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
Qingdao, 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
Beijing, 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
Beijing, 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.

MMSports '22, October 10, 2022, Lisboa, Portugal

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, 17 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 the final 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. The first 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, field 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 the field 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 Re-identification. 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 multiple fine-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

Table 7: Action spotting leaderboard. 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. The baseline description may be found in https://github.com/SoccerNet/sn-spotting.

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
mt_sdu_action^{S4}	62.26	67.48	45.04	69.86	73.81	59.15
Rkrystal	61.84	67.39	48.71	74.75	78.29	69.02
arturxe^{S6}	60.56	65.75	53.00	71.72	75.15	69.91
cihe^{S7}	59.97	64.51	53.80	72.95	76.29	71.95
GUC^{S8}	58.71	63.70	51.86	70.49	73.46	70.11
abcdefg	56.07	62.97	46.51	67.88	72.54	66.37
intro- and inter	53.97	60.04	47.52	67.75	71.16	70.12
memory	53.03	57.94	43.16	67.15	69.20	68.28
stargazer^{S12}	52.04	60.18	32.06	60.86	66.64	48.46
heaven	51.85	59.85	31.62	60.88	66.67	48.45
lczazu	49.56	56.82	31.60	60.86	66.56	48.51
Baseline*	49.56*	54.42	45.42	74.84	78.58	71.52
zqing	47.54	51.75	41.65	66.66	69.06	67.17
welkin	42.74	49.91	20.67	50.90	56.48	35.38
DUT	40.65	43.87	43.10	68.40	71.68	68.53
sshinde5	36.71	39.33	21.26	51.36	55.29	35.34
SIT^{S20}	21.60	26.55	16.83	29.92	34.92	25.22

A APPENDIX

In this appendix, the participants provide a short summary of their methods. Only teams who provided a technical report at the end of the challenge that has been peer-reviewed by the organizers were able to submit a summary. This is to ensure that the presented methods followed the challenge rules.

A.1 Action Spotting

The complete leaderboard is provided in Table 7.

S3 - Temporal Action Detection.
Wei Li, Shimin Chen, Jianyang Gu, Chen Chen, Ning Wang, and Yandong Guo
liwei19@oppo.com, chenshimin1@oppo.com, gu_jianyang@zju.edu.cn, chenchen@oppo.com, wangning12@mail.ecust.edu.cn, guoyandong@oppo.com

We apply temporal action detection (TAD) method following Faster-TAD [2] to the action spotting task. We use a 16 seconds sliding window with 8 seconds stride to detect a 4 seconds soccer action. The starting position of an action is the officially provided timestamp. We first generate multiple VideoSwinTransformer [19] feature extractors pre-trained on different customized label styles which are designed to be sensitive to action boundary. Then, we generate temporal proposals and semantic labels in a unified network following Faster-TAD, learning useful context for each proposal. We involve cross entropy loss and triplet loss [23] for explicit constraints of embedded feature distributions. After collecting all TAD results for each sliding window, we use NMS to generate final results of a game. With some ensemble methods, we finally reached 68.46% and 64.88% tight Average-mAP on SoccerNet test and challenge sets, respectively.

S4 - Group-wise Multi-scale action Detector (GMDet).
Zhiheng Li, Wei Zhang, Yujie Zhong, Dengjie Li, Feng Yan, Xiaolin Wei, Ran Song, and Lin Ma
zhihengli@mail.sdu.edu.cn, davidzhang@sdu.edu.cn, zhongyujie@meituan.com, lidengjie@meituan.com, yanfeng05@meituan.com, weixiaolin02@meituan.com, ransong@sdu.edu.cn, linma@alumni.cuhk.net

The Group-wise Multi-scale action Detector (GMDet) aims at detecting/spotting actions in untrimmed videos. It consists of an ensembled feature encoder (proposed by Baidu, the 1st place last year), a multi-scale transformer encoder and a CNN-based decoder (which predicts the action classes and temporal locations). In order to handle various actions with very different properties, GMDet divides actions into three action groups, where similar actions belong to the same group. The group-wise operation enables a specific network architecture and a tailored training strategy for each group. First, we adopt decoders of different depth for the three action groups. Second, Mix-up and adaptive loss weights are leveraged for the action group with few training examples. At inference, one or more models (in a way of model ensemble) are used to predict action instances for each group independently. The proposed GMDet demonstrates superior performance on the action spotting task.

