SoccerNet 2023 challenges results

Anthony Cioppa, Silvio Giancola, Vladimir Somers, Floriane Magera, Xin Zhou, Hassan Mkhallati, Adrien Deliège, Jan Held, Carlos Hinojosa, Amir Mansourian, et al.

SoccerNet 2023 Challenges Results

Anthony Cioppa[†1,2*], Silvio Giancola[†2*], Vladimir Somers[†3,4,5], Floriane Magera[†1,6], Xin Zhou[†7], Hassan Mkhallati[†8], Adrien Deliège[†1], Jan Held[†1], Carlos Hinojosa[†2], Amir M. Mansourian[†9], Pierre Miralles[†10], Olivier Barnich[†6], Christophe De Vleeschouwer[†4], Alexandre Alahi[†5], Bernard Ghanem[†2], Marc Van Droogenbroeck[†1], Abdullah Kamal[11], Adrien Maglo[12], Albert Clapés[13,14], Amr Abdelaziz[11], Artur Xarles[13,14], Astrid Orcesi[12], Atom Scott[15], Bin Liu[16], Byoungkwon Lim[17], Chen Chen[18], Fabian Deuser[19], Feng Yan[20], Fufu Yu[21], Gal Shitrit[22], Guanshuo Wang[21], Gyusik Choi[23], Hankyul Kim[17], Hao Guo[16], Hasby Fahrudin[17], Hidenari Koguchi[24], Håkan Ardö[25], Ibrahim Salah[11], Ido Yerushalmy[22], Iftikar Muhammad[17], Ikuma Uchida[26], Ishay Be'ery[22], Jaonary Rabarisoa[12], Jeongae Lee[23], Jiajun Fu[27], Jianqin Yin[27], Jinghang Xu[27], Jongho Nang[23], Julien Denize[12,28], Junjie Li[21,29], Junpei Zhang[30], Juntae Kim[23], Kamil Synowiec[31], Kenji Kobayashi[24], Kexin Zhang[30], Konrad Habel[19], Kota Nakajima[24], Licheng Jiao[30], Lin Ma[20], Lizhi Wang[27], Luping Wang[16], Menglong Li[32], Mengying Zhou[18,33], Mohamed Nasr[11], Mohamed Abdelwahed[11], Mykola Liashuha[12], Nikolay Falaleev[34], Norbert Oswald[19], Qiong Jia[21], Quoc-Cuong Pham[12], Ran Song[35], Romain Hérault[28], Rui Peng[30], Ruilong Chen[34], Ruixuan Liu[18], Ruslan Baikulov[36], Ryuto Fukushima[24], Sergio Escalera[13,14,37], Seungcheon Lee[38], Shimin Chen[18], Shouhong Ding[21], Taiga Someya[24], Thomas B. Moeslund[37], Tianjiao Li[39], Wei Shen[18], Wei Zhang[35], Wei Li[18], Wei Dai[32], Weixin Luo[20], Wending Zhao[27], Wenjie Zhang[35], Xinquan Yang[18], Yanbiao Ma[30], Yeeun Joo[38], Yingsen Zeng[20], Yiyang Gan[20], Yongqiang Zhu[32], Yujie Zhong[20], Zheng Ruan[18,33], Zhiheng Li[35], Zhijian Huang[40], Ziyu Meng[35]

[1]University of Liege (ULiège). [2]King Abdullah University of Science and Technology (KAUST). [3]Sportradar. [4]UCLouvain. [5]EPFL. [6]EVS Broadcast Equipment. [7]Baidu Research. [8]Université Libre de Bruxelles (ULB). [9]Sharif University of Technology. [10]Footovision. [11]Zewail City of Science, Technology and Innovation. [12]Université Paris-Saclay, CEA, List. [13]Universitat de Barcelona. [14]Computer Vision Center. [15]Nagoya University. [16]Research Center for Applied Mathematics and Machine Intelligence, Zhejiang Lab. [17]AIBrain. [18]OPPO Research Institute. [19]University of the Bundeswehr Munich - Institute for Distributed Intelligent Systems (VIS). [20]Meituan. [21]Tencent Youtu Lab. [22]Amazon Prime Video Sport. [23]Sogang University. [24]The University of Tokyo. [25]Spiideo. [26]University of Tsukuba. [27]School of Artificial Intelligence, Beijing University of Posts and Telecommunications. [28]Normandie Univ, INSA Rouen, LITIS. [29]Shanghai

Jiao Tong University. [30]Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, Xidian University. [31]NASK – National Research Institute. [32]Robo Space. [33]Tongji University. [34]Sportlight Technology. [35]School of Control Science and Engineering, Shandong University. [36]lRomul. [37]Aalborg University. [38]Turing AI Cultures GmbH. [39]Information Systems Technology and Design, Singapore University of Technology and Design. [40]Sun Yat-sen University.

*Corresponding author(s). E-mail(s): anthony.cioppa@uliege.be; silvio.giancola@kaust.edu.sa;

Abstract

The SoccerNet 2023 challenges were the third annual video understanding challenges organized by the SoccerNet team. For this third edition, the challenges were composed of seven vision-based tasks split into three main themes. The first theme, broadcast video understanding, is composed of three high-level tasks related to describing events occurring in the video broadcasts: (1) action spotting, focusing on retrieving all timestamps related to global actions in soccer, (2) ball action spotting, focusing on retrieving all timestamps related to the soccer ball change of state, and (3) dense video captioning, focusing on describing the broadcast with natural language and anchored timestamps. The second theme, field understanding, relates to the single task of (4) camera calibration, focusing on retrieving the intrinsic and extrinsic camera parameters from images. The third and last theme, player understanding, is composed of three low-level tasks related to extracting information about the players: (5) re-identification, focusing on retrieving the same players across multiple views, (6) multiple object tracking, focusing on tracking players and the ball through unedited video streams, and (7) jersey number recognition, focusing on recognizing the jersey number of players from tracklets. Compared to the previous editions of the SoccerNet challenges, tasks (2-3-7) are novel, including new annotations and data, task (4) was enhanced with more data and annotations, and task (6) now focuses on end-to-end approaches. More information on the tasks, challenges, and leaderboards are available on https://www.soccer-net.org. Baselines and development kits can be found on https://github.com/SoccerNet.*Denotes equal contributions and †the challenges organizers.

Keywords: Soccer, artificial intelligence, computer vision, datasets, challenges, video understanding

1 Introduction

The field of video understanding continues to captivate the focus of computer vision research. As part of the commitment to advance video analysis tools within the context of sports, the SoccerNet dataset has already introduced a set of eleven tasks related to video understanding. In 2023, seven of those tasks were part of yearly open challenges aimed at the broader research community. This paper describes the conclusive outcomes of the SoccerNet 2023 challenges, briefly showcasing the solutions proposed by the participants of each challenges.

1.1 SoccerNet dataset

Originally introduced by Giancola *et al.* [1] in 2018, SoccerNet evolved into a substantial dataset tailored for reproducible research in soccer video understanding. The dataset initially had two objectives: offer a comprehensive benchmark for research in soccer video understanding and introduce the novel task of action spotting, focusing on the temporal localization of soccer actions among 500 videos for three major actions: goals, cards, and substitutions. Subsequent advancements emerged with the introduction of SoccerNet-v2 by Deliège *et al.* [2], which expanded the annotations to a complete set of

common soccer actions such as penalties, clearances, ball out of play, and more. SoccerNet-v2 also featured insights into camera transitions encompassing 13 camera classes for the task of camera shot segmentation. If a camera shot featured an action replay, it was linked to its corresponding action timestamp during the live broadcast, which defined a replay grounding task.

In 2022, Cioppa *et al.* [3] introduced SoccerNet-v3, including new spatial annotations covering players, the ball, field lines, and goal parts from various viewpoints of the same scene. Alongside, three new tasks were proposed: pitch localization, camera calibration, and player re-identification. The same year, SoccerNet-Tracking [4], introduced the task of multiple object tracking across video clips, fostering long-term tracking and including metadata such as jersey numbers and team affiliations.

Finally, SoccerNet received two upgrades in 2023. The first one, SoccerNet-Captions [5], introduced natural language descriptions of events in the broadcast games, defining a new dense video captioning task. The second one, SoccerNet-MVFouls [6], introduced a multi-view dataset for foul recognition and characterization.

