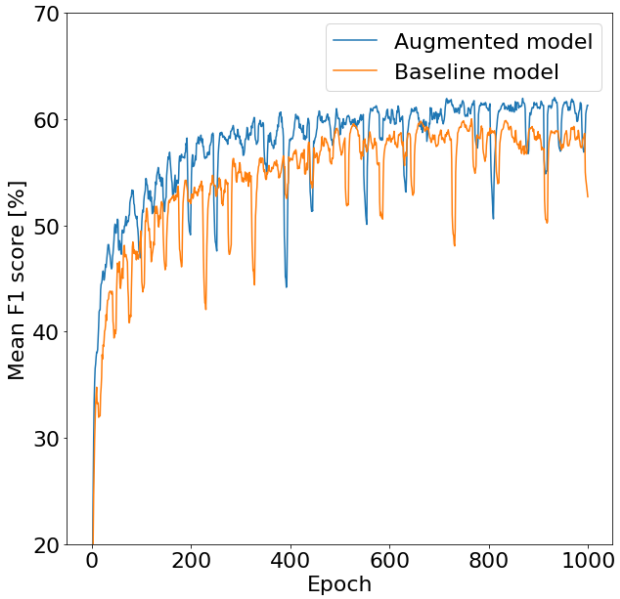

Figure 4. Exemplary smoothed trend of the mean F1 score in the test quarter for both models over 1000 iterations. Both models reach a plateau after approximately 500 training epochs.

For the final evaluation we repeat the training 5 times for each quarter as test set with different seeds for the network parameters (weights and biases) to ensure that the results are not due to good or bad initial values. To evaluate both models we calculate the confusion matrix after each epoch for the testing quarter. These matrices are summed up epoch wise over all runs for both models respectively, resulting in one confusion matrix per model and epoch. These matrices are then used to calculate further quality measurements like the F1 score, recall and precision per class as well as the mean F1 score and the overall accuracy.

7. RESULTS

In the first experiment the resulting mean overall accuracy of the test set classification reached 25.5 % which is slightly above the performance of a random classifier (7 classes, corresponding to 14.3 %). It is worth to note that the performance of the LCZ class 2 (table 4) is far above random classification.

Class	F1 score[%]	Recall [%]	Precision [%]
2	48.8	61.1	40.6
3	18.5	16.6	21.0
5	16.8	15.2	19.7
6	19.8	19.0	21.4
8	21.0	20.5	22.2
9	26.9	33.2	22.7
A	15.7	13.0	20.8

Table 4. Results of the First experiment.

The results of the second experiment are shown in the last two columns of table 3. There we compare the F1 scores of each class for both models. For 11 out of the 13 classes which are represented in the data, the *augmented model* reaches a higher F1 score. This trend is also represented in the mean F1 scores, shown in the last row. The overall accuracy is also 1.3 % higher in the *augmented model* with a value of 78.6 %. The absolute differences in the F1 scores per class are visualized in Figure 5.

Although the improvement of the *augmented model* is below 5 % for most classes, the classes 1 and 2 show an improvement of approx. 10 % and 20 % respectively. The F1 score for the classes 4 and E decreased by less than 5 % in comparison to the *baseline model*. In Figure 6 the mean F1 scores for both models are shown for each epoch. The *augmented model* over performs the *baseline model* for most epochs.

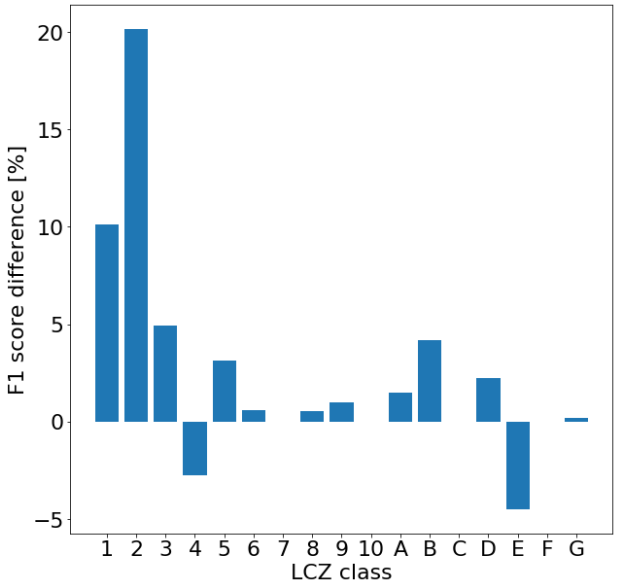

Figure 5. Difference in F1 score per class. Positive value indicates a better score for the *augmented model*.

