

Multi-task Learning for Joint Re-identification, Team Affiliation, and Role Classification for Sports Visual Tracking

Amir M. Mansourian*

Sharif University of Technology
Tehran, Iran
amir.mansurian@sharif.edu

Christophe De Vleeschouwer

UCLouvain
Louvain-la-Neuve, Belgium
christophe.devleeschouwer@uclouvain.be

Vladimir Somers*

UCLouvain & EPFL & Sportradar
Louvain-la-Neuve, Belgium
vladimir.somers@uclouvain.be

Shohreh Kasaei

Sharif University of Technology
Tehran, Iran
kasaei@sharif.edu

ABSTRACT

Effective tracking and re-identification of players is essential for analyzing soccer videos. But, it is a challenging task due to the non-linear motion of players, the similarity in appearance of players from the same team, and frequent occlusions. Therefore, the ability to extract meaningful embeddings to represent players is crucial in developing an effective tracking and re-identification system. In this paper, a multi-purpose part-based person representation method, called PRTreID, is proposed that performs three tasks of role classification, team affiliation, and re-identification, simultaneously. In contrast to available literature, a single network is trained with multi-task supervision to solve all three tasks, jointly. The proposed joint method is computationally efficient due to the shared backbone. Also, the multi-task learning leads to richer and more discriminative representations, as demonstrated by both quantitative and qualitative results. To demonstrate the effectiveness of PRTreID, it is integrated with a state-of-the-art tracking method, using a part-based post-processing module to handle long-term tracking. The proposed tracking method, outperforms all existing tracking methods on the challenging SoccerNet tracking dataset.

CCS CONCEPTS

- Computing methodologies → Tracking.

KEYWORDS

Computer Vision, Deep Learning, Sports Videos, Re-Identification, Multi-Object Tracking, Soccer, SoccerNet, Part-based Re-Identification, Team Affiliation, Multi-task Learning, Deep Metric Learning, Representation Learning.

* Authors contributed equally to this work.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MMSports '23, October 29, 2023, Ottawa, ON, Canada

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 979-8-4007-0269-3/23/10...\$15.00
<https://doi.org/10.1145/3606038.3616172>

ACM Reference Format:

Amir M. Mansourian, Vladimir Somers, Christophe De Vleeschouwer, and Shohreh Kasaei. 2023. Multi-task Learning for Joint Re-identification, Team Affiliation, and Role Classification for Sports Visual Tracking . In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports (MMSports '23), October 29, 2023, Ottawa, ON, Canada*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3606038.3616172>



Figure 1: Automated analytics for team sports requires tracking, re-identification, role classification (e.g. player, referee, staff, ...), and team affiliations (such as Team A or Team B) of all detected persons throughout an entire video of a game.

1 INTRODUCTION

Automated sport analysis involves the use of computer vision and deep learning techniques to analyze sports events and extract meaningful insights. It shows potential applications in a variety of areas, such as sports broadcasting [9, 43, 48], autonomous and personalized production [5, 7, 31], coaching [1, 6, 17, 42], and performance analysis [11, 20, 40, 41, 44, 54]. One of the key components of automated sport analysis is object tracking, which involves the identification of players and objects all along a video. Tracking is essential for automated sport analysis for three main reasons. Firstly, it allows for the identification and distinction of players, enabling the extraction of personalized information. Secondly, it

creates a spatio-temporal representation of the game, providing a detailed overview that can identify key events, game patterns, and trends. Lastly, automatic tracking reduces the manual effort required for game strategy analysis, allowing analysts to focus on higher-level tasks.

Tracking is challenging due to complex and dynamic motion, occlusions, inter-appearance similarities and intra-appearance variations. Re-identification, on the other hand, provides an important cue to solve tracking challenges after an occlusion or a move outside the camera field-of-view.

Moreover, end-user applications typically demand that the upstream tracking system provides comprehensive high-level data regarding the individuals involved in sports, including roles (like player, referee, staff, goalkeeper, and so on) and team affiliations (such as Team A or B). Consequently, as illustrated in Figure 1, a desirable sports tracking system should simultaneously tackle tracking, re-identification, team affiliation, and role classification challenges.