1.2 SoccerNet challenges

The 2023 edition of the SoccerNet challenges proposed a set of seven tasks related to soccer video understanding. They were grouped into three main themes, painting a comprehensive picture that spanned the majority of soccer video analyses. The first theme, **broadcast video understanding**, included three major high-level tasks: (1) action spotting, focusing on retrieving all timestamps related to global actions in soccer, (2) ball action spotting, focusing on retrieving all timestamps related to the soccer ball change of state, and (3) dense video captioning, focusing on describing the broadcast with natural language and anchored timestamps. The second theme, **field understanding**, is centered on the task of (4) camera calibration, focusing on retrieving the intrinsic and extrinsic camera parameters from images. The third and last theme, **player understanding**, introduced three tasks centered on players: (5) re-identification, focusing on retrieving the same players across multiple views, (6) multiple object tracking, focusing on tracking players and the ball through unedited video streams, and (7) jersey number recognition, focusing on recognizing the jersey number of players from tracklets. Compared to previous editions of the SoccerNet challenges [7], tasks (2-3-7) are novel, introducing extra annotations and data. Task (4) was made richer with additional data and annotations, while task (6) changed its focus towards a more comprehensive approach.

1.3 Individual contributions

Anthony Cioppa and Silvio Giancola are the lead organizers of the SoccerNet challenges. They are also the lead task organizers for the action spotting and ball action spotting challenges. Vladimir Somers is the lead task organizer of the tracking challenge with Xin Zhou, as well as the re-identification and jersey number recognition challenges. Floriane Magera is the lead task organizer for the camera calibration challenge. Hassan Mkhallati is the lead task organizer for the dense video captioning challenge. Adrien Deliège co-initiated the action spotting, camera calibration, tracking, and re-identification challenges. Jan Held and Carlos Hinojosa helped with the practical organization and communication of the challenges. Amir M. Mansourian formatted the jersey number recognition dataset. Pierre Miralles provided the ball action spotting dataset, including the videos and annotations. Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem and Marc Van Droogenbroeck are the supervisors and provided funds for the annotations or the human resources to organize those challenges. The remainder of the authors are the participants who provided a summary of their method in the paper, listed in alphabetical order.

1.4 Manuscript organization

The manuscript is organized into seven sections, each summarizing the findings of a task. For each section, we describe the task and the metric used to evaluate the participant. We highlight the leaderboard of our participants for this year's challenges, and the leader provides a summary of their method. We conclude each section with the significant findings for that task. The remainder of the summaries can be found in the appendix.

2 Action Spotting

2.1 Task description

Action spotting consists in localizing the exact moments when actions of interest occur, anchored by single timestamps (*e.g.* a free-kick is defined by the precise moment when the player kicks the ball). Similarly to the two past editions of this challenge [7], the dataset comprises 500 games, with a total of 110,458 actions spanning 17 categories. In addition, a set of an extra 50 games with segregated annotations is used as the challenge set.

2.2 Metrics

We use the Average-mAP [1] metric to evaluate action spotting. A predicted action spot is considered 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 Precision-Recall (PR) curves is computed then averaged over the classes (mAP), after which the Average-mAP is calculated as the AUC of the mAP at different tolerances δ. The *loose Average-mAP* uses tolerances δ ranging from 5 to 60 seconds [1] and the *tight Average-mAP* stricter tolerances δ ranging from 1 to 5 seconds.

2.3 Leaderboard

This year, 10 teams participated in the action spotting challenge for a total of 55 submissions, with an improvement from 68.33 to 71.31 tight average mAP. The leaderboard may be found in Table 1.

2.4 Winner

The winners for this task are Wenjie Zhang *et al.*. A summary of their method is given hereafter.

A1 - MEDet: Multi-Encoder Fusion for Enhanced Action Spotting Task

Wenjie Zhang, Ran Song, Ziyu Meng, Zhiheng Li, Tianjiao Li, and Wei Zhang
{*zwjie, mziyu, zhihengli*}*@mail.sdu.edu.cn,*
{*ransong, davidzhang*}*@sdu.edu.cn,*
tianjiao_li@mymail.sutd.edu.sg

MEDet is designed for the task of detecting action instances in long untrimmed videos. We delve into the capabilities of convolutional neural networks (CNNs) and transformer neural networks in capturing local and global features and design three different encoders including Conv-based, Transformer-based and CNN-Transformer Hybrid architectures. To handle properties of various actions, MEDet divides the whole actions into three groups corresponding to different encoders. Furthermore, we proposed an improved feature pyramid network to obtain enhanced multi-scale features. Finally, our decoder utilizes a CNN-based classification head and a Trident regression head to obtain action labels and action boundaries. At inference, all result sets generated by three branches are merged and then filtered by the Non-Maximum Suppression (NMS) algorithm (in a way of model ensemble). Without using any additional datasets, the proposed MEDet demonstrates superior performance on the action spotting task.

Table 1: 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 or in Section 2.4 for the winning team.

Participants	**tight Average-mAP**			loose Average-mAP		
	main	vis.	inv.	main	vis.	inv.
SDU_VSISLAB^{S1}	**71.31**	76.29	54.09	78.56	81.67	69.13
mt_player^{S2}	71.10	**77.22**	58.50	78.79	82.02	77.62
ASTRA^{S3} [9]	70.10	75.00	57.98	**79.21**	**81.69**	75.36
team_ws_action	69.17	75.18	59.12	76.95	80.39	75.92
CEA LVA^{S5}	68.38	74.79	47.68	73.98	78.57	61.75
Baseline [8]*	68.33	73.22	**60.88**	78.06	80.58	**78.32**
DVP	66.95	74.68	53.81	73.61	79.15	67.38
JAMY2 (AF_GRU)	51.97	58.05	44.29	63.12	65.98	61.66
tyru (GRU_CALF)	51.38	57.50	41.82	62.88	66.30	56.57
JAMY (LocPoint)	45.83	49.68	45.71	61.80	64.23	63.48
test_YYQ	12.73	14.13	11.21	54.21	58.75	48.55

2.5 Results

Even though this was the third edition of the action spotting challenge, we saw an improvement for 5 out of the 10 teams over last year's state-of-the-art results by Soares *et al.* [8]. Specifically, these teams proposed to use (i) different encoder branches for different types of actions that have similar dynamics, (ii) multi-scale pyramidal architectures, (iii) fusions of features including CLIP and Video MAE features, and (iv) self-supervised pre-training.

3 Ball Action Spotting

3.1 Task description

This novel task of ball action spotting consists in localizing the exact time when two types of actions related to the soccer ball occur: *pass* and *drive*. These actions are anchored with a single timestamp corresponding to the exact time the ball leaves a player's foot for a *pass* event and when a player touches the ball to control it for a *drive* event. The dataset is provided by Footovision and is composed of 7 games with a total of 11,041 annotated timestamps, as well as 2 extra segregated games for the challenge set.

3.2 Metrics

Due to the fast nature of the event, we evaluate the performance of the methods based on the tight average mAP as well as the mAP at different δ thresholds ranging from 1 to 5 seconds. We choose to rank the methods with respect to the mAP@1, meaning that participants need to localize all ball events very precisely.

3.3 Leaderboard

This year, 5 teams participated in the action spotting challenge for a total of 102 submissions, with an improvement from 62.72 to 86.47 average mAP@1. The leaderboard may be found in Table 2.

3.4 Winner

The winner of this task is Ruslan Baikulov. A summary of his method is given hereafter.

B1 - Ruslan Baikulov

Ruslan Baikulov ruslan1123@gmail.com

The model architecture and multi-stage training significantly contribute to the overall metric outcome of the solution. The architecture utilizes a slow fusion approach, incorporating 2D and 3D convolutions. The model consumes sequences of grayscale frames stacked in sets of three. The shared 2D encoder independently processes these input threes, producing visual features. The following 3D encoder processes visual features, producing temporal features. Then, a linear classifier predicts the presence of actions. Multi-stage training was carried out in four steps. The first stage is basic training with the 2D encoder initialized with pre-trained ImageNet weights. The second stage is the same training but on the Action Spotting Challenge dataset. The third stage uses predictions from the first stage for hard negative sampling and encoders weights from the second stage for initialization. The fourth stage is fine-tuning weights from the third stage on long sequences (33 frames instead of 15 before).

