SoccerNet 2022 Challenges Results

Silvio Giancola[*,†]
silvio.giancola@kaust.edu.sa
KAUST
Thuwal, Saudi Arabia

Anthony Cioppa[*,†]
anthony.cioppa@uliege.be
University of Liège
Liège, Belgium

Adrien Deliège[†]
adrien.deliege@uliege.be
University of Liège
Liège, Belgium

Floriane Magera[†]
f.magera@evs.com
University of Liège, & EVS Broadcast Equipment
Liège, Belgium

Vladimir Somers[†]
v.somers@sportradar.com
Sportradar, & UCLouvain, & EPFL
London, United Kingdom

Le Kang[†]
deepconv@gmail.com
Baidu Research
Sunnyvale, USA

Xin Zhou[†]
chow459@gmail.com
Baidu Research
Sunnyvale, USA

Olivier Barnich[†]
o.barnich@evs.com
EVS Broadcast Equipment
Liège, Belgium

Christophe De Vleeschouwer[†]
christophe.devleeschouwer@uclouvain.be
UCLouvain
Louvain-la-Neuve, Belgium

Alexandre Alahi[†]
alexandre.alahi@epfl.ch
EPFL
Lausanne, Switzerland

Bernard Ghanem[†]
bernard.ghanem@kaust.edu.sa
KAUST
Thuwal, Saudi Arabia

Marc Van Droogenbroeck[†]
M.VanDroogenbroeck@uliege.be
University of Liège
Liège, Belgium

Abdulrahman Darwish
abdulrahman.darwish@guc.edu.eg
German University in Cairo
New Cairo City, Egypt

Adrien Maglo
adrien.maglo@cea.fr
Université Paris-Saclay, CEA, List
Paris, France

Albert Clapés
alcl@create.aau.dk
Aalborg University
Aalborg, Denmark

Andreas Luyts
andreas.luyts@rebatch.be
ReBatch
Kontich, Belgium

Andrei Boiarov
andrei.boiarov@sit.team
Schaffhausen Institute of Technology
Schaffhausen , Switzerland

Artur Xarles
arturxe@gmail.com
Universitat de Barcelona
Barcelona, Spain

Astrid Orcesi
astrid.orcesi@cea.fr
Université Paris-Saclay, CEA, List
Paris, France

Avijit Shah
avijit.shah@yahooinc.com
Yahoo Research
Sunnyvale, USA

Baoyu Fan
fanbaoyu@inspur.com
Inspur Electronic Information Industry Co., Ltd. State Key Laboratory of High-end Server Storage Technology
Jinan, China

Bharath Comandur
cjrbharath@gmail.com
Purdue University
West Lafayette, USA

Chen Chen
chenchen@oppo.com
OPPO Research Institute
Shenzhen , China

Chen Zhang
zhangchen4@oppo.com
OPPO Research Institute
Shenzhen , China

Chen Zhao
zhaochen03@baidu.com
Department of Augmented Reality Technology (ART), Baidu Inc
Beijing, China

Chengzhi Lin
linchzh3@mail2.sysu.edu.cn
Sun Yat-sen University
Guangzhou, China

Cheuk-Yiu Chan
cy3chan@cihe.edu.hk
Caritas Institute of Higher Education
Tseung Kwan O, Hong Kong, SAR

Chun-Chuen Hui
cchui@cihe.edu.hk
Caritas Institute of Higher Education
Tseung Kwan O, Hong Kong, SAR

Dengjie Li
lidengjie@meituan.com
Meituan Inc.
Beijing, China

Fan Yang
fan.yang@fujitsu.com
Fujitsu Research
Kawasaki, Japan

Fan Liang
liangfan02@meituan.com
Meituan Inc.
Beijing, China

Fang Da
fang@qcraft.ai
QCraft Inc.
Beijing, China

Feng Yan
yanfeng05@meituan.com
Meituan Inc.
Beijing, China

Fufu Yu
fufuyu@tencent.com
Tencent Youtu Lab
Shanghai, China

Guanshuo Wang
mediswang@tencent.com
Tencent Youtu Lab
Shanghai, China

H. Anthony Chan
hhchan@cihe.edu.hk
Caritas Institute of Higher Education
Tseung Kwan O, Hong Kong, SAR

He Zhu
zhuh20@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Hongwei Kan
kanhongwei@inspur.com
Inspur Electronic Information
Industry Co., Ltd. State Key
Laboratory of High-end Server
Storage Technology
Jinan, China

Jiaming Chu
chujiaming886@bupt.edu.cn
OPPO Research Institute, & Beijing
University of Posts and
Telecommunications
Shenzhen , China

Jianming Hu
hujm@mail.tsinghua.edu.cn
Tsinghua University
Beijing, China

Jianyang Gu
gu_jianyang@zju.edu.cn
OPPO Research Institute, & Zhejiang
University
Shenzhen , China

Jin Chen
chenjing@mgtv.com
MGTV
Changsha, China

João V. B. Soares
jvbsoares@yahooinc.com
Yahoo Research
Sunnyvale, USA

Jonas Theiner
theiner@l3s.de
L3S Research Center, Leibniz
University Hannover
Hannover, Germany

Jorge De Corte
jorge.decorte@rebatch.be
ReBatch
Kontich, Belgium

José Henrique Brito
jbrito@ipca.pt
2Ai – School of Technology IPCA
São Martinho, Portugal

Jun Zhang
bobbyjzhang@tencent.com
Tencent Youtu Lab
Shanghai, China

Junjie Li
serenitycapo@gmail.com
Tencent Youtu Lab, & Shanghai Jiao
Tong University
Shanghai, China

Junwei Liang
junweiliang1114@gmail.com
Tencent Youtu Lab
Shanghai, China

Leqi Shen
lunarshen@gmail.com
Tsinghua University
Beijing, China

Lin Ma
linma@alumni.cuhk.net
Meituan Inc.
Beijing, China

Lingchi Chen
lingchi@mgtv.com
MGTV
Changsha, China

Miguel Santos Marques
a18888@alunos.ipca.pt
2Ai – School of Technology IPCA
São Martinho, Portugal

Mike Azatov
mazatov@gmail.com
Arsenal FC
London, United Kingdom

Nikita Kasatkin
nk@sit.team
Schaffhausen Institute of Technology
Schaffhausen, Switzerland

Ning Wang
wangning12@mail.ecust.edu.cn
OPPO Research Institute, & East China University of Science and Technology
Shenzhen , China

Qiong Jia
boajia@tencent.com
Tencent Youtu Lab
Shanghai, China

Quoc-Cuong Pham
quoc-cuong.pham@cea.fr
Université Paris-Saclay, CEA, List
Paris, France

Ralph Ewerth
ralph.ewerth@tib.eu
L3S Research Center Leibniz University Hannover, & TIB - Leibniz Information Center for Science and Technology
Hannover, Germany

Ran Song
ransong@sdu.edu.cn
School of Control Science and Engineering, Shandong University
Jinan, China