Despite substantial advancements in sports analysis methods, as shown in recent studies [8, 14, 19, 27, 30, 45, 53], the majority of current tracking methods do not tackle all these tasks together. Solving each task individually is also not optimal as it overlooks the common objectives shared by all three tasks for accurately representing the individual, which could potentially benefit from a unified approach. Most published sports tracking papers tend to concentrate on one or two tasks (like tracking and re-identification), often neglecting team and role identification, or the other way around. Moreover, in those approaches, the role classification task is typically accomplished by an upstream object detector. Most state-of-the-art pre-trained detectors are limited to the "person" class and cannot distinguish between various sports roles (player, goalkeeper, referee, staff, etc). Similarly, the team affiliation task is often incorrectly formulated as a classification task with predefined teams [48, 55], preventing the model to be used to distinguish new teams, unseen at the training phase. Finally, Re-Identification (ReID) models trained solely with identity labels may not be suitable for team affiliation, as they may consider players with similar attributes, such as skin color, more similar to each other than players from the same team, resulting in poor team clustering.

To address these limitations, a novel representation model is proposed that generates a multi-purpose representation from a single backbone, enabling to jointly solve the three tasks of person re-identification, team affiliation, and role classification. The problems of re-identification and team affiliation are formulated as deep metric learning tasks, where players with the same identities are matched according to their similarity (i.e., distance) in the feature space. The team affiliation within a video is performed by clustering players' representations into two groups, thereby being applicable to the recognition of teams that were not seen at training. On the other hand, sports person's role prediction is formulated as a classification task. Our method enjoys two key benefits; i) by integrating three distinct representation objectives during training, the model generates richer representations resulting in superior re-identification performance compared to models trained with a single objective and ii) it is efficient both in terms of speed and memory as it jointly solves three tasks using a single backbone and a single representation.

In practice, a state-of-the-art body part-based re-identification (BPBreID) model [35] is used as the baseline on top of which the proposed joint model is built by adding two objectives dedicated to role classification and team affiliation. A multi-task approach is then utilized to train the model jointly with all three objectives of re-identification, role classification, and team affiliation. At inference, the model produces part-based features that can be used to solve these three tasks, simultaneously. The method is called *PRTreID* for **P**art-based **R**ole classification **T**eam affiliation and person **ID**entification

Furthermore, the proposed PRTreID model is integrated with a state-of-the-art re-identification-based tracking method, namely StrongSORT [12], and the post-processing step of this tracker is replaced with the proposed part-based tracklet merging module.

In summary, the main contributions of this work are as follows:

- The PRTreID, a novel multi-task sports person representation model, is proposed to address re-identification, team affiliation, and role classification, simultaneously. To the best of our knowledge, this work is the first to address these three significant sport representation tasks with a single model and to demonstrate the benefit of multi-task learning in enhancing the richness of representation features.
- The PRT-Track, a novel StrongSORT-based tracking method, is proposed to leverage the multi-task sports person representation model to produce long-term tracks.
- Extensive experiments are conducted to validate the effectiveness of the proposed method and demonstrate the key benefits of multi-task learning.

The prepared codebase and dataset will be released to encourage further research on joint representation learning for sports.

2 RELATED WORK

The literature most relevant to this work includes researches surrounding the part-based re-identification and multiple object tracking.

2.1 Part-Based Re-Identification

Part-based re-identification methods have recently achieved state-of-the-art results in re-identification tasks, particularly in scenarios with occlusions, due to their ability to use local features for each part. However, one of the challenges faced by these methods is the need to localize each part accurately. Two different approaches are commonly used to address this problem. The first approach uses a fixed parts, which does not require any additional information for localizing parts [39, 49, 51]. In the second approach, however, the method usually involves a pose estimation model to extract body parts, which is trained concurrently with the re-identification part [21, 36]. For example, [35] trains a model to find body parts of a person, using human parsing labels, and extracts embeddings for each part simultaneously. [34] proposes a part-based approach specific to sport works for player identification in basketball.