Table 2: Ball Action spotting leaderboard. The main metric for the leaderboard and best performances are in bold. Team names with a superscript have provided a summary that may be found in Appendix A or in Section 3.4 for the winning team.

Participants	mAP @1	mAP @2	mAP @3	mAP @4	mAP @5	Average-mAP tight
Ruslan Baikulov^{B1}	**86.47**	**87.98**	**88.28**	**88.18**	**87.95**	**87.91**
FDL@ZLab^{B2}	83.39	85.19	85.81	86.00	86.19	85.45
BASIK^{B3}	82.06	83.39	83.86	84.04	83.91	83.57
FC Pixel Nets^{B4}	81.89	83.22	83.97	83.85	84.02	83.50
play	79.74	82.58	84.06	84.49	84.34	83.29
Baseline [10]*	62.72	69.24	72.57	74.29	74.80	71.21

3.5 Results

This new task brings three novel difficulties to the realm of action spotting. First, it focuses on fast and subtle events that provide few visual cues compared to the overall video. Second, the events are much denser compared to the action spotting challenge. It is, therefore, hard to differentiate between two actual close events or a wrong double detection. Finally, the amount of provided data is small compared to the action spotting challenge, encouraging participants to try semi-supervised, self-supervised, or transfer learning paradigms using the 500 broadcast games.

The participants focused on several aspects for this first edition of the challenge. First, some participants showed they could improve their performance by performing pre-training on the action spotting videos and later fine-tuning the network on the ball action spotting task. Next, stacked sequences of grayscale images in the RGB channels and label expansion with focal loss helped improve the performance. Finally, model ensembling helped improve the performance further by generating several network variants.

4 Dense Video Captioning

4.1 Task description

Dense video captioning is a new task introduced by Mkhallati *et al.* [5]. Given a long untrimmed video, the task consists in spotting all instants where a comment should be anchored and generating sentences describing the events occurring around that time using engaging natural language. The SoccerNet-Caption dataset comprises 471 untrimmed broadcast games at 720p resolution and 25fps. This first edition of the challenge focuses on the anonymized comments provided with the dataset, for a total of 36,894 timestamped comments. In addition, a set of 42 extra games with segregated annotations is used as the challenge set.

4.2 Metrics

Established captioning evaluation metrics such as METEOR [11], BLEU [12], ROUGE [13], and CIDEr [14] are adapted to estimate the language similarity between all generated captions with any ground-truth caption for which its timestamps fall within a δ tolerance. Then, the performances are averaged over the video and the dataset. We choose the METEOR@30, corresponding to a time tolerance of 30 seconds around the ground-truth captions, as the primary ranking metric for this challenge.

4.3 Leaderboard

This year, 4 teams participated in the camera calibration challenge for a total of 34 submissions, with an improvement from 21.25% to 26.66% METEOR@30. The leaderboard may be found in Table 3.

4.4 Winner

The winners for this task are Zheng Ruan *et al.*. A summary of their method is given hereafter.

D1 - OPPO

Zheng Ruan, Ruixuan Liu, Shimin Chen, Mengying Zhou, Xinquan Yang, Wei Li, Chen Chen, and Wei Shen

{*liuruixuan, chenshimin1, yangxinquan, liwei19, chenchen, shenwei12*}*@oppo.com* {*2130730, 2130904*}*@tongji.edu.cn*

Table 3: Dense video captioning leaderboard. The main metric for the leaderboard and best performances are in bold. Team names with a superscript have provided a summary that may be found in Appendix A or in Section 4.4 for the winning team.

Participants	Metric@30						
	METEOR	BLEU_1	BLEU_2	BLEU_3	BLEU_4	ROUGE_L	CIDEr
OPPO^{D1}	**26.66**	**35.55**	**31.03**	**28.13**	**25.65**	**33.23**	**69.73**
HZC	21.30	29.73	24.52	21.44	19.13	24.56	24.76
Baseline$_2$*	21.25	30.01	24.80	21.74	19.44	24.65	25.68
justplay^{D3}	21.20	29.83	24.68	21.66	19.38	24.34	25.89
aisoccer	21.02	29.53	24.42	21.42	19.15	24.31	23.72
Baseline$_1$*	15.24	11.91	9.97	8.83	7.97	10.69	16.33

For video captioning, we modified Blip [15] as our framework. We pick a second every 10 seconds and select the first frame of each of the 16 seconds before and after that second as input. We apply ViT [16] as the vision encoder to extract features. To better use spatial-temporal information, we add a perceiver resampler [17] after the vision encoder, which can map spatial-temporal visual features to a fixed length learned latent vectors. Then we use BERT [18] as the vision-grounded text decoder for captions generation. For localization, we use 0.875 as the threshold to filter the low-quality caption. When the confidence of the output caption is higher than the threshold, the caption will be selected. Compare with other methods [17, 19] our framework performs the best on both CIDEr and Meteor. After adding filtering, the performance is improved by +5.124% in Meteor, which shows the effectiveness of filtering. References

4.5 Results

As this was the first edition of the challenge on a novel task officially presented on arXiv in April 2023, only a few teams were able to provide results on the challenge set in time for the challenge. Nevertheless, two teams were able to improve the performance compared to our second baseline presented in the work of Mkhallati *et al.* [5]. Specifically, these methods leveraged pre-trained video-language transformers and ensembling methods to select the most suitable generated captions. As one can see from Table 3, the winning team significantly improved on all metrics compared to the rest of the methods, especially for the CIDEr@30 metric.

5 Camera Calibration

5.1 Task description

Camera calibration consists in estimating the intrinsic and extrinsic camera parameters based on an image. Similarly to last year's challenge, the pinhole camera model with optional distortion (tangential, radial, and thin prism) is imposed. This year's dataset was complemented with extra images and annotations compared to the 2022 edition. Specifically, the dataset contains 25,506 images and 226,305 annotated polylines. In addition, a set of 2,690 images with segregated annotation is used as the challenge set.

5.2 Metrics

Since no camera parameter ground truths are available for our soccer images, the methods are evaluated using the re-projection error with the manually annotated field lines in the 2D image plane. Compared to last year's challenge, we simplified the ranking metric (named *combined metric*) by computing it as the multiplication of the *accuracy*@5 and the *completeness score*. All details on how to compute those two metrics are given in Section 5 of Giancola *et al.* [7].

5.3 Leaderboard

This year, 7 teams participated in the camera calibration challenge for a total of 66 submissions, with an improvement from 8% to 55% for the combined metric. The leaderboard may be found in Table 4.

Table 4: Camera calibration leaderboard. The main metric for the leaderboard and best performances are in bold. Team names with a superscript provided a summary that can be found in Appendix A, or in Section 5.4 for the winner.

Participants	**Combined**	ACC@5	Completeness
Sportlight^{C1}	**0.55**	**73.22**	75.59
Spiideo^{C2}	0.53	52.95	**99.96**
SAIVA_Calibration^{C3}	0.53	60.33	87.22
BPP	0.50	69.12	72.54
ikapetan	0.43	53.78	79.71
NASK^{C6}	0.41	53.01	77.81
MA & JT	0.41	58.61	69.34
Baseline*	0.08	13.54	61.54

5.4 Winner

The winners for this task are Nikolay Falaleev *et al.*. A summary of their method is given hereafter.

C1 - Sportlight

Nikolay Falaleev and Ruilong Chen
nikolay.falaleev@sportlight.ai,
ruilong.chen@sportlight.ai

Our solution was based on a combination of keypoint and line detections. In total, there were 57 keypoints, most of which were defined by intersecting fitting results of lines and ellipses from annotations. To increase the overall number of available points, we utilized the correspondence between tangent points of the circles to add 8 tangent points. Additionally, 13 points were added through homography projection. We used an HRNetV2-w48-based neural network for detecting the keypoints as heatmaps, which have peaks at the keypoints locations. Similarly, 23 lines were detected, and each line was represented by a heatmap with two peaks indicating the location of the line's extremities. The camera parameters were determined using keypoints and line intersections. We applied the standard OpenCV camera calibration algorithm to subsets of points selected with various heuristics. The final camera parameters were chosen through a heuristic voting mechanism based on the reprojection error.