Rengang Li
lirg@inspur.com
Inspur Electronic Information Industry Co., Ltd. State Key Laboratory of High-end Server Storage Technology
Jinan, China

Rikke Gade
rg@create.aau.dk
Aalborg University
Aalborg, Denmark

Ruben Debien
ruben.debien@rebatch.be
ReBatch
Kontich, Belgium

Runze Zhang
zhangrunze@inspur.com
Inspur Electronic Information Industry Co., Ltd. State Key Laboratory of High-end Server Storage Technology
Jinan, China

Sangrok Lee
lsrock1@yonsei.ac.kr
Graduate school of information yonsei university, & MODULABS
Seoul, Korea

Sergio Escalera
sergio.escalera.guerrero@gmail.com
Universitat de Barcelona, & Computer Vision Center, & Aalborg University
Barcelona, Spain

Shan Jiang
jiang.shan@fujitsu.com
Fujitsu Research
Kawasaki, Japan

Shigeyuki Odashima
sodashima@fujitsu.com
Fujitsu Research
Kawasaki, Japan

Shimin Chen
chenshimin1@oppo.com
OPPO Research Institute
Shenzhen , China

Shoichi Masui
masui.shoichi@fujitsu.com
Fujitsu Research
Kawasaki, Japan

Shouhong Ding
ericshding@tencent.com
Tencent Youtu Lab
Shanghai, China

Sin-wai Chan
chansinwai@cihe.edu.hk
Caritas Institute of Higher Education
Tseung Kwan O, Hong Kong, SAR

Siyu Chen
chensiyu25@meituan.com
Meituan Inc.
Beijing, China

Tallal El-Shabrawy
tallal.el-shabrawy@guc.edu.eg
German University in Cairo
New Cairo City, Egypt

Tao He
kevin.92.he@gmail.com
Tsinghua University
Beijing, China

Thomas B. Moeslund
tbm@create.aau.dk
Aalborg University
Aalborg, Denmark

Wan-Chi Siu
enwcsiu@polyu.edu.hk
Caritas Institute of Higher Education, & Hong Kong Polytechnic University
Tseung Kwan O, Hong Kong, SAR

Wei Zhang
davidzhang@sdu.edu.cn
School of Control Science and Engineering, Shandong University
Jinan, China

Wei Li
liwei19@oppo.com
OPPO Research Institute
Shenzhen , China

Xiangwei Wang
wangxiangwei@baidu.com
Department of Augmented Reality
Technology (ART), Baidu Inc
Beijing, China

Xiao Tan
tanxiao01@baidu.com
Department of Computer Vision
Technology (VIS), Baidu Inc
Beijing, China

Xiaochuan Li
lixiaochuan@inspur.com
Inspur Electronic Information
Industry Co., Ltd. State Key
Laboratory of High-end Server
Storage Technology
Jinan, China

Xiaolin Wei
weixiaolin02@meituan.com
Meituan Inc.
Beijing, China

Xiaoqing Ye
yexiaoqing@baidu.com
Department of Computer Vision
Technology (VIS), Baidu Inc
Beijing, China

Xing Liu
liuxing12@baidu.com
Department of Augmented Reality
Technology (ART), Baidu Inc
Beijing, China

Xinying Wang
xinying@mgtv.com
MGTV
Changsha, China

Yandong Guo
guoyandong@oppo.com
OPPO Research Institute
Shenzhen , China

Yaqian Zhao
zhaoyaqian@inspur.com
Inspur Electronic Information
Industry Co., Ltd. State Key
Laboratory of High-end Server
Storage Technology
Jinan, China

Yi Yu
yuyi@mgtv.com
MGTV
Changsha, China

Yingying Li
liyingying05@baidu.com
Department of Computer Vision
Technology (VIS), Baidu Inc
Beijing, China

Yue He
heyue04@baidu.com
Department of Computer Vision
Technology (VIS), Baidu Inc
Beijing, China

Yujie Zhong
zhongyujie@meituan.com
Meituan Inc.
Beijing, China

Zhenhua Guo
guozhenhua@inspur.com
Inspur Electronic Information
Industry Co., Ltd. State Key
Laboratory of High-end Server
Storage Technology
Jinan, China

Zhiheng Li
zhihengli@mail.sdu.edu.cn
School of Control Science and
Engineering, Shandong University
Jinan, China

ABSTRACT

The SoccerNet 2022 challenges were the second annual video understanding challenges organized by the SoccerNet team. In 2022, the challenges were composed of 6 vision-based tasks: (1) action spotting, focusing on retrieving action timestamps in long untrimmed videos, (2) replay grounding, focusing on retrieving the live moment of an action shown in a replay, (3) pitch localization, focusing on detecting line and goal part elements, (4) camera calibration, dedicated to retrieving the intrinsic and extrinsic camera parameters, (5) player re-identification, focusing on retrieving the same players across multiple views, and (6) multiple object tracking, focusing on tracking players and the ball through unedited video streams. Compared to last year's challenges, tasks (1-2) had their evaluation metrics redefined to consider tighter temporal accuracies, and tasks (3-6) were novel, including their underlying data and annotations.

[*] Both authors contributed equally to this research.
[†] SoccerNet Organizers.

This work is licensed under a Creative Commons Attribution International 4.0 License.

MMSports '22, October 10, 2022, Lisboa, Portugal
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9488-8/22/10.
https://doi.org/10.1145/3552437.3555703

More information on the tasks, challenges and leaderboards are available on https://www.soccer-net.org. Baselines and development kits are available on https://github.com/SoccerNet.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**.

KEYWORDS

datasets, challenges, computer vision, video understanding, neural networks, soccer

ACM Reference Format:
Silvio Giancola, Anthony Cioppa, Adrien Deliège, Floriane Magera, Vladimir Somers, Le Kang, Xin Zhou, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem, Marc Van Droogenbroeck, Abdulrahman Darwish, Adrien Maglo, Albert Clapés, Andreas Luyts, Andrei Boiarov, Artur Xarles, Astrid Orcesi, Avijit Shah, Baoyu Fan, Bharath Comandur, Chen Chen, Chen Zhang, Chen Zhao, Chengzhi Lin, Cheuk-Yiu Chan, Chun-Chuen Hui, Dengjie Li, Fan Yang, Fan Liang, Fang Da, Feng Yan, Fufu Yu, Guanshuo Wang, H. Anthony Chan, He Zhu, Hongwei Kan, Jiaming Chu, Jianming Hu, Jianyang Gu, Jin Chen, João V. B. Soares, Jonas Theiner, Jorge De Corte, José Henrique Brito, Jun Zhang, Junjie Li, Junwei Liang, Leqi Shen, Lin Ma, Lingchi Chen, Miguel Santos Marques, Mike Azatov, Nikita Kasatkin, Ning Wang, Qiong Jia, Quoc-Cuong Pham, Ralph Ewerth, Ran Song, Rengang Li, Rikke Gade, Ruben Debien, Runze Zhang, Sangrok Lee, Sergio Escalera, Shan Jiang, Shigeyuki Odashima, Shimin Chen, Shoichi Masui, Shouhong Ding, Sin-wai Chan, Siyu Chen, Tallal El-Shabrawy, Tao He, Thomas B. Moeslund, Wan-Chi Siu, Wei Zhang, Wei Li, Xiangwei Wang, Xiao Tan, Xiaochuan Li, Xiaolin Wei, Xiaoqing Ye, Xing Liu, Xinying Wang, Yandong Guo, Yaqian Zhao, Yi Yu, Yingying Li, Yue He, Yujie Zhong, Zhenhua Guo, and Zhiheng Li. 2022. SoccerNet 2022 Challenges Results. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports (MMSports '22), October 10, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3552437.3555703