In the context of team sports, there are also other cues that can be helpful in re-identification of a person, such as team affiliation [20, 23, 48], role information [23], and jersey number [26, 46, 47]. [20] trains a network that can output embeddings that are close for players on the same team and far from each other for players from

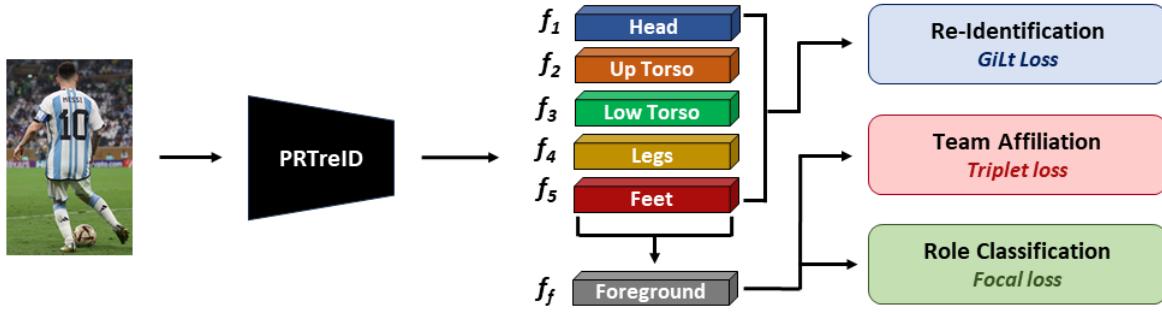


Figure 2: Diagram of the proposed PRTreID method. An input image is fed into a shared backbone, which outputs an embedding for each part of the body. The foreground mask is created by combining all parts embeddings. The re-identification objective is trained with triplet loss and cross entropy loss on the body parts embeddings. Additionally, team affiliation and role classification objectives are trained with triplet loss and focal loss, respectively, on the foreground embedding. At inference time, the resulting multi-purpose embeddings can be utilized for person re-identification, team affiliation, and role classification tasks.

different teams. In [48] an automated system containing tracking, team classification, and jersey number recognition but used distinct networks to extract those multiple information. In contrast, similar to our work, [55] presents a deep player model for re-identification using jersey number, team class, and pose estimation features with a shared backbone. Our work goes one step further by adopting part-based representation features (and simply pose parameters), and by making those embedded features compliant with role and team affiliation prediction.

This paper utilizes the re-identification model introduced in [35] to extract body-part-based features for persons in soccer videos, and adds two objectives for training the model: team affiliation and role classification. Training the model jointly for these three tasks leads to more meaningful part-based embeddings.

2.2 Multiple Object Tracking

The dominant paradigm in existing multiple object tracking methods is track-by-detection, where objects are detected in each frame before being associated across frames. SORT [3] used a Kalman filter motion model and Intersection over Union (IoU) criteria for object association. Deep SORT [50] improved upon SORT by adding appearance features alongside motion. Recent works such as OC-SORT [4] have made changes to SORT to handle non-linear motions, while others such as CBIoU [52] use a buffered version of IoU to extend bounding boxes for better object matching. More recently, re-identification-based trackers, like StrongSORT [12], have emerged, which focus on extracting more discriminative features for object re-identification to improve tracking results [2, 29, 57]. Some works have also been specifically developed for tracking team sport players [16, 22, 30, 32, 33]. [55] proposes a deep learning-based approach for multi-camera multi-player tracking in sports videos, leveraging deep player identification to improve tracking accuracy and consistency across multiple cameras. [22] addresses the specificity of use cases where appearance features are quite discriminant but only available sporadically, like it happens for the digits printed on the shirt of team sport players.

In this work, the Strong-SORT method [12] is adopted. To improve the re-identification part, its appearance embeddings are

replaced with part-based ones. Additionally, a post-processing step is proposed to further improve the tracking results by merging short tracklets with similar part-based features.

3 PROPOSED METHODS

In this section, first the body part-based re-identification baseline, initially proposed in [35] is explained. Subsequently, two new components are introduced. They correspond to the *team affiliation* training objective and to the multi-role classifier head. Finally, a tracking paradigm exploiting our proposed part-based features.

3.1 PRTreID

The overall architecture of our part-based ReID method *PRTreID* is illustrated in Figure 2. Compared to global Re-Identification (ReID) methods that produce a single embedding for each input sample, part-based methods produce multiple embeddings, each corresponding to a specific body region of the target person. Part-based methods have shown superior performance when facing occlusions [13, 38] and are therefore a promising solution to address the multi-object tracking process.