S6 - Hierarchical Multimodal Transformers for Action Spotting (HMTAS).
Artur Xarles, Sergio Escalera, Albert Clapés Thomas B. Moeslund, and Rikke Gade
arturxe@gmail.com, sergio.escalera.guerrero@gmail.com, alcl@create.aau.dk, tbm@create.aau.dk, rg@create.aau.dk

Hierarchical Multimodal Transformers for Action Spotting (HMTAS) is built on 6 different precomputed embeddings, the 5 visual ones from TPN, GTA, VTN, irCSN, and I3D-Slow backbones (provided by the winners of the action spotting challenge at CVPR 2021 ActivityNet workshop [32]), plus the audio ones from a VGGish fine-tuned on SoccerNet-v2. Differently from [32], which concatenated the visual embeddings channel-wise, the sequences of the different embeddings accounting for 3-seconds-long temporal windows are here independently evolved through their own Transformer encoder. The evolved embeddings are then temporally max-pooled and concatenated into a 5+1 sequence of global embeddings that is fed to a multimodal Transformer encoder producing the final multilabel classification into the 18 different spotting classes. HMTAS minimizes the Negative Log-Likelihood Loss (NLLL) of the final classification, but also the auxiliary NLLLs supervising the intermediate classifications attempted by the unimodal encoders. During inference, Non-Maximum Suppression is used to reduce the number of spotting candidates.

S7 - Mixed Spatial and Temporal Attention for Soccer Game Action Spotting (CIHE).
Cheuk-Yiu Chan, Chun-Chuen Hui, Wan-Chi Siu, Sin-wai Chan, and H. Anthony Chan

Table 8: Field localization leaderboard. 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. The baseline description may be found in https://github.com/SoccerNet/sn-calibration.

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
channings	66.87	78.19	81.65	73.05
test222	61.8	77.42	81.45	70.21
eidos.ai	63.42	75.94	78.32	70.04
Mike Azatov^{P7}	62.09	73.47	76.79	68.28
goahead	50.67	82.87	92.54	68.28
L3S	20.57	45.44	64.38	35.85
test26	15.91	44.02	61.29	32.56
tactica	15.90	44.0	61.24	32.54
B1	14.65	40.41	56.47	29.94
Baseline*	13.32	38.28	53.87	28.14*

cy3chan@cihe.edu.hk, cchui@cihe.edu.hk, enwcsiu@polyu.edu.hk, chansinwai@cihe.edu.hk, hhchan@cihe.edu.hk

A novel approach on mixing spatial and temporal attention is proposed. Our model consists of two similar stages, with the 2nd stage designed for improving spatio-temporal representation capability by enlarging receptive fields. Each stage has initially a Transformer network (Temporal-Grouped Attention (TGA)) with different sets of parameters to be trained for different groups of channels, and another transformer with only one set of parameters formed by grouping all channels together for the advantage of extracting temporal domain features. Each transformer contains a concurrent "Temporal- Grouped Local Attention" network and a "Temporal-Grouped Global Attention' network. A Selective Feature Aggregation (SFA) structure is proposed to select the final weights between the two attention networks intelligently which forms an additional attention-based gate controlling mechanism, and allows a further cooperation of the spatial and temporal features for making good decisions on soccer game action spotting. Our network has achieved a tight Average mAP of 60.51%. *Acknowledgement:* Work supported by Hong Kong UGC Grant: UGC/IDS(C)11/E01/20) via CIHE.

S8 - STE.