5.5 Results

This year, the participants improved their method by following three main research directions. First, some participants detected new pitch elements such as circles and lines intersections which improved their pitch element localization and therefore the derived camera calibration parameters. Second, some participants used keypoints spread uniformly on the ground to improve the estimated camera parameters. Finally, we saw the first solution based on differential rendering, optimizing directly the camera parameters given the synthetic projection of pitch zones and image segmentations.

6 Player Re-Identification

6.1 Task description

Person re-identification [20], or simply ReID, is a person retrieval task that 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. Re-identification is a challenging task, because person images generally suffer from background clutter, inaccurate bounding boxes [21], luminosity changes [22], and occlusions [23] from street objects or other people. 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 datasets, the SoccerNet-v3 ReID dataset is particularly challenging because soccer players from the same team have very similar appearances, which makes it difficult to tell them apart. On the other hand, each identity has a few amount of samples, which makes the model even more difficult 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) [24] at Rank-1 and the mean average precision [25] (mAP). The participants in the 2023 SoccerNet re-identification challenge are ranked according to their mAP score on a segregated challenge set.

6.3 Leaderboard

This year, 5 teams participated in the re-identification challenge, for a total of 28 submissions, with an improvement from 91.68% to 93.26% mAP. The leaderboard may be found in Table 5.

Table 5: Re-identification leaderboard. The main metric for the leaderboard and best performances are in bold. The winning team provided a summary that can be found in Section 6.4.

Participants	**mAP**	R-1	Participants	**mAP**	R-1
UniBw Munich - VIS[R1]	**93.26**	**91.26**	MTVACV	90.11	87.04
Baseline (Inspur)*	91.68	89.41	ErxianBridge	85.76	82.33
sjtu-lenovo	91.51	89.17	cm_test	42.60	28.73

6.4 Winner

The winners for this task are Konrad Habel *et al.*. A summary of their method is given hereafter.

R1 - CLIP-ReIdent: Contrastive Training for Player Re-Identification

Konrad Habel, Fabian Deuser, and Norbert Oswald

konrad.habel@unibw.de, fabian.deuser@unibw.de, norbert.oswald@unibw.de

Our approach is mainly based on our paper CLIP-ReIdent: Contrastive Training for Player Re-Identification [26]. This approach is also the 1^{st} place winning solution for the Player Re-Identification challenge 2022 [27] at the ACM Multimedia MMSports Workshop. For the SoccerNet challenge we use an ensemble of three CLIP-based models from OpenCLIP [28] and OpenAI [29]. Our approach utilizes a custom sampling strategy during training to sample players of the same action together. Furthermore, we use a self-designed re-ranking per action as post-processing. For the ensemble the vision encoders of the CLIP models are fine-tuned with contrastive training and InfoNCE loss as training objective on the data of the Train and Validation set. Our solution achieves on the Test set a mAP of 93.51% and on the Challenge set 93.26%.

6.5 Results

Last year, participants came up with various innovative ideas and achieved outstanding performances despite the difficulty of the task. Last year's winning solution was used as the baseline, but this year's winning team managed to improve upon it and set a new state-of-the-art performance. This year's solutions were mostly based on ViT [30], and most teams adopted the foundation model CLIP [29]. Moreover, participants used model ensembling, per-action player re-ranking techniques, and custom action-based training samplers to improve ranking results.

7 Multiple Player Tracking

7.1 Task description

The task of multiple player tracking aims to track individual subjects (*e.g.* players or ball) across frames. It is useful for downstream applications such as player highlights and player statistics. Different from last year's challenge [7], we did not provide any ground truth bounding boxes for the objects to track, making this task more challenging. Instead, the participants have to solve both detection and association. Compared to most tracking datasets, re-identification present extra challenges since player appearances are very similar, fast moving, and may leave the frames and come back.

7.2 Metrics

We keep the HOTA metric to rank the participants as proposed in Luiten *et al.* [31]. The metric combines a detection accuracy (DetA) and an association accuracy (AssA).

7.3 Leaderboard

This year, 7 teams participated in the multiple player tracking challenge for a total of 83 submissions, with an improvement from 42.38% to 75.61% HOTA. Table 6 presents the leaderboard.

Table 6: Tracking leaderboard. The main metric for the leaderboard and best performances are in bold. Team names with a superscript have provided a summary that may be found in Appendix A, or in Section 7.4 for the winners.

Participants	**HOTA**	DetA	AssA
Kalisteo^{T1}	**75.61**	**75.38**	**75.94**
MTIOT^{T2}	69.54	75.18	64.45
MOT4MOT^{T3}[32]	66.27	70.32	62.62
ICOST^{T4}	65.67	73.07	59.17
SAIVA_Tracking^{T5}	63.20	70.45	56.87
ZTrackers^{T6}	58.69	68.69	50.25
scnu	58.07	64.77	52.23
Baseline*	42.38	34.41	52.21

7.4 Winner

The winners for this task are Adrien Maglo *et al.*. A summary of their method is given hereafter.

T1 - Kalisteo

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

After having detected the players with a YOLO-X model, TrackMerger v2 generates player tracklets by sequentially processing the video frames. The current frame detections are matched to existing tracklets bounding boxes with two successive Hungarian assignment algorithms. The Intersection-Over-Union between bounding boxes and the distance between their center are used as criteria. A Kalman filter with camera motion compensation predicts the positions of existing tracklets in the current frame. The generated tracklets are subsequently split if they cross each other to remove most association errors. These non-ambiguous tracklets are used to fine-tune a Multiple Granularity Network re-identification model with a triplet loss formulation. Positive samples are extracted from the same tracklets as the anchor while negative samples come from concomitant tracklets. To generate full tracks, tracklets are iteratively merged according to the distance between their re-identification vectors, preventing player duplication and teleportation.

7.5 Results

For the detection step, most participants fine-tuned the YoloX detector. The maturity of the Yolo framework and its lightweight made it the go-to choice, allowing to achieve a DetA close to 75. Since no ground-truth detections were provided, lower detection scores affected the association step. As the association step was more difficult, tackling these challenges contributed to increase the AssA score. More specifically, multiple participants went beyond simple position prediction methods to compensate for fast player motions, camera motions, and deblurring. Both two-stage (such as ByteTrack, SORT based) and end-to-end tracking methods achieved strong performance. From the end-result point of view, the proposed methods perform much better than off-the-shelf open-source baselines. However, they are still some room for improvement in both detection and association.

8 Jersey Number Recognition

8.1 Task description

The task consists in identifying the jersey number of a player from a short video tracklet showing the soccer players. This jersey number recognition task is challenging because of the low quality of the thumbnails (*i.e.* low resolution and high motion blur) and because the jersey numbers might be visible on a minimal subset of the whole tracklet. The SoccerNet Jersey Number dataset comprises 2,853 tracklets of players extracted from the SoccerNet tracking videos. The challenge set comprises 1,211 separate players' tracklets with segregated annotations. The target classes are therefore all the jersey numbers from 1 to 99, and one extra "-1" class when no jersey number is visible in the tracklet.

8.2 Metrics

Since the jersey number recognition challenge was formulated as a classification task, with one class for each possible jersey number in the $[0, 99]$ range, we use the overall classification accuracy as a target metric to rank participants' solutions. Therefore, we compute the jersey number prediction accuracy as the number of correctly predicted jersey numbers (including -1 for non-visible numbers) over the total number of tracklets in the challenge set.

8.3 Leaderboard

This year, 15 teams participated in the jersey number recognition challenge, for a total of 157 submissions, with the best method reaching 92.85% accuracy. The leaderboard may be found in Table 7.

Table 7: Jersey Number Recognition leaderboard. The main metric for the leaderboard and best performance are in bold. Team names with a superscript have provided a summary that may be found in Appendix A, or in Section 8.4 for the winning team.