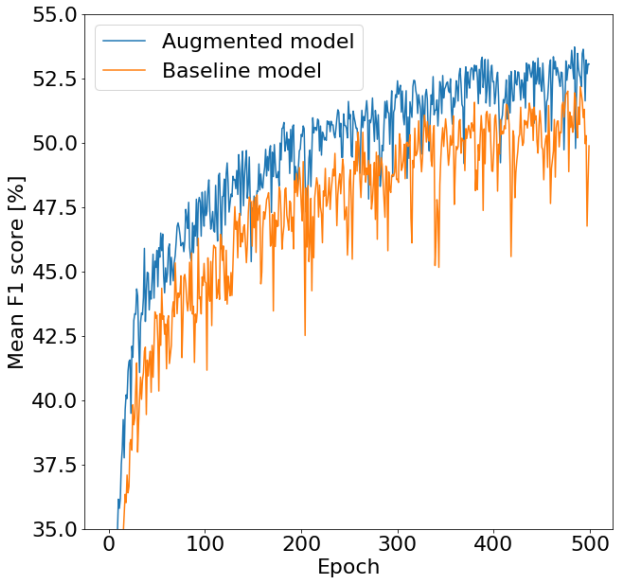

Figure 6. Mean F1 score for both models over 500 training epochs. The values are based on the summed confusion matrices over all runs and quarters per epoch and model.

8. DISCUSSION

Since the overall accuracy achieved in the first experiment is better than a random classifier, we conclude that Twitter data contains information suitable for LCZ classification. As expected,

the quality is low, due to missing explicit content information and noisy nature of the Twitter data. In second experiment we gathered additional evidence for the positive effect of Twitter data on a more realistic classification scenario. The results also imply, that our developed architecture is capable of fusing the dense satellite images and the sparse Twitter data. In both experiments this improvement was most prominent for LCZ class 2. Against our expectations we did not observe a dependency between the amount of Twitter data and the improvement of the prediction of a class. We assume, that this is due to the fact, that the absence of Twitter data itself is a valuable information for the implemented classification model.

Prominent is the F1-score decrease of the class 4 and E. As described in section 6 we combat the imbalanced distribution of the data by applying weight penalties to over presented classes. This penalties are proportional to the count of the particular class in the whole training area. The class 4 and E are under represented in Twitter data so that few noise instances of Twitter data could cause the performance loss.

As a conclusion we state that we proved our initial hypothesis about the beneficial contribution of Twitter data on land use classification and our classification approach is suitable for data fusion.

9. OUTLOOK

In this work we use most simple and straight forward Twitter data features. It would be a logical step to investigate, if more complex features derived from tweet text or pictures improve the classification results even more. Since we assume that many tweets are not related to their tagged location, those features may also help to provide more specific or additional locally related information to the model. On the other hand we did not investigate, if all of the used features derived from the Twitter data are actually relevant. In future work we want to train the model using only a subset of the features to get a better understanding about the relevant information.

Furthermore our goal was to investigate the improvement of land use classification using Twitter data. Since we proved this hypothesis, we now want to develop a more sophisticated and optimized classification model. Considering more training and testing data, e.g. including different cities, we could evaluate our approach on a more realistic scenario. This would additionally allow us to see whether the incorporation of Twitter data can improve the models capability of generalization.

ACKNOWLEDGEMENTS

This work was partially funded by the Federal Ministry of Education and Research, Germany (Bundesministerium für Bildung und Forschung, Förderkennzeichen 01IS17076). We gratefully acknowledge this support.

REFERENCES

Anderson, J., 1976. *A Land use and land cover classification system for use with remote sensor data*. Geological Survey professional paper, U.S. Govt. Print. Off.

Bechtel, B., Alexander, P. J., Böhner, J., Ching, J., Conrad, O., Feddema, J., Mills, G., See, L. and Stewart, I., 2015. Mapping local climate zones for a worldwide database of the form and function of cities. *ISPRS International Journal of Geo-Information* 4(1), pp. 199–219.

Bechtel, B. and Daneke, C., 2012. Classification of local climate zones based on multiple earth observation data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5(4), pp. 1191.

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M. et al., 2015. Global land cover mapping at 30 m resolution: A pok-based operational approach. *ISPRS Journal of Photogrammetry and Remote Sensing* 103, pp. 7–27.