Abdulrahman Darwish and Tallal El-Shabrawy
abdulrahman.darwish@guc.edu.eg, tallal.el-shabrawy@guc.edu.eg

STE is a proposed deep model for action spotting in soccer videos. The model reads the pre-extracted Baidu soccer embeddings, generated from SoccerNet-v2 dataset. It encodes the spatial features then the temporal features at different scales. Its architecture consists of 3 blocks: the spatial encoder, the temporal encoder then the prediction block. The first block extracts the spatial semantics through 1D max-pooling, a fully connected layer and 1D convolutional layer. The temporal encoder then extracts features across frames by applying 3 1D convolutional layers with 1D max-pooling after the first layer. The last block predicts two outputs, the first prediction is the class event that occurred in the input window. Two fully connected layers are used to map the extracted semantics to predicted class. The second output is the time frame where the event occurred. In a parallel branch, 2 fully connected layers predict the timestamp of the event.

S12 - Stargazer.

He Zhu, Junwei Liang, Chengzhi Lin, and Jun Zhang, Jianming Hu
zhuh20@mails.tsinghua.edu.cn, junweiliang1114@gmail.com, linchzh3@mail2.sysu.edu.cn, bobbyjzhang@tencent.com, hujm@mail.tsinghua.edu.cn

We propose Stargazer, a transformer-based system which can efficiently exploit the rich temporal features about the soccer action information. We first extract a certain number of sequential frames from the video as clips, and then put them into the action recognition module pretrained on Kinetics. It is an improved network based on multi-scale vision transformer, and it can learn a hierarchy of robust representations. It processes the proposal clips and classifies them into one of the soccer actions or background. We also use spatial and temporal multi-crop data augmentation to facilitate the training. Finally, we utilize the action recognition module and overlapped sliding window strategy to extract the features of the video and sent them to model of NetVLAD++ to get the result.

S20 - Simultaneous Supervised Contrastive Learning for Spatio-Temporal Action Recognition.

Andrei Boiarov and Nikita Kasatkin
andrei.boiarov@sit.team, nk@sit.team

Our approach is a combination of a spatio-temporal backbone and metric learning for the action recognition problem. The model consists of two networks: backbone (frozen pretrined on Kinetics-400 SlowFast-R101 [12]) and head. Head network: L_2 normalization, two fully connected (FC) blocks (Dropout, FC layer with hidden size 2048, BatchNorm and ELU activation), embedding layer (output size 512) and classification FC layer. Dataset videos have been split as clips of 2 seconds duration. We propose Simultaneous Supervised Contrastive (Equation 3) loss for simultaneous training embedding and classifier in a [1] fashion. In contrast to the standard SupCon (Equation 2) [17], function (Equation 3) is calculated only for positive classes.

$$\mathcal{L}_{CE} = -\sum_{i \in I} \log \frac{\exp(W_{\mathbf{y}_i}^{\top} \mathbf{z}_i)}{\sum_{j=1}^{n+1} \exp(W_j^{\top} \mathbf{z}_i)} \tag{1}$$

$$\mathcal{L}_{SC} = \sum_{\substack{i \in I \\ \mathbf{y}_i < n+1}} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \mathbf{z}_a / \tau)} \tag{2}$$

$$\mathcal{L} = \mathcal{L}_{CE} + \frac{\lambda}{2} \mathcal{L}_{SC}, \tag{3}$$

n — positive classes, $n+1$ — negative class, I — batch, $A(i) = I \setminus \{i\}$, $P(i)$ — one class indices, $\mathbf{z}_i$ — ith embedding, $\mathbf{y}_i$ — correct label, W_j — jth column of the classifier weights matrix, τ — temperature.

A.2 Field Localization

The complete leaderboard is provided in Table 8.

Table 9: Camera calibration leaderboard. 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. The baseline description may be found in https://github.com/SoccerNet/sn-calibration.