Participants	**acc**	Participants	**acc**
ZZPM^{J1}	**92.85**	Kalisteo	58.35
UniBw Munich - VIS^{J2}	90.95	FindNum	54.91
zzzzz^{J3}	88.08	jn	47.55
Mike Azatov	82.05	Surya	28.40
MT-IOT^{J5}	81.70	tony506672558	20.06
justplay^{J6}	77.77	lfriend	5.68
AIBrain Global Team^{J7}	75.18	zhq	4.07
SARG UWaterloo	73.77	Baseline (Random)*	3.93

8.4 Winner

The winners for this task are Rui Peng *et al.*. A summary of their method is given hereafter.

J1 - ZZPM

Rui Peng, Kexin Zhang, Junpei Zhang, Yanbiao Ma, and Licheng Jiao
{22171214876,22171214672,22171214671, ybmamail}@stu.xidian.edu.cn,
lchjiao@mail.xidian.edu.cn

The task requires the identification of players' numbers, which become difficult due to high motion blur and low quality. Our solution was to first perform text detection on the training set, filtering out a portion of the data that clearly did not have numbers, using a modified pre-trained DBNet++ model for the training set, and removing images from the training set that did not contain a detection box with a confidence level of 90 or higher. Data augmentation was then performed on the dataset, including image rotation and flipping, image scaling and cropping, color scrambling, noise addition, and multi-frame image overlay methods. We found that the data augmentation method of multi-frame fusion can provide more information and increase the robustness to features such as number shape, color and texture, thus improving the accuracy of recognition. The enhanced data were then trained for text recognition using multiple models, including SVTR-tiny, SVTR-small, SATRN, NRTR, and ASTER. The final result consists of a fusion of votes from multiple models.

8.5 Results

Most teams adopted a standard three-stage approach. A first text detection step is performed with various state-of-the-art methods (DBNet++, MMOCR, Deepsolo, YOLO) to predict the jersey number location in the image. A second text recognition step is performed to recognize the corresponding number using fine-tuned state-of-the-art OCR methods (PP-OCRv3, PaddleOCR). Finally, majority voting is employed to aggregate image-level results within a tracklet and output the final video-level prediction.

9 Conclusion

This paper summarizes the outcome of the SoccerNet 2023 challenges. In total, we present the results on seven tasks: action spotting, ball action spotting, dense video captioning, camera calibration, player re-identification, player tracking, and jersey number recognition. 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 over our proposed baselines or previously published state of the art. Some tasks such as action spotting or player re-identification are reaching promising results for industrial use, while novel tasks such as dense video captioning may still require further investigation. In the future, we will keep on extending the set of tasks, challenges, and benchmarks related to video understanding in sports.

Acknowledgement. A. Cioppa is funded by the F.R.S.-FNRS. This work was partly supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research through the Visual Computing Center (VCC) funding and the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI).

Declarations

Availability of data and code. The data and code are available at these addresses https://github.com/SoccerNet https://www.soccer-net.com.

Conflict of interest. The authors declare no conflict of interest.

Open Access.

References

[1] Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: SoccerNet: A scalable dataset for action spotting in soccer videos. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), pp. 1792–179210. Inst. Electr. Electron. Eng. (IEEE), Salt Lake City, UT, USA (2018). https://doi.org/10.1109/cvprw.2018.00223

[2] Deliège, A., Cioppa, A., Giancola, S., Seikavandi, M.J., Dueholm, J.V., Nasrollahi, K., Ghanem, B., Moeslund, T.B., Van Droogenbroeck, M.: SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In: IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports, Nashville, TN, USA, pp. 4508–4519 (2021). https://doi.org/10.1109/CVPRW53098.2021.00508

[3] Cioppa, A., Deliège, A., Giancola, S., Ghanem, B., Van Droogenbroeck, M.: Scaling up SoccerNet with multi-view spatial localization and re-identification. Sci. Data **9**(1), 1–9 (2022) https://doi.org/10.1038/s41597-022-01469-1

[4] Cioppa, A., Giancola, S., Deliege, A., Kang, L., Zhou, X., Cheng, Z., Ghanem, B., Van Droogenbroeck, M.: SoccerNet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In: IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports, pp. 3490–3501. Inst. Electr. Electron. Eng. (IEEE), New Orleans, LA, USA (2022). https://doi.org/10.1109/cvprw56347.2022.00393

[5] Mkhallati, H., Cioppa, A., Giancola, S., Ghanem, B., Van Droogenbroeck, M.: SoccerNet-caption: Dense video captioning for soccer broadcasts commentaries. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), pp. 5074–5085. Inst. Electr. Electron. Eng. (IEEE), Vancouver, Can. (2023). https://doi.org/10.1109/cvprw59228.2023.00536

[6] Held, J., Cioppa, A., Giancola, S., Hamdi, A., Ghanem, B., Van Droogenbroeck, M.: VARS: Video assistant referee system for automated soccer decision making from multiple views. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), pp. 5086–5097. Inst. Electr. Electron. Eng. (IEEE), Vancouver, Can. (2023). https://doi.org/10.1109/cvprw59228.2023.00537

[7] Giancola, S., Cioppa, A., Deliège, A., Magera, F., Somers, V., Kang, L., Zhou, X., Barnich,

O., De Vleeschouwer, C., Alahi, A., Ghanem, B., Van Droogenbroeck, M., Darwish, A., Maglo, A., Clapés, A., Luyts, A., Boiarov, A., Xarles, A., Orcesi, A., Shah, A., Fan, B., Comandur, B., Chen, C., Zhang, C., Zhao, C., Lin, C., Chan, C.-Y., Hui, C.C., Li, D., Yang, F., Liang, F., Da, F., Yan, F., Yu, F., Wang, G., Chan, H.A., Zhu, H., Kan, H., Chu, J., Hu, J., Gu, J., Chen, J., Soares, J.V.B., Theiner, J., De Corte, J., Brito, J.H., Zhang, J., Li, J., Liang, J., Shen, L., Ma, L., Chen, L., Santos Marques, M., Azatov, M., Kasatkin, N., Wang, N., Jia, Q., Pham, Q.C., Ewerth, R., Song, R., Li, R., Gade, R., Debien, R., Zhang, R., Lee, S., Escalera, S., Jiang, S., Odashima, S., Chen, S., Masui, S., Ding, S., Chan, S.-w., Chen, S., El-Shabrawy, T., He, T., Moeslund, T.B., Siu, W.-C., Zhang, W., Li, W., Wang, X., Tan, X., Li, X., Wei, X., Ye, X., Liu, X., Wang, X., Guo, Y., Zhao, Y., Yu, Y., Li, Y., He, Y., Zhong, Y., Guo, Z., Li, Z.: SoccerNet 2022 challenges results. In: Int. ACM Work. Multimedia Content Anal. Sports (MMSports), pp. 75–86. ACM, Lisbon, Port. (2022). https://doi.org/10.1145/3552437.3558545

[8] Soares, J.V.B., Shah, A., Biswas, T.: Temporally precise action spotting in soccer videos using dense detection anchors. In: IEEE Int. Conf. Image Process. (ICIP), pp. 2796–2800. Inst. Electr. Electron. Eng. (IEEE), Bordeaux, France (2022). https://doi.org/10.1109/icip46576.2022.9897256

[9] Xarles, A., Escalera, S., Moeslund, T.B., Clapés, A.: ASTRA: An Action Spotting TRAnsformer for Soccer Videos. In: Int. ACM Work. Multimedia Content Anal. Sports (MMSports). ACM, Ottawa, Canada (2023). https://doi.org/10.1145/3606038.3616153

[10] Hong, J., Zhang, H., Gharbi, M., Fisher, M., Fatahalian, K.: Spotting temporally precise, fine-grained events in video. In: Eur. Conf. Comput. Vis. (ECCV). Lect. Notes Comput. Sci., vol. 13695, pp. 33–51. Springer, Tel Aviv, Israël (2022). https://doi.org/10.1007/978-3-031-19833-5_3

[11] Lavie, A., Agarwal, A.: Meteor. In: Workshop on Statistical Machine Translation (StatMT), pp. 228–231. Association for Computational Linguistics, Prague, Czech Republic (2007). https://doi.org/10.3115/1626355.1626389

[12] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU. In: Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL), pp. 311–318. Association for Computational Linguistics, Philadelphia, PA, USA (2002). https://doi.org/10.3115/1073083.1073135

[13] Lin, C.-Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (2004)

[14] Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: Consensus-based image description evaluation. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). Inst. Electr. Electron. Eng. (IEEE), Boston, MA, USA (2015). https://doi.org/10.1109/cvpr.2015.7299087