1 INTRODUCTION

The topic of video understanding drew a lot of attention in computer vision research. In order to push research towards better video analysis tools in sports, the SoccerNet dataset introduces six tasks related to video understanding, which are supported by open challenges for the community. This paper presents thefi nal results of the SoccerNet 2022 challenges and gives voice to the participants who briefly present their solution.

1.1 SoccerNet dataset

Giancola et al. [14] introduced SoccerNet in 2018. The objective was to share a large-scale dataset for reproducible research in soccer video understanding and to define a new task of action spotting for the temporal localization of sports activities defined with single timestamps. Originally, the dataset contained 500 videos of complete broadcast soccer games, totalling almost 800 hours of videos from the six major European championships (Seria A, La Liga, Premier League, Ligue 1, Bundesliga, and Champion's League) from 2014 to 2017. Thefi rst annotations cover temporal timestamps of three main actions in soccer: goal, cards, and substitutions. The annotations were scrapped from websites with a one minute resolution and later manually refined to a one second precision.

Figure 1: Word cloud generated from the word occurrence on the SoccerNet website. One can spot the different tasks, baselines, sponsors, communication channels, and venues that are parts of the SoccerNet 2022 challenges.

Later, Deliège et al. [9] introduced SoccerNet-v2, which significantly increased the number of annotations for the action spotting task by completing the set with common actions in soccer such as penalties, clearances, ball out of play, etc., for a total of 110,458 actions split into 17 classes. In addition to these extended annotations, SoccerNet-v2 integrates annotations for all camera changes among 13 camera classes and three transition classes: abrupt, smooth, or logo. Finally, each camera shot is annotated by specifying if it contains a replay of an action, and if so, links it to its corresponding action timestamp during the live feed. Three tasks were proposed with this dataset: action spotting, camera shot segmentation and boundary detection, and replay grounding. Challenges were organized in 2021 for action spotting and replay grounding, the two tasks focusing on retrieving specific moments in videos, using 50 extra games with segregated annotation as challenge set.

This year, Cioppa et al. [4] introduced SoccerNet-v3, scaling up previous efforts by introducing spatial annotations and tasks. SoccerNet-v3 leveraged the redundancy of the actions from SoccerNet-v2 that were shown in both live and replay moments of the broadcast, and introduced spatial annotations for players, ball,fi eld lines and goal parts from multiple views of the same scene. The frames corresponding to live and replayed actions were manually synchronized to the same salient moment of the action, totaling 33,986 frames. Three novel tasks are defined based on those frames with spatial annotations. First, a pitch localization task that aims to recover semantic pitch elements such as thefi eld lines and

[*] Equal contributions (silvio.giancola@kaust.edu.sa, anthony.cioppa@uliege.be).

goal parts. Second, a camera calibration task aiming at estimating the intrinsic and extrinsic camera parameters, and finally, a player re-identification task focusing on retrieving the same player across multiple camera views.

As the latest release, SoccerNet-Tracking [6] introduces spatio-temporal annotations on a new set of 12 complete games captured from a single main camera. The dataset includes 200 30-seconds long clips extracted around key actions and a complete 45-minute half-time for long-term tracking with all objects annotated with bounding boxes, tracklet IDs, jersey numbers, and team tags. SoccerNet-Tracking is one of the largest multi-object tracking dataset and the largest one related to soccer, accounting for more than 3.6 M bounding boxes and more than 5,000 unique tracklets.

1.2 SoccerNet challenges

In the 2022 edition of the SoccerNet challenges, we proposed 6 vision-based tasks: (1) action spotting, focusing on retrieving action timestamps in long untrimmed videos, (2) replay grounding, focusing on retrieving the live moment of an action shown in a replay, (3) pitch localization, focusing on detecting line and goal part elements, (4) camera calibration, focusing on retrieving the intrinsic and extrinsic camera parameters, (5) player re-identification, focusing on retrieving the same players across multiple views, and (6) multiple object tracking, focusing on tracking players and the ball through unedited video streams. The data for each challenge is split in 4 sets: a training set and a validation set for training models, a public test set for benchmarking in scientific publications, and a private challenge set for ranking participants, whose annotations are kept segregated to avoid any cheating.

To help participants get started with these challenges, we provide sample code on different SoccerNet GitHub repositories (https://github.com/SoccerNet) to download the data, run task-specific baselines, and evaluate their performance.

To facilitate interactions between participants, we created a Discord server, gathering more than 300 researchers during the 2022 challenges. Furthermore, we organized several live tutorials with Q&As and published explanatory videos on YouTube (https://www.youtube.com/c/acadresearch) to further attract the interest of the community. In total, 67 teams competed on the 6 proposed tasks and submitted 637 results files. We offered prizes for the winner of each task sponsored by Sportradar (2,000 $ for tasks 1-2&5), EVS Broadcast Equipment (1,000 $ for tasks 3-4), and Baidu Research (1,000 $ for task 6).

In the following, we present a detailed analysis of each task, including its description and metric, the final leaderboard, a presentation of the best performing method by the team itself, and an analysis of the results. Each other participant team was entitled to present its method in the Appendix.

2 ACTION SPOTTING

2.1 Task description

Action spotting can be considered one of the highest level of understanding for a soccer broadcast. It consists of localizing temporally when specific actions of interest occur (*e.g.* penalty, kick-off, goal, etc.). Unlike other temporal localization tasks in video understanding (*e.g.* temporal activity localization), the actions to spot are defined with single timestamps, based on soccer rules. For example, a goal is defined as the exact timestamp the ball crosses the goal line and a corner as the precise moment the player kicks the ball from the corner of the field.

Spotting soccer actions can be the building block of several applications in soccer video understanding, such as automatic video summarization and salient moment retrieval in live broadcasts. Furthermore, in its lowest level of granularity, it can support the generation of extended statistics for players and teams.

In this year's challenge, we leveraged the videos and annotations from SoccerNet-v2 [9]. The data consists of 500 games, each of them split into two half-time videos of 45 min plus eventual extra time. The annotations amount to 110,458 actions from 17 classes, anchored with a single timestamp. In addition to these annotated data, we reserved extra 50 games for the scope of this challenge, with segregated annotations to impede any participant team to train or overfit on this set.