3.1.1 Part-based Re-Identification Baseline. In this work, the body part-based re-identification method, introduced in [35], is employed as a baseline to re-identify persons in soccer videos. The so-called BPBReID model consists of a CNN feature extractor backbone, along with two modules for body-part attention and ReID, used for person re-identification. The backbone receives an input image and produces a spatial feature map ($G \in \mathbb{R}^{C \times H \times W}$) that is utilized in both modules. In the body part attention module, for each pixel (h, w) , a pixel-wise part classifier predicts whether this pixel belongs to the background or one of the K body parts. This process is similar to a segmentation task for each body part, with $K + 1$ classes corresponding to one of the K body parts or to the background. This module employs a pixel-wise cross-entropy loss and utilizes human parsing labels (obtained from the Open Part Intensity Field, Part Association Field (OpenPifPaf) [24] pose estimation model) for training. The overall loss to train this attention mechanism is

called the *part-prediction loss* ℓ_{pa} , and is the sum of all pixel-wise cross-entropy loss on the spatial feature map G:

$$\ell_{pa} = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} \ell_{CE}(G(h, w)) \quad (1)$$

This body part attention module produces two outputs for each body part: a mask ($M_k \in \mathbb{R}^{H \times W}; k \in \{1, \dots, K\}$) and a binary visibility score ($v_k; k \in \{1, \dots, K\}$) that indicates if the part is visible in the image. The masks of all parts are then merged to create a foreground mask, denoted M_f . The second module takes the outputs of the first module and performs a weighted average pooling of the spatial feature map using each attention mask, to produce $K + 1$ embeddings; i.e., one for each of the K parts and the foreground ($f_k; k \in \{f, 1, \dots, K\}$). In addition to these part-based embeddings, two other global features are extracted; one by concatenating all embeddings of K parts ($f_c = \text{concat}(f_1, \dots, f_K)$), and the other by performing the global average pooling on $G(f_g)$. Finally, the Global identity Local triplet (GiLt) loss, introduced in [35], is employed as the re-identification objective to train our model. GiLt is a re-identification loss designed especially for part-based ReID methods, that is robust to occluded and non-discriminative body parts:

$$\ell_{reid} = \ell_{GiLt}(f_g, f_c, f_f, f_1, \dots, f_K) \quad (2)$$

The overall training objective of BPBReID is therefore the sum of two losses: the GiLt loss and the part prediction loss.

Finally, at inference, the distance between two person's images "q" and "g" is computed as the average distance of their mutually visible body parts as

$$\text{dist}^{qg} = \frac{\sum_{i \in \{f, 1, \dots, K\}} (v_i^q \cdot v_i^g \cdot \text{dist}_{eucl}(f_i^q, f_i^g))}{\sum_{i \in \{f, 1, \dots, K\}} (v_i^q \cdot v_i^g)}, \quad (3)$$

where $v_i^{q|g}$ is the visibility score of body part i and dist_{eucl} is the Euclidean distance between two features. A low distance dist^{qg} between two person's images corresponds to a high similarity.

For more information about the BPBReID baseline, refer to [35]. In this work, the the aforementioned model is utilized to improve learning performance and extract richer player embeddings through the addition of a new role classification head and two new training objectives.

3.1.2 Role Classification. In the proposed method, a role classification head is added to the Re-Identification model. This new head is a fully connected layer for multi-class classification purposes. It is designed to classify individuals in soccer videos into four classes: player, goalkeeper, referee, and staff. The role classification head contributes to better representations and adds new semantics to the embeddings. The Focal loss (ℓ_{focal}) [25] is employed to train it; because of the imbalance in the data between the player class and other classes. The role classification loss is denoted as ℓ_{role} and is applied on the foreground embedding.

$$\ell_{role} = \ell_{focal}(f_f), \quad (4)$$

3.1.3 Team Affiliation. An additional loss is added to the Re-Identification model. It aims at clustering players' embeddings according to their team affiliation. This loss function brings the embeddings of players from a same team closer to each other and pulls away the ones from distinct. This feature organization strategy is complementary to the ReID objective, which pulls together the embeddings from the same person and pulls away the embeddings of different persons. The triplet loss [18] is employed as the team affiliation loss. It is defined as

$$\ell_{team} = \ell_{tri}(f_f), \quad (5)$$

where players from the same team are assigned the same team ID and are considered as positive samples, while players from different teams are considered as negative samples. Last, but not least, this team affiliation loss is only applied to i) foreground embedding and ii) training samples having the "player" role; therefore, excluding goalkeepers, referees, and staff.