[15] Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. CoRR **abs/2201.12086** (2022) https://doi.org/10.48550/arXiv.2201.12086

[16] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. CoRR **abs/2010.11929** (2020) https://doi.org/10.48550/arXiv.2010.11929

[17] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, M., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for

few-shot learning. CoRR **abs/2204.14198** (2022) https://doi.org/10.48550/arXiv.2204.14198

[18] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, MN, USA (2019). https://doi.org/10.18653/v1/N19-1423

[19] Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. CoRR **abs/2301.12597** (2023) https://doi.org/10.48550/arXiv.2301.12597

[20] Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H.: Deep learning for person re-identification: A survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 2872–2893 (2022) https://doi.org/10.1109/tpami.2021.3054775

[21] Zheng, W.-S., Li, X., Xiang, T., Liao, S., Lai, J., Gong, S.: Partial person re-identification. In: IEEE Int. Conf. Comput. Vis. (ICCV), pp. 4678–4686. Inst. Electr. Electron. Eng. (IEEE), Santiago, Chile (2015). https://doi.org/10.1109/iccv.2015.531

[22] Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer GAN to bridge domain gap for person re-identification, 79–88 (2018) https://doi.org/10.1109/cvpr.2018.00016

[23] Somers, V., De Vleeschouwer, C., Alahi, A.: Body part-based representation learning for occluded person Re-Identification. In: IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), pp. 1613–1623. Inst. Electr. Electron. Eng. (IEEE), Waikoloa, HI, USA (2023). https://doi.org/10.1109/wacv56688.2023.00166

[24] Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P.: Shape and appearance context modeling. In: IEEE Int. Conf. Comput. Vis. (ICCV), pp. 1–8. Inst. Electr. Electron. Eng. (IEEE), Rio de Janeiro, Brazil (2007). https://doi.org/10.1109/iccv.2007.4409019

[25] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: IEEE Int. Conf. Comput. Vis. (ICCV), pp. 1116–1124. Inst. Electr. Electron. Eng. (IEEE), Santiago, Chile (2015). https://doi.org/10.1109/iccv.2015.133

[26] Habel, K., Deuser, F., Oswald, N.: CLIP-ReIdent: Contrastive training for player re-identification. In: Int. ACM Work. Multimedia Content Anal. Sports (MMSports), pp. 129–135. ACM, Lisbon, Port. (2022). https://doi.org/10.1145/3552437.3555698

[27] Van Zandycke, G., Somers, V., Istasse, M., Don, C.D., Zambrano, D.: DeepSportradar-v1: Computer vision dataset for sports understanding with high quality annotations. In: Int. ACM Work. Multimedia Content Anal. Sports (MMSports), pp. 1–8. ACM, Lisbon, Port. (2022). https://doi.org/10.1145/3552437.3555699

[28] Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: OpenCLIP. Zenodo (2021). https://doi.org/10.5281/zenodo.5143773

[29] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. (ICML), pp. 8748–8763 (2021)

[30] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. CoRR **abs/2010.11929**

(2020) https://doi.org/10.48550/arXiv.2010.11929

[31] Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: HOTA: A higher order metric for evaluating multi-object tracking. Int. J. Comput. Vis. **129**(2), 548–578 (2021) https://doi.org/10.1007/s11263-020-01375-2

[32] Shitrit, G., Be'ery, I., Yerhushalmy, I.: SoccerNet 2023 tracking challenge – 3rd place MOT4MOT team technical report. CoRR **abs/2308.16651** (2023) https://doi.org/10.48550/arXiv.2308.16651

[33] Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: VideoMAE v2: Scaling video masked autoencoders with dual masking. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 14549–14560. Inst. Electr. Electron. Eng. (IEEE), Vancouver, Can. (2023). https://doi.org/10.1109/cvpr52729.2023.01398

[34] Zhou, X., Kang, L., Cheng, Z., He, B., Xin, J.: Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection. CoRR **abs/2106.14447** (2021) https://doi.org/10.48550/arXiv.2106.14447

[35] Wang, L., Guo, H., Liu, B.: A boosted model ensembling approach to ball action spotting in videos: The runner-up solution to CVPR'23 SoccerNet challenge. CoRR **abs/2306.05772** (2023) https://doi.org/10.48550/arXiv.2306.05772

[36] Liu, S., Chen, W., Li, T., Li, H.: Soft rasterizer: A differentiable renderer for image-based 3D reasoning. In: IEEE Int. Conf. Comput. Vis. (ICCV), pp. 7707–7716. Inst. Electr. Electron. Eng. (IEEE), Seoul, South Korea (2019). https://doi.org/10.1109/iccv.2019.00780

[37] Zhang, Y., Wang, T., Zhang, X.: MOTRv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. CoRR **abs/2211.09791** (2022) https://doi.org/10.48550/arXiv.2211.09791

[38] Yan, F., Luo, W., Zhong, Y., Gan, Y., Ma, L.: Bridging the gap between end-to-end and non-end-to-end multi-object tracking. CoRR **abs/2305.12724** (2023) https://doi.org/10.48550/arXiv.2305.12724

[39] Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: Exceeding YOLO series in 2021. CoRR **abs/2107.08430** (2021) https://doi.org/10.48550/arXiv.2107.08430

[40] Cao, J., Pang, J., Weng, X., Khirodkar, R., Kitani, K.: Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 9686–9696. Inst. Electr. Electron. Eng. (IEEE), Vancouver, Can. (2023). https://doi.org/10.1109/cvpr52729.2023.00934

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 ensures that the presented methods followed the challenge rules.

Action Spotting

S2 - mt_player

Yingsen Zeng, Yujie Zhong, Zhijian Huang, Feng Yan, Lin Ma
{*zengyingsen, zhongyujie, yanfeng05*}*@meituan.com, huangzhj56@mail2.sysu.edu.cn, linma@alumni.cuhk.net*

Our proposed method aims to improve the accuracy of action detection in untrimmed videos through multi-scale and multi-feature fusion. The model is based on an encoderdecoder structure, including a fully convolutional encoder, a multi-scale feature pyramid network, and a lightweight decoder for action classification and location. To tackle single timestamp labeling for action locations, we employ a soft-NMS approach based on Euclidean distance after IoU-based NMS. Additionally, we discover that different video features exhibit biases towards different action categories, and therefore, we explore three methods of feature fusion: early fusion, mid-level fusion, and late fusion. Four pre-computed video features we use are Baidu, ResNet (both from Soccernet Challenge), CLIP [29], and VideoMAE [33]. Ultimately, we achieve the best performance by combining mid-level fusion and class-wise late fusion. Our method is extensively tested and demonstrates its effectiveness, achieving 71.1% tight-mAP and 78.8% loose-mAP in challenge set.

S3 - ASTRA [9]

Artur Xarles, Sergio Escalera, Thomas B. Moeslund, Albert Clapés
arturxe@gmail.com, sescalera@ub.edu, tbm@create.aau.dk, aclapes@ub.edu

Action Spotting TRAnsformer (ASTRA) leverages pre-computed visual embeddings from five video classification backbones provided by Baidu Research [34]. ASTRA combines these embeddings in a transformer encoder-decoder module with learnable queries on the decoder side. This design enables the model to handle different input and output temporal dimensions, resulting in improved performance with higher temporal resolution for the outputs. To further enhance ASTRA's capability to capture fine-grained details, we introduce the concept of temporally local attention within the transformer encoder. Following Soares et al. [8], two prediction heads are employed to predict the action classification and the displacement offsets with respect to the predefined anchors. To enhance generalization and address the long-tail distribution of the data, ASTRA incorporates a balanced mix-up technique. This technique generates data mixtures during training using an action-balanced data distribution stored in a queue, which is updated during each batch iteration. More information is available in our paper published at MMSports'23 [9].