2.2 Metrics

We use the Average-mAP [14] metric for action spotting. A predicted action spot is considered as a true positive if it falls within a given tolerance δ of a ground-truth timestamp from the same class. The Average Precision (AP) based on PR curves is computed then averaged over the classes (mAP), after what the Average-mAP is the AUC of the mAP computed at different tolerances δ. We define the *loose Average-mAP* using the original tolerances δ ranging from 5 to 60 seconds [14]. We introduce a novel *tight Average-mAP* with stricter tolerances δ ranging from 1 to 5 seconds, to evaluate for a more precise spotting.

Moreover, we differentiate between actions that are visible in the broadcast video, versus the actions that are not directly shown. For instance, several throw-ins and indirect free-kicks are not shown in the broadcast but can still be inferred from the dynamic of a game, after a ball went out of play or after a foul occurred. Spotting unshown actions requires a more abstract level of understanding involving the learning of causality and game logic.

2.3 Leaderboard

This year, 19 teams participated to the action spotting challenge for a total of 167 submissions, with an improvement from 49.56 to 67.81 tight Average-mAP. The leaderboard reporting the top-3 perfomances may be found in Table 1.

2.4 Winner

The winners for this task are João Soares *et al.* from the Yahoo Research, USA. A summary of their method is given hereafter.

S1 - Dense Detection Anchors.
João V. B. Soares and Avijit Shah
jvbsoares@yahooinc.com, avijit.shah@yahooinc.com

Soares et al. [25] proposed an anchor-based approach, defining an anchor as a pair formed by a time instant and an action class, with time instants sampled densely. For each anchor, both a detection confidence and a fine-grained temporal displacement were

Table 1: Top-3 action spotting leaderboard, complete leaderboard available in Table 7 in the appendix. Main metric for the leaderboard and best performances in bold. Team names with a superscript have provided a summary that may be found in Appendix A.1 or in Section 2.4 for the winning team.

Participants	tight Average-mAP			loose Average-mAP		
	main	vis.	inv.	main	vis.	inv.
Yahoo Research^{S1}	**67.81**	72.84	60.17	78.05	80.61	78.05
PTS	66.73	74.84	53.21	73.62	79.16	67.42
AS&RG^{S3}	64.88	70.31	53.03	72.83	76.08	72.35
Baseline*	49.56*	54.42	45.42	74.84	78.58	71.52

inferred, with the displacement indicating exactly when an action was predicted to happen. The approach resulted in a substantial improvement to temporal precision, reaching 60.7 tight average-mAP. Specifically for the challenge, changes were introduced that led to the final 67.8 tight average-mAP on the challenge set, as detailed in a follow-up report [24]. While their method uses pre-computed features, for the challenge, two different feature types (Baidu and ResNet) were combined using a standard late fusion approach, after resampling them to the desired temporal frequency of two feature vectors per second. In addition, they applied a soft version of non-maximum suppression for post-processing, while optimizing the corresponding suppression window size.

2.5 Results

This year's challenge participants focused on improving video encoders and spotting heads. The video encoders evolved from CNN to transformers, learning spatial and/or temporal self-attention mechanisms. Some methods investigated multi-modality reasoning with additional audio encoders. The spotting heads were mostly adapted from temporal activity localization methods, with dense detection anchors and hierarchical action grouping.

It is worth noting that the leading method in tight Average-mAP (Yahoo Research) also performs best in the loose metric. However, on the subset of visible actions, the 2nd best method (PTS) outperforms the leader. We believe PTS primarily relies on visual cues while Yahoo Research's method has a deeper understanding of soccer rules, with best results on actions unshown in the broadcasts.

3 REPLAY GROUNDING

3.1 Task description

Replays of salient moments are regularly shown in broadcast soccer games to emphasis the importance of an action, visualized under a more informative angle. Being able to link replays with their corresponding actions is thus a great tool for ranking actions by their impact on the game, which may be used to generate highlights of the game.

Given a replay clip, the goal of the replay grounding task is to spot the same action during the live game. The action timestamp correspond to the ones of the action spotting task, and thus follow the same annotation format. Thus, the dataset consists of the same 500 broadcast games from the action spotting task from which all replays have been retrieved. An extra 50 games with segregated annotations compose the challenge set.

Table 2: Replay grounding leaderboard. Main metric for the leaderboard and best performances in bold. The winning team summary may be found in Section 3.4. The baseline description may be found in https://github.com/SoccerNet/sn-grounding.

Participants	tight Average-AP		loose Average-AP	
	Challenge	Test	Challenge	Test
AS&RG^{G1}	**45.33**	52.31	61.07	68.57
Baseline*	19.12*	25.55	71.90	76.00

3.2 Metrics

The replay grounding task may be viewed as retrieving a single timestamp in a long untrimmed video. Hence, the same metrics as the ones used for the action spotting challenge may be used for this task. However, unlike action spotting, replay grounding does not consider the action class in its evaluation. Hence both the tight and loose average mean-Average Precision metrics are adapted by removing the averaging over the classes. These new metrics are called the tight and loose Average-AP.

For the tight Average-AP, we consider intervals of 1 to 5 seconds with a step of 1 seconds, and for the loose Average-AP, we consider intervals of from 5 to 60 seconds with a step of 5 seconds, following the action spotting metrics.

3.3 Leaderboard

This year, a single team submitted results on the replay grounding challenge set. Their performance may be found in Table 2, alongside the baseline performance.

3.4 Winner

The winners for this task are Shimin Chen *et al.* from the OPPO Research Institute, China. A summary of their method is given hereafter.

G1 - Video Action Location.

Shimin Chen, Wei Li, Jiaming Chu, Chen Chen, Chen Zhang, and Yandong Guo
chenshimin1@oppo.com, liwei19@oppo.com,
chujiaming886@bupt.edu.cn, chenchen@oppo.com,
zhangchen4@oppo.com, guoyandong@oppo.com

In order to make full use of video information, we transform the replay grounding problem into a video action location problem. We select 120 seconds clip before replay timestamps as input clip, and we set the timestamp label as the starting second of the segment labels with 3 seconds length. In this way, the predicted live stream timestamp corresponding to replay moment is equivalent to the start position of our detected result. As for temporal action detection, we first train VideoSwinTransformer [19] to extract video features. Then, we apply a unified network Faster-TAD [2] proposed by us to get segments. To get more samples for training, we randomly synthesize positive samples. Finally, by observing the data distribution of the training data, we refine results to get the final submission. Our method reached a tight mAP of 52.31% in test of SoccerNet Challenge 2022, bringing a gain of 26.76% mAP relative to last year's top result.