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
2AI-IPCA^{C4}	51.31	65.47	71.43	60.57	54.03
imgo	45.11	61.38	68.63	77.98	51.93
test26	13.05	28.49	41.66	68.90	21.30*
Baseline*	12.94	29.14	43.48	58.95	21.00*

P2 - MGTV.
Lingchi Chen, Yi Yu, Xinying Wang, and Jin Chen
lingchi@mgtv.com, yuyi@mgtv.com,
xinying@mgtv.com, chenjing@mgtv.com

This solution comes from MGTV, we are a video media company from China. There is three main parts to introduce our method: firstly, We modify the LaneNet network structure for semantic segmentation, which include a segmentation branch and instance branch. The second is About data augmentation, we have tried some enhancement methods, and the experiments show that the following three types of operations are most helpful for the prediction effect: left and right mirroring, randomcrop, and force the label of the image behind the goal to the left. The above three operations have brought about an improvement of nearly 8% in the AF@5 score. Finally, the postprocessing of taking endpoints from line segments, also includes operations such as extending, correcting categories, and merging line under the guidance of the soccerpitch template. The line correction operations have brought about an improvement of nearly 12% in the AF@5 score.

P3 - Hierarchical Line Extremity Segmentation Unet.
José Henrique Brito and Miguel Santos Marques
jbrito@ipca.pt, a18888@alunos.ipca.pt

Our method directly infers line extremities using a CNN based on Unet. The input and outputs are 288×512. It produces three outputs, the first with 3 channels (Background, Floor Lines, Goal Lines), the second 26 channels (line classes), the third 25 channels, (class extremities except "Circle central"). Output 1 has a Softmax activation, and the others use Sigmoid. Outputs are hierarchical, *i.e.*, output 2 takes as input the concatenation of the last backbone feature map and the output 1 feature map. Output 3 takes as input the concatenation of the last backbone feature map and the output 2 feature map. The global loss is a weighted average the 3 output entropy losses. Line masks for training were simply drawn, extremity masks were generated with Gaussian heatmaps centered on extremity coordinates with standard deviation of 1. Inferred extremity masks are thresholded and the highest pair of peaks are used as predictions.

P7 - Two-step procedure.
Mike Azatov
mazatov@gmail.com

The approach follows a two-step procedure where in the first step I find the pitch element segmentation of the image and in the second part, I find the locations of the extremities. The segmentation is improved by using a superior segmentation model coupled with splitting the dataset into a few sub-datasets based on the camera view. Multiple camera views challenge segmentation algorithms as different pitch elements can appear in different positions with respect to each other depending on the view. The extremity localization is done by fitting lines and ellipses to individual pitch elements and finding their intersection points with each other as well as with the image borders.

A.3 Camera Calibration

The complete leaderboard is provided in Table 9.

C2 - Camera Calibration for Broadcast Soccer Videos.
Jonas Theiner and Ralph Ewerth
theiner@l3s.de, ralph.ewerth@tib.eu

We propose the differentiable *segment reprojection loss* function that aims to approximate the 6-DoF camera pose (position and orientation) and focal length given segment correspondences (lines and point clouds) between image and calibration object (3D soccer pitch model). Initialized with camera parameters representing *central* views from multiple camera locations, the *segment reprojection loss* induced by the camera parameters is iteratively minimized via gradient descent. Estimates that are likely invalid (*e.g.*, erroneous pitch element localization, local minima, out of camera distribution) are automatically discarded by thresholding the loss. For the segmentation of pitch elements and subsequent selection of relevant pixels, we apply a retrained baseline segmentation model and randomly sample additional points to get more stable gradients.

C3 - Self-Assessing Algorithm.
Mike Azatov
mazatov@gmail.com

The core idea in the camera calibration algorithm is to combine the location of extremities with pitch segmentation results to create a self-assessing algorithm that can accurately predict how well the camera is calibrated and decide on whether further improvements are necessary. The point and line correspondences provided a candidate solution for camera parameters. The self-assessment mechanism decides whether the results are sufficiently strong, and proceeds to optimize the camera calibration until it we get satisfactory results

C4 - 2Ai-IPCA.
José Henrique Brito and Miguel Santos Marques
jbrito@ipca.pt and a18888@alunos.ipca.pt

Our camera calibration method uses the line extremity estimates produced by our pitch localization method and feeds those point coordinates as input to the baseline method for camera calibration provided by the SoccerNet team. The baseline camera calibration method then computes the homography between the detected 2D markings on the image and the known 3D measurements of the markings in a soccer field, and subsequently computes the camera calibration parameters.