The proposed method enjoys the advantage of not restricting it to some predefined classes/teams, and therefore can be applied to any match with unseen teams. At the inference phase, a clustering algorithm with two clusters is conducted on the player's embeddings to assign their team label

3.1.4 Overall Training Procedure. The final loss of the model is a combination of all the previously mentioned losses, and is defined as

$$\ell_{total} = \lambda_{pa}\ell_{pa} + \lambda_{reid}\ell_{reid} + \lambda_{team}\ell_{team} + \lambda_{role}\ell_{role}, \quad (6)$$

where ℓ_{total} is the total loss of the re-identification network, and λ_{pa} , λ_{reid} , λ_{team} , and λ_{role} are hyperparameters that specify the scaling factors for each loss.

3.2 PRT-Track

In this section, the utilization of part-based player representation in supporting player tracking is described. The tracking method used as a baseline is first detailed, followed by an explanation of how both its online and offline modules are adapted to leverage the new part-based re-identification model.

3.2.1 StrongSORT Baseline. For multi-person tracking in soccer videos, the Strong-SORT [12], a recent method that uses an extended version of the Kalman filter to predict bounding boxes in the next frame, and a re-identification model to extract appearance features is employed. It generates a cost matrix using the IoU and the appearance similarity between the Kalman filter predictions and current detections. Linear assignment is then used in an online fashion to associate detections from the new frame with previous tracklets. The method also includes two lightweight post-processing modules; Appearance-Free Link (AFLink) and Gaussian-Smoothed Interpolation (GSI). In fact, AFLink performs global association with spatio-temporal information of tracklets instead of their appearance, while GSI is a Gaussian-smoothed interpolation that relieves missing detections.

3.2.2 Online Tracking Module. Our proposed PRTreID model is integrated into Strong-SORT by replacing its global ReID features with our part-based features. The detection-to-detection, detection-to-tracklet, and tracklet-to-tracklet ReID distances are computed

using Eq.(3). An exponential moving average (EMA) is applied to update the part-based features of the tracklet as

$$e_k^t = \alpha e_k^{t-1} \cdot v_k^{t-1} + (1 - \alpha) f_k^t \cdot v_k^t, \forall k \in \{f, 1, \dots, K\}, \quad (7)$$

where e_k^t and f_k^t are the appearance features of the k -th body part of the tracklet and the matched detection, respectively. Parameter α is the momentum term and v_k^t is the visibility score of the k -th part of the body, where $k \in \{f, 1, \dots, K\}$. The visibility score v_k^{t-1} of body part k of a tracklet at time $t-1$ is set to 1 if the corresponding body part is visible in at least one of the detections that are part of the tracklet.

3.2.3 Offline Post-Processing Module. The output of the first online tracking part consists of a set of short tracklets that need further processings to be merged into long tracks. To address this problem, an additional *offline* post-processing step based on part-based features is proposed to merge tracklets based on their underlying appearance in order to form long trajectories. To this end, a part-based appearance cost matrix for tracklets is build as

$$A_{ij} = \begin{cases} +\infty & i = j \\ dist_{total}^{ij} & \text{otherwise,} \end{cases} \quad (8)$$

where $dist^{ij}$, the distance between tracklet i and j , was introduced in Eq.(3). The matrix $A \in \mathbb{R}^{M \times M}$ is the appearance cost matrix and M is the number of tracklets obtained from the first online tracking step. The linear assignment problem minimizing the sum of matching costs is then solved by the Hungarian algorithm. This post-processing step aims at improving the results and lower the identity switches. The output of this step is a set of complete trajectories, where each trajectory corresponds to a unique identity that has been tracked throughout the video sequence. The proposed part-based tracking method is called *PRT-Track*.

Finally, it is worth mentioning that the team and role information are not explicitly used in the association stage. However, these details are implicitly encoded in the re-identification embeddings, owing to the multi-task training of the model. As a result, individuals belonging to different teams or having different roles are well-separated in the embedding space, with distances greater than the assigned association threshold, to prevent wrong associations.