3.5 Results

The baseline performance correspond to last year's winner [32]. As shown in Table 2, this year's winning method significantly improved the spotting performance for tight intervals in both the challenge and test sets. These results show that the temporal activity module has a much better localization capability compared to the baseline. However, the loose average-AP significantly drops compared to the baseline. This may be due to the fact that the winner's method also makes several other guesses with high confidence, while the baseline usually focuses on a single instant, even though it is not perfectly localized.

4 FIELD LOCALIZATION

4.1 Task description

In the context of live sports events, camera calibration has many applications. One of them is to insert graphics in augmented reality for storytelling or to enforce the rules of the game (*e.g.* drawing the offside line). The automatic calibration of a camera can be done leveraging correspondences between a known 3D representation of the scene, named a calibration pattern, and its image. In soccer, the field has a specific shape and appearance, which makes it a convenient calibration pattern. Therefore, in order to achieve camera calibration, we propose a first task consisting in the localization of the soccer field elements in the image.

Given an image, the goal of the field localization task is to detect each class of soccer field element present in the image, and also to predict the 2D points in the image representing the extremities of every soccer field element detected. The soccer field elements are the set of soccer field line or circle markings, and the three posts constituting each goal. Note that the extremity of an element is defined as either its true end, or the intersection of the object with the border of the image.

The dataset has been annotated with polylines, a sequential list of 2D points that fits any soccer field element of rectilinear or circular nature. In this task, the objective is to retrieve the first and the last element, *i.e.* the extremities, of each annotated polyline.

4.2 Metrics

As there might be some uncertainty on the true exact location of an extremity, we threshold the Euclidean distance between a predicted extremity and its corresponding annotation in order to assess its validity. This thresholding strategy allows us to frame the problem as a detection task that can be evaluated by an accuracy metric dependent on the threshold value (t). We evaluate the predictions at different threshold levels. Concretely, we define that a point x belonging to the predictions of class C is a true positive (TP) if : $x \in TP : \min_i \|x, \hat{x}_i\|_2 < t$ with $\hat{x}_i$ being the set of extremities annotated for the class C in the image. The predicted extremities that do not meet that condition are counted as false positives (FP), along with predicted extremities that do not have a matching class in the annotations. Lastly, the false negatives (FN) are the extremities present in the annotations unmatched with any prediction. We define the Accuracy of the Field localization task within a tolerance of t pixels $AF@t$ as: $AF@t = \frac{TP}{TP+FP+FN}$. The final evaluation is a weighted sum defined as $0.5\,AF@5 + 0.35\,AF@10 + 0.15\,AF@20$.

Table 3: Top-3 field localization leaderboard, complete leaderboard available in Table 8 in the appendix. Main metric for the leaderboard and best performance in bold. Team names with a superscript provided a summary that can be found in Appendix A.2, or in Section 4.4 for the winner.

Participants	AF@5	AF@10	AF@20	**Final score**
ONEDAY^{P1}	84.40	90.24	92.17	**87.61**
imgo^{P2}	74.19	84.59	87.62	79.84
2Ai-IPCA^{P3}	71.01	76.18	77.60	73.81
Baseline*	13.32	38.28	53.87	28.14*

4.3 Leaderboard

For this first edition of the field localization challenge, 12 teams competed on the challenge set, for a total of 163 submissions. The top-3 performances are reported in Table 3.

4.4 Winner

The winners for this task are Yue He *et al.* from Baidu Inc, China. A summary of their method is given hereafter.

P1 - Pitch Localization Detector (PLD).
Yue He, Xiangwei Wang, Xing Liu, Xiaoqing Ye, Yingying Li, Chen Zhao, and Xiao Tan
heyue04@baidu.com, wangxiangwei@baidu.com,
liuxing12@baidu.com, yexiaoqing@baidu.com,
liyingying05@baidu.com, zhaochen03@baidu.com,
tanxiao01@baidu.com

The task evaluation is dependent on the distance for the various class lines extremities. Besides, we observe that each line is unique, that is, there is at most one instance of a category of objects for a given image from the soccer pitch. Therefore, we treat it as an instance segmentation task at first that can correctly handle occlusions where an object is spilled into two separate regions. In this way, we build the framework of Pitch Localization Detector (PLD) with a Mask2Former [3], a state-of-the-art universal image segmentation model to identify the lines category, and a PP-YOLOv2 [16] detection model for optimizing extremities locations followed with a series of optimization strategy steps which include refinement with point results, dealing with left-right ambiguities, merging intersection points, geometry-based check, and merging output results. Therefore, our PLD method predicts the extremities of the soccer pitch elements present in each image.

4.5 Results

As can be seen in Table 3, the winner team obtains a significant performance gain compared to other teams. It can be explained by their combination of two modalities, *i.e.* soccer field element instance segmentation and extremities detection, whereas other participants relied on semantic segmentation only. Another differentiating factor between the winning team and other participants is the use of recent neural networks architectures, such as a transformer for the segmentation of soccer field elements.

5 CAMERA CALIBRATION

5.1 Task description

As previously mentioned in Section 4, the automatic calibration of broadcast cameras is a game-changer to bring augmented reality graphics into live production. The goal of the task is to retrieve intrinsic and extrinsic camera parameters based on a single frame. The pinhole camera model is imposed, with some flexibility regarding the distortion parameters of the lens. Indeed, participants can choose to provide tangential, radial and thin prism distortion.

Following the previous task, we provide a 3D model of the soccer field to allow the mapping of the extremities located in the previous task to the 3D points of the field. This 3D model is further used in the evaluation.

For this task, the annotations are the same as in the previous section, but this time we keep all the annotated points of the polylines whilst before, we selected only each polyline's extremities. We emphasize the absence of any ground-truth concerning the extrinsic and intrinsic camera parameters. The evaluation is only based on metrics measuring the reprojection error in the image.

5.2 Metrics

In order to assess the quality of a submission, we provide several metrics. First, we must take into account the fact that there are some calibration methods that will fail to provide results on certain images, which is why we introduce a "Completeness Ratio" (CR) that is the ratio of the dataset images for which the method provides camera parameters. Then the other metrics are based on the accuracy of the projection of each soccer field element in the image. Using our provided soccer field model, we sample 3D points regularly along each soccer field element, then project each point in the image using the predicted camera parameters for a specific frame. In this way we obtain a set of 2D polylines that we can compare to the annotated polylines. Given a point in the 3D world $\mathbf{X}$ that has been sampled along a soccer field element of our 3D soccer field model, we use the predicted camera parameters to derive its projection in the image $\mathbf{x}$. The projection first transforms the point $\mathbf{X}$ to the camera reference system using the predicted rotation matrix R and translation vector $\mathbf{t} : (X_c, Y_c, Z_c)^T = \begin{bmatrix} \mathrm{R} & \mathbf{t} \end{bmatrix} (X, Y, Z)^T$. Then the point is projected in the normalized image plane : $(x', y') = \left(\frac{X_c}{Z_c}, \frac{Y_c}{Z_c}\right)$, where distortion can be applied using the set of predicted distortion coefficient $r : (x_d, y_d) = \psi_r(x', y')$ where ψ_r is the function applying radial, tangential and thin prism distortion. Finally, we obtain the final pixel coordinates of $\mathbf{x}$ using the predicted focal lengths f_x and f_y as well as the principal point $(c_x, c_y) : (x, y) = (f_x x_d + c_x, f_y y_d + c_y)$.