Table 10: Re-identification leaderboard. Main metric for the leaderboard and best performance in bold. Team names with a superscript have provided a summary that can be found in Appendix A.4, or in Section 6.4 for the winner. The baseline description may be found in https://github.com/SoccerNet/sn-reid.

Participants	**mAP**	R-1	Participants	**mAP**	R-1
Inspur^{R1}	**91.68**	89.41	tianchao	87.39	84.25
MGSoccer^{R2}	91.48	89.21	126_187	86.96	83.04
MTVACV^{R3}	90.11	87.04	Blackghost	78.25	71.50
gunners^{R4}	89.47	86.42	1234567	65.98	58.40
MIG^{R5}	89.44	86.11	rasengan	64.79	54.27
MMLab	89.01	85.68	baba	60.32	48.91
ReBatch^{R7}	88.36	84.82	Baseline*	59.11*	48.41

A.4 Player Re-Identification

The complete leaderboard is provided in Table 10.

R2 - MGSoccer-ViT.
Guanshuo Wang, Junjie Li, Fufu Yu, Qiong Jia, Shouhong Ding
mediswang@tencent.com, serenitycapo@gmail.com,
fufuyu@tencent.com, boajia@tencent.com,
ericshding@tencent.com

We apply a hybrid architecture named MGSoccer-ViT to solve the player re-identification task. The player image is respectively represented by CNN and ViT backbones. CNN backbone is designed according to the Multiple Granularity Network, which represents global and local features by independent branches with different levels of partitions. ViT backbone is referred to the TransReID architecture, which divides the images into equal patches as input token sequences of Transformer networks. All the backbones are initialized with self-supervised pretrained models on large-scale LUPerson dataset. ArcFace and hard triplet losses are applied during training. Towards the special data distribution of SoccerNet, we remove all the long-tail PIDs and apply an action-based sampler to sample equal numbers of PIDs in different actions. Various data augmentation methods are employed including random flipping, erasing, cropping after padding, AutoAugment and AugMix. Our model achieves 91.28% mAP on the Test set, and 91.48% on the Challenge set.

R3 - Domain-Aware Self-Supervised Pre-Training.
Siyu Chen, Dengjie Li, Yujie Zhong, Fan Liang, Xiaolin Wei, Lin Ma
chensiyu25@meituan.com, lidengjie@meituan.com,
zhongyujie@meituan.com liangfan02@meituan.com,
weixiaolin02@meituan.com, forest.linma@gmail.com

The SoccerNet ReID dataset mainly includes two difficulties: high appearance similarity of players and serious mutual occlusions between players. To alleviate these two problems, we first propose a target-domain-oriented self-supervised pre-training strategy to obtain a strong pretrained model. Namely, we refine the images of LUPerson dataset and retain those having high Catastrophic Forgetting Score (CFS) with the SoccerNet ReID dataset to form the pre-training set. DINO is adopted to train a ViT-B on the pre-training set. Second, in order to better distinguish different players, we design a visibility-and-importance-aware feature learning framework to extract discriminative features of different body parts of the players. Finally, we design a model fusion strategy, which simply averages the distance matrices of different models, to further boost the performance. Our method achieves 90.1 mAP on the SoccerNet ReID challenge dataset.

R4 - Hierarchical sampling and centroid loss to improve player re-identification.
Bharath Comandur
cjrbharath@gmail.com

Instead of random sampling, we use a novel procedure to select samples for each batch. We use a loss function with centroids to better separate the embeddings. Finally we create a model-soup [28] to further increase mAP. Our models are pretrained on only ImageNet. With these simple ideas, we achieve an mAP of 89.47 on the SoccerNet challenge set. More details can be found at [7].

R5 - MIG.
Leqi Shen and Tao He
lunarshen@gmail.com, kevin.92.he@gmail.com

We propose an improvement of the classical sampling strategy. We constitute batches by randomly sampling A actions, I identities and K images per identity, thus resulting in a batch of A × I × K images. The large number of overall actions and the randomness can prevent different identities of the same player from appearing at the same batch. We utilize non-parametric InfoNCE loss within batches. We use ViT with ICS in TransReID-SSL [22] as our backbone. Based on self-attention in Transformer, a features fusion method is proposed to generate the refined feature for an original query feature. Model ensemble is also an effective method to boost performance. Five models (different input sizes and pretrain models) with the fusion method are ensembled to generate the final distance matrix.