Table 1: Some numbers for the proposed SoccerNet tracking-based re-identification dataset.

subset	# ids	# images	# cameras
train	1352	16217	57
query	1146	3432	49
gallery	1146	13719	49

4 EXPERIMENTS

In this section, first, some details on the datasets, evaluation metrics, and implementation setups are given. Next, the detailed results of the proposed re-identification, team affiliation, role classification,

Table 2: Comparison of the proposed PRTreID method with some existing re-identification methods on the test set of our proposed re-identification dataset. R1 denotes rank-1 accuracy. Only PRTreID is designed to address all three tasks. PRTreID_{team}/PRTreID_{role} refers to a variant of PRTreID where only the team/role objective is used at training.

Method	Re-ID		Team Aff		Role Cls	
	mAP	R1	mAP	R1	Acc	Prec
BoT [28]	62.63	80.24	-	-	-	-
PCB [39]	63.61	80.39	-	-	-	-
BPBreID [35]	71.43	89.31	-	-	-	-
PRTreID _{team}	-	-	91.48	96.68	-	-
PRTreID _{role}	-	-	-	-	93.64	71.45
PRTreID	72.59	89.57	92.89	97.60	94.27	74.36

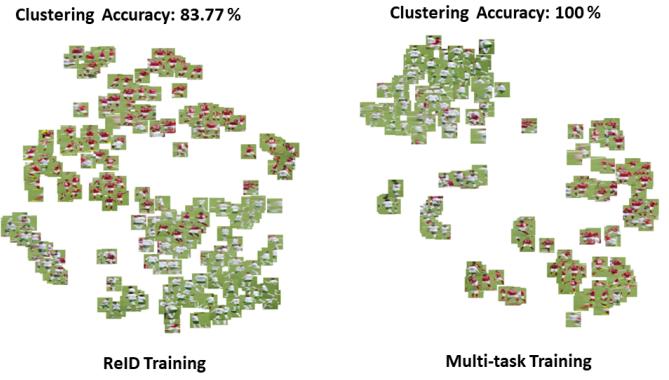


Figure 3: t-SNE visualization of player embeddings in a 2D space for a specific video, with and without multi-task training of the model. It can be observed that proposed multi-task model improves the team players clustering.

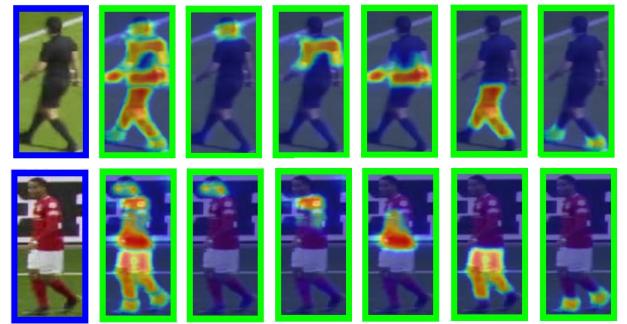


Figure 4: Visualizations of two images from proposed ReID dataset and their attention maps of the foreground and each body part.

and tracking methods are presented and analyzed. Finally, the ablation studies and qualitative assessments are discussed in order to further validate the effectiveness of the proposed method.

Table 3: Comparison of the overall tracking performance of the proposed part-based tracking method (PRT-Track) with recent tracking methods, from the 2022 SoccerNet Tracking challenge, on the test set of the SoccerNet-Tracking dataset using oracle detections and detections from YOLOv8. (The symbol \dagger indicates that the results are reported from [10]).

Oracle detections using ground truth							
Method	Setup	HOTA \uparrow	DetA \uparrow	AssA \uparrow	MOTA \uparrow	IDF1 \uparrow	IDs \downarrow
DeepSORT † [50]	w/ GT	69.52	82.62	58.66	94.84	-	-
ByteTrack † [56]	w/ GT	71.5	84.34	60.71	94.57	-	-
OC-SORT[4]	w/ GT	80.94	97.81	66.98	96.76	74.79	6079
StrongSORT[12]	w/ GT	83.75	95.08	73.78	94.67	79.13	2815
StrongSORT++	w/ GT	84.08	95.07	74.36	94.62	79.76	2619
CBiouU[52]	w/ GT	89.20	99.40	80.00	99.40	86.10	-
PRT-Track	w/ GT	90.77	99.83	82.53	98.66	88.47	3355
YOLOv8 detections							
Method	Setup	HOTA \uparrow	DetA \uparrow	AssA \uparrow	MOTA \uparrow	IDF1 \uparrow	IDs \downarrow
DeepSORT † [50]	w/o GT	36.63	40.02	33.76	33.91	-	-
FairMOT † [57]	w/o GT	43.91	46.31	41.77	50.69	-	-
ByteTrack † [56]	w/o GT	47.22	44.49	50.25	31.74	-	-
OC-SORT[4]	w/o GT	54.60	63.47	47.07	76.18	62.52	3593
StrongSORT[12]	w/o GT	54.86	62.19	48.79	74.52	65.1	2178
StrongSORT++	w/o GT	56.21	62.89	50.27	75.02	66.53	2106
PRT-Track	w/o GT	59.77	61.09	58.55	73.07	74.44	1428