S5 - COMEDIAN: Long-Context Transformer Pretraining Through Spatio-Temporal Knowledge Distillation for Action Spotting

Julien Denize, Mykola Liashuha, Jaonary Rabarisoa, Astrid Orcesi, Romain Hérault
{*julien.denize, mykola.liashuha, jaonary.rabarisoa, astrid.orcesi*}*@cea.fr, romain.herault@insa-rouen.fr*

COMEDIAN is an approach to pretrain spatiotemporal transformer architectures to enrich local spatiotemporal information with a larger context before specializing in the action spotting task. These transformer architectures contain two encoders. The first is spatial as it takes input frames from a short video to embed their information in one output token. The second encoder is a temporal encoder that takes as input the spatial output tokens of sub-videos from a large video. To pretrain the architecture, knowledge distillation is performed between the output tokens of the temporal encoder and their provided temporally aligned Baidu features [34]. The architecture is then finetuned to the action spotting task by performing the classification of each action independently on each timestamp associated with the temporal output tokens. For inference, a sliding window is performed on videos to provide predictions of each action at each timestamp followed by a soft NMS per class.

Ball Action Spotting

B2 - Boosted Model Ensembling (BME)

Luping Wang, Hao Guo, and Bin Liu
{*wangluping, guoh, liubin*}*@zhejianglab.com*

Our method, Boosted Model Ensembling (BME), is based on the end-to-end baseline model, E2E-Spot, as presented in [10]. We generate several variants of the E2E-Spot model to create a candidate model set and propose a strategy for selecting appropriate model members from this set while assigning appropriate weights to each selected model. More details can be found in [35]. The resulting ensemble model takes into account uncertainties in event length, optimal network architectures, and optimizers, making it more robust than the baseline model. Our approach has the potential to handle various video event analysis tasks.

B3 - BASIK

Juntae Kim, Gyusik Choi, Jeongae Lee, and Jongho Nang
{*jtkim1211, gschoi, jalee3, jhnang*}*@sogang.ac.kr*

Our primary focus is on tackling the class imbalance in the dataset of nine 90-minute soccer games, most frames of which are labelled as *background*. We successfully implement Label Expansion to extend the labels of *PASS* and *DRIVE* frames to adjacent frames. The best performance is consistently achieved with a window size of four frames. The class imbalance is further mitigated through the application of the Focal Loss function, achieving optimal results with α=1 and γ=2. Furthermore, we substitute the Gated Recurrent Unit in the original model with a Transformer encoder for better temporal reasoning. The final model is an ensemble of three models with different temporal reasoning architectures, contributing to a substantial 13.99% improvement in test set performance compared to the baseline model.

B4 - FC Pixel Nets

Ikuma Uchida, Atom Scott, Taiga Someya, Kota Nakajima, Kenji Kobayashi, Hidenari Koguchi, and Ryuto Fukushima
uchida.ikuma@image.iit.tsukuba.ac.jp,
{*atom.james.scott,atokota1022*}*@gmail.com,*
{*taiga98-0809, kobayashi-kenji, hidenari-1108-hk, fukushima-ryuto0407*}*@g.ecc.u-tokyo.ac.jp*

In this Challenge, we aimed to enhance the E2E-Spot baseline method. We introduced RandomAffine, RandomPerspective and Mixup data augmentation during training, and adopted Focal Loss as the loss function. We conducted various experiments using the improved E2E-Spot model, including changes in input clip length and image size. We selected the top two models and applied Test-Time Augmentation to average their output class probabilities. Post-processing included the Savitzky-Golay filter, peak detection, and Non-Maximum Suppression. These improvements led to a significant increase in the mAP@1 metric score, reaching 83.53%.

Additionally, we implemented a two-stage action recognition architecture using E2E-Spot models trained on optical flow, grayscale stacked frames, and RGB frames. We also tried combining YOLOv8 player detection with PySceneDetect for scene transition identification, to extract replay scenes by identifying camera zoom-ins based on a thresholded average ratio of the player's bounding box to image size. However, these attempts had limited impact on performance.

Dense Video Captioning

D3 - justplay

Wei Dai, Yongqiang Zhu, and Menglong Li
loveispdvd@gmail.com, alexzhu.vip@gmail.com, mlli8803@163.com

The baseline for this task is Temporally-Aware Feature Pooling for Dense Video Captioning in Video Broadcasts [5]. This approach divides the task of dense video captioning (DVC) into two stages: spotting and captioning. The first stage involves locating the events that need to caption, while the second stage involves generating captions for these events. Therefore, the baseline requires training two models: a spotting model and a captioning model. Pre-trained weights for both models are provided in the GitHub repository for this task. To improve the performance of our captioning model, we experimented with replacing the pooling layer in the model. Additionally, we fine-tuned both the spotting and captioning models for a few epochs. The resulting METEOR score was 21.2, which was higher than the score of baseline1 but still lower than the score of baseline2.

Camera Calibration

C2 - Spiideo

Håkan Ardö
hakan.ardo@spiideo.com

The camera parameters are estimated from the images in two steps. A pixel-level segmentation followed by a camera optimization using differential rendering. Code is available at https://github.com/Spiideo/soccersegcal. The segmentation is performed using a DeepLabV3 CNN, segmenting the image into 6 classes, representing different parts of the field. Sigmoid activations are used, which allows the different classes to overlap. Ground truth segmentations are generated from the SoccerNet annotations using floodfill operations. A synthetic image segmentation is generated by rendering a 3D soccer-pitch and given camera parameters. The differential renderer SoftRasterizer[36] is used and the camera parameters updated using the local optimizer AdamW. To initiate the process a set of predefined start cameras are found by clustering all the cameras produced by the SoccerNet baseline method on the training data into 20 clusters using k-means. They are then pre-optimized using a loss that align the centers of the measured and rendered segments.

C3 - SAIVA_Calibration

Hankyul Kim
harry.kim@aibrain.co.kr

SAIVA (Soccer AI Virtual Assistant) is an advanced AI platform designed for soccer match video analysis, with a key feature being the camera calibration module that supports other SAIVA modules with precise location data. This module offers two calibration methods: object detection, trained on a homography matrix with field transformation data, and keypoint detection, trained on 221 annotated soccer field keypoints. These methodologies were synergized in the SoccerNet Camera Calibration 2023 Challenge. Here, the predicted homography matrix assists in choosing keypoints, evaluated based on distance from segmentation lines. The module scored 0.52 and 0.53 respectively in the test and challenge set evaluations, demonstrating its proficiency.

C6 - NASK

Kamil Synowiec
kamil.synowiec@nask.pl

The proposed approach consists of two main parts. Firstly, annotated points belonging to soccer pitch elements were transformed to segmentation masks by utilizing curve fitting techniques. These masks served as input for an instance segmentation model - Mask2Former with Swin-S hierarchical vision transformer as its backbone. Model was trained to detect and classify various field elements, including lines, conics, and goal parts. Subsequently, specific pitch points were localized by identifying the intersection of lines and ellipses derived from output segmentation masks. To compute the homography matrix, at least four image points mapped to corresponding points from the 3D pitch model are required. This matrix was estimated using RANSAC solver with different values of maximum reprojection threshold. Based on calculated homography, camera parameters were extracted and further refined using Perspective-n-Point (PnP) solver to obtain final results.

Multiple Player Tracking

T2 - Enhance End-to-End Multi-Object Tracking by CO-MOT

Feng Yan, Weixin Luo, Yiyang Gan, Yujie Zhong, and Lin Ma
{yanfeng05, luoweixin, ganyiyang, zhongyujie, malin11}@meituan.com

We propose a effective approach to enhance end-to-end multi-object tracking based on motrv2 [37], by incorporating a coopetition label assignment proposed by CO-MOT [38]. To address the imbalance between positive and negative samples for detection queries caused by Tracking Aware Label Assignment, especially in the closed environment of SoccerNet matches where all objects are detected in the initial frame and there are few new objects in subsequent frames, we introduce the Copetition Label Assignment. In the first five layers of the decoder, detection queries are responsible not only for detecting new objects but also for detecting already tracked objects. This significantly increases the number of positive samples and effectively trains the detection queries. Thanks to the self-attention mechanism in the decoder, the performance of the detection queries is transferred to the tracking queries, further improving tracking performance. We achieved a 69.5 HOTA on the Tracking 2023

challenge data without using any additional data at https://github.com/BingfengYan/CO-MOT

T3 - MOT4MOT [32]

Ishay Be'ery, Gal Shitrit, and Ido Yerushalmy
{ishaybee, galshi, idoy}@amazon.com

For player tracking, we employ a state-of-the-art online multi-object tracker DeepOCSORT along with a fine-tuned YOLO8 object detector. We finetuned an appearance model on a well curated dataset from the training set. To overcome the limitations of the online approach, we incorporate a post-processing stage that includes interpolation and appearance free track merging. Additionally, an appearance-based track merging technique is used to handle track termination and creation far from the image boundaries. For ball tracking, we treat it as a single object detection problem and utilize a fine-tuned YOLOv8l detector. For training the detector we curate the training data from erroneous labels using pre-trained ball detector and use only well annotated frames. In addition, we used filtering techniques such as estimating the ball trajectory as 3rd order polynomial with a large temporal window and rejecting detections that are far from this trajectory to enhance detection precision. More information is available in our technical report [32].