R7 - Person re-identification task with RESNET ensemble.
Jorge De Corte, Andreas Luyts, and Ruben Debien
jorge.decorte@rebatch.be, andreas.luyts@rebatch.be,
ruben.debien@rebatch.be

RE(s)IDnet is an ensemble of models with RESNET variants as backbone. These individual models are trained using a weighted sum of the Euclidian Triplet loss and an ID/classification loss, the latter only having a very small contribution to the overall loss. Images from the same action ID were kept together in batches to ensure that the best/hardest triplets possible were used. A cyclical learning rate schedule, starting with a warmup cycle, was used during the training process. Pre- and postprocessing have a big impact on model performance. Various augmentation techniques, such as coarse dropout, are used in the preprocessing phase. Post-processing involves changing the gallery rankings after prediction. If all models agree on the highest ranked image for a certain query, then we can move this image backwards in the ranking for all other queries.

A.5 Multiple Player Tracking

The complete leaderboard is provided in Table 11.

T2 - Cascaded Buffered IoU (BIoU).
Fan Yang, Shigeyuki Odashima, Shoichi Masui, and Shan Jiang

Table 11: Tracking leaderboard. 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. The baseline description may be found in https://github.com/SoccerNet/sn-tracking.

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
FGV	92.49	99.76	85.74
smot	91.49	99.77	83.90
tianchao	89.42	99.62	80.27
who	88.99	99.74	79.39
tomo	88.94	99.77	79.28
dk^{T9}	88.65	99.70	78.82
1p	88.55	99.68	78.67
Baseline*	70.89*	82.97	60.68
ret-1	57.81	70.07	47.89
WOTAICAILE	51.03	60.83	42.96

fan.yang@fujitsu.com, sodashima@fujitsu.com, masui.shoichi@fujitsu.com, jiang.shan@fujitsu.com

Our method mainly includes two steps. First, we propose a Cascaded Buffered IoU (BIoU) for online short-term tracking. Compared with normal IoU matching, buffers are added to observations in BIoU to relax the matching space, which could alleviate the miss matching caused by motion estimation errors in complicated scenes. Meanwhile, we apply cascaded matching to match simple ones before the hard ones. Then, we link short-term tracklets to long-term ones using Hierarchical Clustering (HC). The distance matrix used in HC is formed by comparing appearance features extracted from short-term tracklets.

T3 - Camera-motion-aware appearance-based ByteTrack.

Fang Da
fang@qcraft.ai

Our submission entry builds on the ByteTrack baseline. We augment the track-detection similarity metric with a track appearance feature extractor combining semantic features from an off-the-shelf object instance segmentation network with manually-engineered features such as color histograms, which improves track association when multiple players, usually players from opposing teams, are in physical contact and have nearly coincident boxes. We implement a track re-identification post-processor that connects tracks of players disappearing and reappearing due to camera field of view, also using the track appearance features above. We develop a field pitch element detector based on the baseline pitch semantic segmentation network in the Camera Calibration track, and with its help implement a camera extrinsic parameter estimator to detect drastic camera movements, such as camera panning to follow a long ball, which helps tracking association by factoring out the camera motions from the screen-space track motions.

T9 - Movement Forecasting(MF) in Soccer Player Tracking.

Sangrok Lee
lsrock1@yonsei.ac.kr

We introduce MF, a novel architecture placed on top of DeepLabV3 to perform movement forecasting of players. In SoccerNet tracking dataset, players move dynamically and cameras also move around to follow a ball. Because of that, it is really hard to apply existing tracking methods like Kalman Filter. To tackle this problem, MF is designed to track dynamically moving players under moving camera conditions. MF predicts 2-dimensional movement vectors of players with only one feed-forwarding step and estimates a future position of objects. Although the MF is precise, it is impossible to recognize the person who disappears and reappears. Therefore, we also adopt a re-identification (re-id) module. Overall tracking process consists of the movement estimation and re-id.