The 2D point $\mathbf{x}$ will be part of the 2D polyline associated with the class of the soccer field element. Our idea is again to frame this evaluation as a detection of soccer field elements in the image. We define that a polyline corresponding to a soccer field element l is correctly detected if the Euclidean distance between every point belonging to the annotated polyline $\hat{l}$ and the projected polyline l is less than t pixels: $\forall \hat{x} \in \hat{l} : \|\hat{x}, l\|_2 < t$. We count each predicted soccer field element that meets this condition as true positives (TP), whilst a predicted soccer field element that is located at more than t pixels from one of the annotated points for this primitive is counted as a false positive (FP), along with projected polylines that do not appear in the annotations. The false negatives (FN) are the polylines annotated that do not have a corresponding prediction. Finally, we define the Accuracy for the Camera calibration task within a tolerance of t pixels as: $AC@t = \frac{TP}{TP+FN+FP}$. We combine, in a weighted average, several levels of $AC@t$ and we apply a trade-off between the completeness rate and this weighted average in order to produce our final evaluation metric. The idea of the trade-off is to encourage participants to focus on improving accuracy rather than robustness as the completeness ratio is increasing. This is ensured by the use of a factor containing a negative exponential of the completeness ratio: an improvement in a small completeness ratio value has a higher positive impact on the metric rather than the same improvement with already satisfying completeness rate. This yields the following final score s defined as $s = (1 - e^{-4\,\mathrm{CR}})(0.5\,AC@5 + 0.35\,AC@10 + 0.15\,AC@20)$.

Table 4: Top-3 camera calibration leaderboard, complete leaderboard available in Table 9 in the appendix. Main metric for the leaderboard and best performance in bold. Team names with a superscript provided a summary that can be found in Appendix A.3, or in Section 5.4 for the winner.

Participants	AC@5	AC@10	AC@20	CR	**Final** s
achengmao^{C1}	82.38	94.80	96.33	72.61	**83.96**
L3S^{C2}	57.83	81.42	90.74	69.32	66.58
MikeAzatov^{C3}	62.25	84.32	90.56	56.41	66.45
Baseline*	12.94	29.14	43.48	58.95	21.00*

5.3 Leaderboard

For this first edition of the camera calibration challenge, 6 teams competed on the challenge set, for a total of 63 submissions. The top-3 performances are reported in Table 4.

5.4 Winner

The winners for this task are Xiangwei Wang *et al.* from Baidu Inc, China. A summary of their method is given hereafter.

C1 - Achengmao.

Xiangwei Wang, Xing Liu, Yue He, Xiaoqing Ye, Yingying Li, Chen Zhao, and Xiao Tan
wangxiangwei@baidu.com, liuxing12@baidu.com, heyue04@baidu.com, yexiaoqing@baidu.com, liyingying05@baidu.com, zhaochen03@baidu.com, tanxiao01@baidu.com

We address the problem of camera calibration for soccer videos. Given a frame extracted from a video, we detect and segment the elements (*e.g.*, lines, conics) of the pitch. We compute five types of landmark, which are line-line intersection, conic-line intersection, field center, vanishing point, and points at curves based on the detection and segmentation results. To ensure accurate landmarks, we: (1) resolve ambiguities caused by the symmetric nature of soccer field, (2) prevent each pair of lines from incorrectly splitting into two from a whole; (3) reject incorrect conic-line intersections. We propose three solvers to estimate the homograph for calibration in parallel. They are all points solver, RANSAC solver w/ and w/o coordinate perturbation. We determine the winner solver with the minimum re-projection error and conduct additional optimizations on it to obtain the optimal result of our method. The proposed method have achieved the first place in SoccerNet 2022 calibration competition.

5.5 Results

Since the algorithm provided for the previous task is used to solve the camera calibration problem, there is a strong dependency between the results obtained on the previous task and those achievable for the current task. It is therefore not surprising that with such a lead in the detection of football field features, the best camera calibration method is that of the best team on the previous task. In a later edition of this challenge, we will consider further disentanglement between the two tasks, in order to evaluate solely the calibration method without implicitly also evaluating the underlying semantic feature detection.

6 PLAYER RE-IDENTIFICATION

6.1 Task description

Person re-identification [29], or simply ReID, is a person retrieval task which aims at matching an image of a person-of-interest, called the *query*, with other person images within a large database, called the *gallery*, captured from various camera viewpoints. ReID has important applications in smart cities, video-surveillance and sport analytics, where it is used to perform person retrieval or tracking.

The goal of the SoccerNet ReID task is to re-identify players and referees across multiple camera viewpoints for a given action at a specific time instant during a soccer game. Our SoccerNet re-identification dataset is composed of 340,993 players thumbnails extracted from image frames of broadcast videos from 400 soccer games within 6 major leagues.

Compared to traditional street surveillance type re-identification dataset, the SoccerNet-v3 ReID dataset is particularly challenging because soccer players from the same team have very similar appearance, which makes it hard to tell them apart. On the other hand, each identity has a few amount of samples, which makes the model harder to train. Finally, there is a big diversity within samples of the dataset in terms of image resolution.

6.2 Metrics

We use two standard retrieval evaluation metrics to compare different ReID models: the cumulative matching characteristics (CMC) [27] at Rank-1 and the mean average precision [30] (mAP). Participants to the SoccerNet ReID challenge have been ranked according to their mAP score on the challenge set.

6.3 Leaderboard

For this first edition of the player ReID challenge, 13 teams competed on the challenge set, for a total of 123 submissions. Their top-3 performances are reported in Table 5.

Table 5: Top-3 leaderboard for the ReID task, complete leaderboard available in Table 10 in the appendix. Main metric for the leaderboard and best performance in bold. Team names with a superscript have provided a summary that can be found in the appendix, or in the next section for the winner.

Participants	**mAP**	R-1
Inspur^{T1}	**91.68**	**89.41**
MGSoccer^{T2}	91.48	89.21
MTVACV^{T3}	90.11	87.04
Baseline	59.11	48.41

6.4 Winner

The winners for this task are Rengang Li *et al.* from Inspur, China. A summary of their method is given hereafter.