T4 - ICOST

Jiajun Fu, Jinghang Xu, Wending Zhao, Lizhi Wang, and Jianqin Yin
{JaakkoFu,xjh_amber,windy,wanglizhi,jqyin}
@bupt.edu.cn

The first step for our method is to detect balls, players, and referees in the image. Since the blurry image caused by camera or player movement will introduce missing detections, we first run a state-of-the-art image deblurring method to pre-preprocess images. Then, we trained a YOLOX detector [39] with the preprocessed training data. This can prevent false detections like the spectators and ball boys. The second step is to run a short-term tracking algorithm. We adopted OC-SORT [40] and introduced a Buffered Complete Intersection of Union (BCIoU) for the association between detections and tracklets. We also integrate Camera Motion Compensation. Finally, we introduce Co-occurrence-aware Hierarchical Clustering to merge tracklets, where the mergence between two tracklets with co-occurrent detections is prohibited. The clustering is conducted by comparing the appearance features. The appearance features are extracted from a reid network.

T5 - SAIVA_Tracking

Byoungkwon Lim, Yeeun Joo, Seungcheon Lee
bklim@aibrain.co.kr, yeeun.joo@turingai.global, sclee@turingai.global

As part of SAIVA (Soccer AI Virtual Assistant), the player's tracking system is based on ByteTrack and four added features - Camera movement estimation, IoU based distance, Order distance and Outside player calibration. The Camera movement estimation focuses on adjusting distances between tracks and detecting objects. This takes place by estimating and comparing paired objects on consecutive frames. The IoU based distance is developed for estimating distances between objects without overlapping. In that case, its distance value will be a constant which makes two areas close together by enlarging its area. The Order distance is being combining with IoU based distance and used to minimize object detection errors. It calculates distances, using each object's X,Y axis order as elements. The Outside player calibration improves focus on the tracking objects by eliminating inactive tracks if the duration of outside detection is over a certain threshold. SAIVA tracking system records HOTA 63.19 about SoccerNet Player Tracking Challenge 2023

T6 - ZTrackers

Ibrahim Salah, Mohamed Abdelwahed, Abdullah Kamal, Mohamed Nasr, and Amr Abdelaziz
{s-ibrahimsaad,s-mohamedabdelwahed,
s-abdullahkamal,s-mohamed_nasr,
s-amr.ragab1041}@zewailcity.edu.eg

Our methodology involved combining the YOLOv5 Medium object detection model with the ByteTrack algorithm. We trained the YOLOv5 Medium model on annotated soccer game images to accurately detect players and the ball. The detections obtained from YOLOv5 Medium were then passed to the ByteTrack algorithm. We fine-tuned several parameters, including the tracking confidence threshold, frames buffer, and matching threshold, to optimize tracking accuracy, smoothness, and robustness. Additionally, we fine-tuned the parameters of the Kalman filter to enhance velocity and position estimates. By integrating YOLOv5 Medium, ByteTrack, and the fine-tuned Kalman filter parameters, we developed a robust

system capable of providing real-time and accurate player and ball tracking in the soccer video game

Jersey Number Recognition

J2 - CLIP Zero-Shot Jersey Number Labeling

Konrad Habel, Fabian Deuser, and Norbert Oswald
konrad.habel@unibw.de, fabian.deuser@unibw.de, norbert.oswald@unibw.de

Our approach uses an ensemble of two CLIP [29] pre-trained models fine-tuned on the SoccerNet Re-Identification dataset instead of the tracklets of the actual SoccerNet Jersey Number Recognition dataset. Due to the high amount of label noise in the dataset of the challenge, we decided to train our models not on the given data and labels. Instead, we use the ViT-L14 CLIP model of OpenAI for automatically zero-shot labeling the more diverse SoccerNet Re-Identification dataset without any human annotation. On these pseudo labels per image we fine-tuned the image encoders of the two models. To predict the jersey number on a tracklet basis, we use a majority voting taking only images with a classification probability over 70% for numbers 1 - 99 into account. We achieve on the Test set an accuracy of 90.09% and on the Challenge set 90.95%.

J3 - zzzzz

Junjie Li, Guanshuo Wang, Fufu Yu, Qiong Jia, and Shouhong Ding
serenitycapo@gmail.com, {mediswang, fufuyu, boajia, ericshding}@tencent.com

We approached this challenge by employing various problem formulation methodologies, which involved formulating the task as an optical character recognition (OCR) problem and modeling the recognition of jersey numbers as a sequential prediction based on tracklets. To begin with, we utilized a state-of-the-art text detection model called Deepsolo to obtain initial OCR results. Building upon these predictions at the image level, the sequential prediction model has been designed in a similar manner to transformer, where the image patches are replaced by feature representations of images in a tracklet. To further boost the performance, such as performing model ensemble with multi-input resolution. Finally, the sequential prediction results are refined with the original text detection results to obtain the final predictions.

J5 - MT-IOT

Gan Yiyang, Luo Weixin, Yan Feng, Lin Ma
{ganyiyang, luoweixin, yanfeng05}@meituan.com, forest.linma@gmail.com

The accurate recognition of jersey numbers in soccer broadcasts is important for tracking player movements, evaluating performance, and making strategic decisions during games. To address this task, the authors propose a multi-task video classification approach that leverages temporal and spatial cues from player tracklets. They employ an advanced video transformer network as a powerful backbone to extract features from tracklets and designed a multi-task classification head to address the issue of long-tailed data distribution. The task is divided into two sub-tasks: predicting the two digits of the number separately and predicting the permutation of the digits. The authors use binary cross-entropy loss for the digit head and cross-entropy loss for the permutation head to effectively address the issue of unbalanced data. The proposed approach achieves competitive accuracy of 87.37% on the test set and 81.70% on the challenge set, showcasing its effectiveness.

J6 - justplay

Wei Dai, Yongqiang Zhu, and Menglong Li
loveispdvd@gmail.com, alexzhu.vip@gmail.com, mlli8803@163.com

The first step is to detect the jersey numbers present in the image. This is a necessary step before proceeding with recognition. In the second step, we manually reviewed over 700 folders in both the training and testing sets, which contained the cropped jersey number patches during the detection phase. Each folder only retained the patches with the same number as the folder's annotation, while removing those with different detected numbers and those with irrelevant contents due to false detection. For the folder labeled with no jersey number, we added some patches without jersey numbers. This allows the model to recognize results for no jersey number as well. Finally, we inferred on the challenge set using the fine-tuned detection model and the fine-tuned recognition model. The majority of recognition results in each folder were taken as the predicted result for that folder. After submission, the accuracy score was 77.77.

J7 - AIBrain Global Team

Iftikar Muhammad and Hasby Fahrudin
{*iftikarm,hfahrudin*}*@aibrain.co.kr*

As an effort of SAIVA (Soccer AI Virtual Assistant), we propose an approach to accurately detect and recognize jersey numbers by leveraging player body orientation information. Our method aims to emulate human perception when observing tracklet images to determine the jersey numbers. Instead of relying on a large number of tracklet images, we utilize a confidence sorting algorithm based on the visibility of the jersey number and the quality of the image. First to tackle the low-resolution issue we upscale the image using ESRGAN-based model, and then we utilize keypoints extracted from pose-estimation model to localize the jersey number and get the body orientation. Lastly we use body orientation and image quality assessment to rank the prediction confidence of each tracklet. On SoccerNet Challenge 2023 our work achieved 76.05% on Test-set and 75.17% on Challenge-set in terms of classification accuracy.