Table 4: Comparison of the post-processing techniques used by StrongSORT++ and our TrackMerge module (w/wo part-based features).

Post-processing method	HOTA	Det A	Ass A
StrongSORT++	84.08	95.07	74.36
TrackMerge + global features	85.7	95.82	77.25
TrackMerge + part-based features	90.84	99.82	82.03

4.1 Datasets

4.1.1 SoccerNet Tracking. SoccerNet-Tracking [10] is a publicly available dataset that contains player-tracking data from soccer matches. The dataset consists of 100 video clips, each 30 seconds long (25 frame per second), captured from the main camera. The dataset includes 57 videos from 3 games for the training set and 49 videos from 3 games for the test set. The objects in the dataset belong to the following classes: player team left, player team right, goalkeeper team left, goalkeeper team right, main referee, side referee, staff, and ball.

4.1.2 Re-Identification Dataset. is used to train and evaluate the PRTreID model. A re-identification dataset was created by cropping objects (excluding the ball) from videos in the SoccerNet-Tracking dataset, using the same train/test splits. This resulted in a large number of detections per identity, much more than a standard re-identification dataset, with many of these detections being redundant and providing no additional information as they are from consecutive frames. To reduce the number of detections per identity, a uniform sampling approach was used along the frames. Table 1 presents the characteristics of the ReID dataset generated from SoccerNet-Tracking.

Finally, it is worth mentioning that the SoccerNet ReID [15] dataset was not suitable for our research on video analysis and tracking, since it only includes cross-camera images of the same action without videos or cross-action annotations. This limits the ability to conduct large-scale experiments on team affiliation with multiple crops from two teams within a game video.

4.2 Evaluation Metrics

The performance of a given person re-identification method is typically evaluated as a retrieval problem: given an image of an individual of interest (the "query") and a database of image crops of various persons (the "gallery"), all gallery samples are ranked according to their distance to the query, such that gallery samples with the same identity as the query are at the top of the ranking. In line with standard evaluation practices in ReID research, the cumulative matching characteristics (CMC) at Rank-1 and the mean average precision (mAP) are used to assess the quality of query-gallery rankings. The same two metrics are also employed to evaluate team affiliation, using the same definition as in re-identification, reflecting the model's accuracy in retrieving images from the gallery set with the same *team ID* as the query sample. For the role classification head, accuracy and precision are used as evaluation metrics. In the tracking part, HOTA, IDF1, MOTA, AssA, ID Switches (IDs), and DetA are reported. Additionally, a two-class clustering is performed for the team affiliation part, and the model's accuracy in predicting team labels is reported.

4.3 Implementation Details

For the re-identification part, the BPBReID model was employed with the HRNet-W32 (HR) [37] backbone, which was pre-trained on the ImageNet dataset. The model was trained using weights from the Market 1501 [58] dataset and the number of parts (K) was

Table 5: Ablation for proposed PRTreID method to validate the benefits from multi-task learning.

#	Loss			Re-Identification		Team Affiliation		Role Classification	
	ReID	Team	Role	mAP ↑	Rank-1 ↑	mAP ↑	Rank-1 ↑	Accuracy ↑	Precision ↑
1	✓			71.43	89.31	87.12	97.43	-	-
2		✓		12.34	4.28	91.48	96.68	-	-
3			✓	10.29	4.28	55.38	50.96	93.64	71.45
4	✓	✓		72.53	89.04	89.03	97.40	-	-
5	✓		✓	72.78	89.34	78.51	97.33	93.27	68.79
6	✓	✓	✓	72.59	89.57	92.89	97.60	94.27	74.36

Table 6: Ablation study for team affiliation by applying a two-cluster Kmeans algorithm on foreground embedding of players for each video.