R1 - Optimized Strategy for Player Re-identification.
Rengang Li, Yaqian Zhao, Hongwei Kan, Zhenhua Guo, Baoyu Fan, Runze Zhang, Xiaochuan Li
lirg@inspur.com, zhaoyaqian@inspur.com, kanhongwei@inspur.com, guozhenhua@inspur.com, fanbaoyu@inspur.com, zhangrunze@inspur.com, lixiaochuan@inspur.com

We analyzed that the main challenges are the sample imbalance and unrobustness mainly caused by multi-input resolution. We removed the ids whose images are less 3 and employed the focal loss function to solve sample imbalance. We experimented different combination of ReID network module to choose the best representation ability and selected ResNeSt269, combination of Arc-Softmax and Cos-Softmax. We used Auto-Aug, Color Jittering and Random Erase and all of the data augmentation uses the probability of 0.5. After we optimized the best hyper-parameters of single model, we paid more attention to the common person ReID tricks, such as multi-input resolution model fusion, add test phase dataset as well as unsupervised domain adaptation.

6.5 Results

Participants came up with various innovative ideas and have achieved outstanding performances despite the difficulty of the task. We list here some of the keys ideas shared by participants. **(i)** Apply some pre-processing by removing identities with too few samples in the training set. **(ii)** Design a handcrafted training batch sampling strategies based on additional SoccerNet ReID dataset labels, such as action id and game id. **(iii)** Add standard data augmentation strategies: Horizontal Flip, Random Erasing [31], Random Cropping, AutoAugment [8], AugMix [15], Color Jitter, ... **(iv)** Use a strong baseline such as the TransReID-SSL [21] baseline with ViT [11] backbone and unsupervised pre-training on LUPerson [13] dataset. **(v)** Use specific metric learning loss functions: the Focal Loss [18], a custom Centroid loss, the InfoNCE loss [26], the Arcface loss [10], ... **(vi)** Inference time fine-tuning with unsupervised domain adaptation on the challenge set to further increase final performance. **(vii)** Combine multiple models predictions at inference to compute final distance metric.

Table 6: Top-3 tracking leaderboard, complete leaderboard available in Table 11 in the appendix. Main metric for the leaderboard and best performances in bold. Team names with a superscript have provided a summary that may be found in Appendix A.5, or in Section 7.4 for the winning team.

Participants	**HOTA**	DetA	AssA
Kalisteo^{T1}	**93.64**	99.56	88.06
CBIOUT (CB-IoU)^{T2}	93.25	99.76	87.15
tactica^{T3}	93.17	99.85	86.94
Baseline*	70.89*	82.97	60.68

7 MULTIPLE PLAYER TRACKING

7.1 Task description

Tracking is a hot topic of research, which is far from being solved. In sports, tracking algorithms enable many interesting applications. They can be used to generate player specific highlights and statistics, or be leveraged for holistic video understanding [5].

As defined in the SoccerNet-Tracking dataset, the tracking task is split in two steps: (1) detecting the objects to track and (2) associating the bounding boxes over time to create the tracklets. For this year's challenge, the participants had access to 150 30−seconds clips recorded only from a single camera, with all ground-truth bounding boxes provided. The goal of the task is therefore to associate these bounding boxes over time to create the final tracklets. The complete tracking task, including both detection and association, will be part of the next edition of the SoccerNet challenges.

Compared to most tracking datasets, SoccerNet-Tracking includes several challenges such as long-term re-identification, *i.e.* if an object leaves the frame and comes back, it needs to be associated to the same tracklet. Since most players in the same team have very similar appearances, the re-identification is challenging.

7.2 Metrics

Following the recent work of Luiten *et al.* [20], we use the HOTA metric to rank the participants. This metric may be decomposed into a detection accuracy (DetA) and an association accuracy (AssA). Compared to the previous common MOTA metric, it is much more balanced for the evaluation of detection and association capabilities.

7.3 Leaderboard

For this first edition of the challenge, 12 teams competed on the challenge set, for a total of 103 submissions. The performance of the top-3 teams may be found in Table 6.

7.4 Winner

The winners for this task are Adrien Maglo *et al.* from Université Paris-Saclay, CEA, List, France. A summary of their method is given hereafter.

T1 - TrackMerger.

Adrien Maglo, Astrid Orcesi, and Quoc-Cuong Pham
adrien.maglo@cea.fr, astrid.orcesi@cea.fr, quoc-cuong.pham@cea.fr

The first step of TrackMerger generates player tracks by sequentially processing the video frames. The current frame detections are matched to existing tracks bounding boxes with a Hungarian assignment algorithm using two criteria, the Intersection-Over-Union between bounding boxes and the distance between their center. Only small bounding boxes can extend the ball track. Generated tracks are of good quality as long as the player stay visible. To be able to recognize players who exit and later re-enter the camera field of view, the second step fine-tunes a re-identification network with a triplet loss formulation. Positive samples are extracted from the same track as the anchor while negative samples come from concomitant tracks. The third step merges the tracks according to the distance between their re-identification vectors. It also prevents the duplication of a player's identity in the same frame and teleportation in successive frames.

7.5 Results

Similar to the ReID challenge, participants achieved outstanding performances on this task. Most participants used the standard two phases approach to address long-term tracking: **(i) Short tracklets**: Build short tracklets using an online tracking method relying mainly on spatio-temporal features, such as IoU/BIoU with Kalman filter. **(ii) Long tracks**: Connect these short tracklets in an offline manner using appearance features, in order to solve heavy occlusions or players going out of the camera view. These appearance features are obtained using pre-trained re-identification models, that are fine-tuned on the training set or that are learned at inference in a self-supervised way on the short tracklets generated in the previous step. Some participants used additional priors to further improve HOTA performance, such as physical constraints on ball size or players maximum speed.

8 CONCLUSION

This paper summarizes the outcome of the SoccerNet 2022 challenges. In total, we present the results on six tasks: action spotting, replay grounding, pitch localization, camera calibration, player re-identification, and player tracking. These challenges provide a comprehensive overview of current state-of-the-art methods within each computer vision task. For each challenge, participants were able to significantly improve the performance of our proposed baselines, introducing new architectures, engineering tricks, and soccer-centric priors. Yet, much more effort is still needed to solve the proposed tasks for practical applications. In future editions, we expect to enrich the current sets of annotations and propose further tasks related to video understanding in soccer, introducing multiple modalities, higher level of granularity, and summarization tasks.

ACKNOWLEDGMENTS

This work was supported by the Service Public de Wallonie (SPW) Recherche under the DeepSport project and Grant N^{o}. 2010235 (ARIAC by https://DigitalWallonia4.ai), the FRIA, the FNRS, and KAUST Office of Sponsored Research through the Visual Computing Center funding.

REFERENCES

[1] Andrei Boiarov and Eduard Tyantov. 2019. Large Scale Landmark Recognition via Deep Metric Learning. In *ACM Int. Conf. Inf. Knowl. Manag.* ACM, Beijing China, 169–178. https://doi.org/10.1145/3357384.3357956

[2] Shimin Chen, Chen Chen, Wei Li, Xunqiang Tao, and Yandong Guo. 2022. Faster-TAD: Towards Temporal Action Detection with Proposal Generation and Classification in a Unified Network. *arXiv* abs/2204.02674 (2022), 16 pages. arXiv:2204.02674

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* New Orleans, LA, USA, 1290–1299.