Danylo, O., See, L., Bechtel, B., Schepaschenko, D. and Fritz, S., 2016. Contributing to WUDAPT: A Local Climate Zone Classification of Two Cities in Ukraine. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9(5), pp. 1841–1853.

Diansheng, G. and Chao, C., n.d. Detecting non-personal and spam users on geo-tagged twitter network. *Transactions in GIS* 18(3), pp. 370–384.

Dittrich, A., Vasardani, M., Winter, S., Baldwin, T. and Liu, F., 2015. A classification schema for fast disambiguation of spatial prepositions. In: *Proceedings of the 6th ACM SIGSPATIAL International Workshop on GeoStreaming*, ACM, pp. 78–86.

Feng, Y. and Sester, M., 2018. Extraction of pluvial flood relevant volunteered geographic information (vgi) by deep learning from user generated texts and photos. *ISPRS International Journal of Geo-Information* 7(2), pp. 39.

Foody, G. M., 2002. Status of land cover classification accuracy assessment. *Remote sensing of environment* 80(1), pp. 185–201.

Foody, G. M., See, L., Fritz, S., Van der Velde, M., Perger, C., Schill, C. and Boyd, D. S., 2013. Assessing the accuracy of volunteered geographic information arising from multiple contributors to an internet based collaborative project. *Transactions in GIS* 17(6), pp. 847–860.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J. and Chen, T., 2018. Recent advances in convolutional neural networks. *Pattern Recogn.* 77(C), pp. 354–377.

Guo, D. and Chen, C., 2014. Detecting non-personal and spam users on geo-tagged twitter network. *Transactions in GIS* 18(3), pp. 370–384.

Laso Bayas, J., See, L., Fritz, S., Sturn, T., Perger, C., Dürauer, M., Karner, M., Moorthy, I., Schepaschenko, D., Domian, D. and McCallum, I., 2016. Crowdsourcing in-situ data on land cover and land use using gamification and mobile technology. 8, pp. 905.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E. and Jackel, L. D., 1990. Handwritten digit recognition with a back-propagation network. In: D. S. Touretzky (ed.), *Advances in Neural Information Processing Systems 2*, Morgan-Kaufmann, pp. 396–404.

Lin, J. and Cromley, R. G., 2018. Inferring the home locations of twitter users based on the spatiotemporal clustering of twitter data. *Transactions in GIS* 22(1), pp. 82–97.

Mills, G., Ching, J., See, L., Bechtel, B. and Foley, M., 2015. An Introduction to the WUDAPT project. In: *Proceedings of the 9th International Conference on Urban Climate, Toulouse, France*, pp. 20–24.

See, L., Comber, A., Salk, C. F., Fritz, S., van der Velde, M., Perger, C., Schill, C., McCallum, I., Kraxner, F. and Obersteiner, M., 2013. Comparing the quality of crowdsourced data contributed by expert and non-experts. In: *PloS one*.

Sengstock, C. and Gertz, M., 2012. Latent geographic feature extraction from social media. In: *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL ’12, ACM, New York, NY, USA, pp. 149–158.

Stewart, I. D. and Oke, T. R., 2012. Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society* 93(12), pp. 1879–1900.

Stewart, I. D., Oke, T. R. and Krayenhoff, E. S., 2014. Evaluation of the ‘local climate zone’scheme using temperature observations and model simulations. *International journal of climatology* 34(4), pp. 1062–1080.

Taubenböck, H., Esch, T., Felbier, A., Wiesner, M., Roth, A. and Dech, S., 2012. Monitoring urbanization in mega cities from space. *Remote Sensing of Environment* 117, pp. 162 – 176. Remote Sensing of Urban Environments.

Verdonck, M.-L., Okujeni, A., van der Linden, S., Demuzere, M., Wulf, R. D. and Coillie, F. V., 2017. Influence of neighbourhood information on ‘local climate zone’ mapping in heterogeneous cities. *International Journal of Applied Earth Observation and Geoinformation* 62, pp. 102 – 113.

Yokoya, N., Ghamisi, P., Xia, J., Sukhanov, S., Heremans, R., Tankoyeu, I., Bechtel, B., Le Saux, B., Moser, G. and Tuia, D., 2018. Open data for global multimodal land use classification: Outcome of the 2017 IEEE GRSS Data Fusion Contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11(5), pp. 1363–1377.