[4] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. 2022. Scaling up SoccerNet with multi-view spatial localization and re-identification. *Scientific Data* 9, 1 (June 2022), 1–9. https://doi.org/10.1038/s41597-022-01469-1

[5] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Floriane Magera, Olivier Barnich, Bernard Ghanem, and Marc Van Droogenbroeck. 2021. Camera Calibration and Player Localization in SoccerNet-v2 and Investigation of their Representations for Action Spotting. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Work. (CVPRW), CVsports*. Inst. Elect. and Electron. Engineers (IEEE), Nashville, TN, USA, 4537–4546. https://doi.org/10.1109/CVPRW53098.2021.00511

[6] Anthony Cioppa, Silvio Giancola, Adrien Deliège, Le Kang, Xin Zhou, Cheng Zhiyu, Bernard Ghanem, and Marc Van Droogenbroeck. 2022. SoccerNet-Tracking: Multiple Object Tracking Dataset and Benchmark in Soccer Videos. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Work. (CVPRW), CVsports*. Inst. Elect. and Electron. Engineers (IEEE), New Orleans, LA, USA, 3491–3502.

[7] Bharath Comandur. 2022. Sports Re-ID: Improving Re-Identification Of Players In Broadcast Videos Of Team Sports. *arXiv* abs/2206.02373 (2022), 11 pages. arXiv:2206.02373

[8] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. 2018. AutoAugment: Learning Augmentation Policies from Data. *arXiv* abs/1805.09501 (2018), 14 pages. arXiv:1805.09501

[9] Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. 2021. SoccerNet-v2: A Dataset and Benchmarks for Holistic Understanding of Broadcast Soccer Videos. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Work. (CVPRW), CVsports*. Inst. Elect. and Electron. Engineers (IEEE), Nashville, TN, USA, 4508–4519. https://doi.org/10.1109/CVPRW53098.2021.00508

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *IEEE/CVF Conf. Comput. Vis. and Pattern Recogn. (CVPR)*. Inst. Elect. and Electron. Engineers (IEEE), Long Beach, CA, USA, 4685–4694. https://doi.org/10.1109/cvpr.2019.00482

[11] 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. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* abs/2010.11929 (2021), 22 pages. arXiv:2010.11929

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast Networks for Video Recognition. In *Int. Conf. Comput. Vis.* Inst. Elect. and Electron. Engineers (IEEE), Seoul, South Korea, 6201–6210. https://doi.org/10.1109/iccv.2019.00630

[13] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. 2021. Unsupervised Pre-training for Person Reidentification. In *IEEE/CVF Conf. Comput. Vis. and Pattern Recogn. (CVPR)*. Inst. Elect. and Electron. Engineers (IEEE), Nashville, TN, USA, 14745–14754. https://doi.org/10.1109/cvpr46437.2021.01451

[14] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. 2018. SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Work. (CVPRW), CVsports*. Inst. Elect. and Electron. Engineers (IEEE), Salt Lake City, UT, USA, 1711–1721. https://doi.org/10.1109/CVPRW.2018.00223

[15] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2019. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. *arXiv* abs/1912.02781 (2019), 15 pages. arXiv:1912.02781

[16] Xin Huang, Xinxin Wang, Wenyu Lv, Xiaying Bai, Xiang Long, Kaipeng Deng, Qingqing Dang, Shumin Han, Qiwen Liu, Xiaoguang Hu, et al. 2021. PP-YOLOv2: A practical object detector. *arXiv* abs/2104.10419 (2021), 7 pages. arXiv:2104.10419

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., Virtual conference, 18661–18673.

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 2 (Feb. 2020), 318–327. https://doi.org/10.1109/tpami.2018.2858826

[19] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2021. Video Swin Transformer. *arXiv* abs/2106.13230 (2021), 12 pages. arXiv:2106.13230

[20] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. 2021. HOTA: A higher order metric for evaluating multi-object tracking. *Int. J. Comput. Vis.* 129, 2 (Oct. 2021), 548–578. https://doi.org/10.1007/s11263-020-01375-2

[21] Hao Luo, Pichao Wang, Yi Xu, Feng Ding, Yanxin Zhou, Fan Wang, Hao Li, and Rong Jin. 2021. Self-Supervised Pre-Training for Transformer-Based Person Re-Identification. *arXiv* abs/2111.12084 (2021), 15 pages. arXiv:2111.12084

[22] Haowen Luo, Pichao Wang, Yi Xu, Feng Ding, Yanxin Zhou, Fan Wang, Hao Li, and Rong Jin. 2021. Self-Supervised Pre-Training for Transformer-Based Person Re-Identification. *arXiv* abs/2111.12084 (2021), 15 pages. arXiv:2111.12084

[23] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *IEEE/CVF Conf. Comput. Vis. and Pattern Recogn. (CVPR)*. Inst. Elect. and Electron. Engineers (IEEE), Boston, MA, USA, 815–823. https://doi.org/10.1109/cvpr.2015.7298682

[24] João V. B. Soares and Avijit Shah. 2022. Action Spotting using Dense Detection Anchors Revisited: Submission to the SoccerNet Challenge 2022. *arXiv* abs/2206.07846 (2022), 3 pages. arXiv:2206.07846

[25] João V. B. Soares, Avijit Shah, and Topojoy Biswas. 2022. Temporally Precise Action Spotting in Soccer Videos Using Dense Detection Anchors. *arXiv* abs/2205.10450 (2022), 5 pages. arXiv:2205.10450

[26] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv* abs/1807.03748 (2018), 13 pages. arXiv:1807.03748

[27] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter Tu. 2007. Shape and Appearance Context Modeling. In *Int. Conf. Comput. Vis.* Inst. Elect. and Electron. Engineers (IEEE), Rio de Janeiro, Brazil, 1–8. https://doi.org/10.1109/iccv.2007.4409019

[28] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiplefine-tuned models improves accuracy without increasing inference time. *arXiv* abs/2203.05482 (2022), 34 pages. arXiv:2203.05482

[29] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. 2022. Deep Learning for Person Re-Identification: A Survey and Outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 6 (June 2022), 2872–2893. https://doi.org/10.1109/tpami.2021.3054775

[30] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable Person Re-identification: A Benchmark. In *Int. Conf. Comput. Vis.* Inst. Elect. and Electron. Engineers (IEEE), Santiago, Chile, 1116–1124. https://doi.org/10.1109/iccv.2015.133

[31] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random Erasing Data Augmentation. In *AAAI*, Vol. 34. Association for the Advancement of Artificial Intelligence, New York, USA, 13001–13008. https://doi.org/10.1609/aaai.v34i07.7000

[32] Xin Zhou, Le Kang, Zhiyu Cheng, Bo He, and Jingyu Xin. 2021. Feature Combination Meets Attention: Baidu Soccer Embeddings and Transformer based Temporal Detection. *arXiv* abs/2106.14447 (2021), 7 pages. arXiv:2106.14447