	ReID	ReID + Team	ReID + Team + Role
Accuracy	91.87%	95.17% (+3.3)	95.6% (+3.73)

set to 5 (head, up torso, low torso, legs, and feet) to achieve the best results. For the tracking part, the Strong-SORT tracker was utilized, and its feature extractor was replaced with the proposed PRTreID method. The training configurations were the same as in a previous study [35]. All images were resized to 384×128 and were first augmented with random cropping and 10-pixel padding, followed by random erasing at a 0.5 probability. All networks were trained end-to-end for 120 epochs using the Adam optimizer on a single NVIDIA GeForce RTX 3090 GPU. The learning rate was increased linearly from 3.5×10^{-5} to 3.5×10^{-4} after 10 epochs and was then decayed to 3.5×10^{-5} and 3.5×10^{-6} at the 40th and 70th epochs, respectively. A new batch sampler with a batch size of 32 was used for training. Each batch contained 4 identities from players of the left team, 4 identities from players of the right team, and 3 identities from other roles, with 4 images sampled for each identity which are all from a specific video. For the team affiliation objective, only identities from the player role were used, resulting in a batch size of 24. The triplet loss margin defined in Eq.(2) was set to 0.3 and 0.05 for the re-identification and team affiliation objectives, respectively. The hyperparameters specified in Eq.(6) were optimized by testing various values and selecting the optimal ones. Empirically, the values were set to 0.3, 1, 0.1, and 1.5 for λ_{pa} , λ_{reid} , λ_{team} , and λ_{role} respectively.

4.4 Experimental Results

4.4.1 Re-Identification Results. Table 2 shows the performance of the proposed PRTreID model on the test set (query/gallery) of the proposed ReID dataset in comparison to some existing re-identification baselines. As this table shows, the proposed method outperforms other methods, despite the heavy occlusions in some images and the similar appearance of players from the same team. Additionally, PRTreID also performs well in terms of team affiliation and role classification. It should be noted that other methods do not support the other two tasks, which renders our model unique in its ability to perform multiple tasks simultaneously. We could not compare the team affiliation and role classification performance

of PRTreID with other works, because none provided open-source code allowing us to compute their performance on our dataset.

4.4.2 Tracking Results. In Table 3, the performance comparison of the proposed tracking method to existing tracking methods on the test set videos of the SoccerNet-Tracking dataset is given. Since our work focuses on feature learning and association for tracking, the results are compared to the trackers from the 2022 SoccerNet Tracking challenge. The proposed PRT-Track, with a part-based post-processing step, outperforms other methods on the SoccerNet-Tracking dataset, both with and without ground-truth detections. This demonstrates the importance of rich representations for tracking. Although it can be challenging to discriminate players by their appearance features in soccer due to the similar appearance of players and the distance between the main camera and the players, the proposed part-based tracking method achieves state-of-the-art performance and outperforms all previous methods.

4.5 Ablation Studies

To validate the effectiveness of each component of the PRTreID model, comprehensive ablation studies were conducted. Table 5 shows the impact of the two additional learning objectives on performance. As reported in the table, adding the team affiliation and role classification objectives during training improves the performance on both tasks as well as the re-identification performance. Additionally, we evaluate team affiliation performance at inference using a K-means clustering with two clusters on the foreground embedding of players for each video. After clustering, a team label was assigned to each player. The results in Table 6 demonstrate the improvements of multi-task training on the overall model performance; the joint model outperforms all other partial models on all tasks. Finally, Table 4 reports a comparison of the proposed appearance-based tracklet merging step, with those of StrongSORT++ (AFLink and GSI). The results show that the proposed tracklet merging with global features outperforms the StrongSORT++. Also, the tracklet merging with part-based features achieves the best results.

4.6 Qualitative Assessment

To further validate the effectiveness of the proposed method, some qualitative assessments of the model's performance are conducted. Figure 4 illustrates two images from the proposed ReID dataset and their respective output attention maps. Each attention map highlights a different part; five parts of the body and the foreground. As illustrated here, the part attention module is still performing well when trained with other objectives; as it correctly extracts

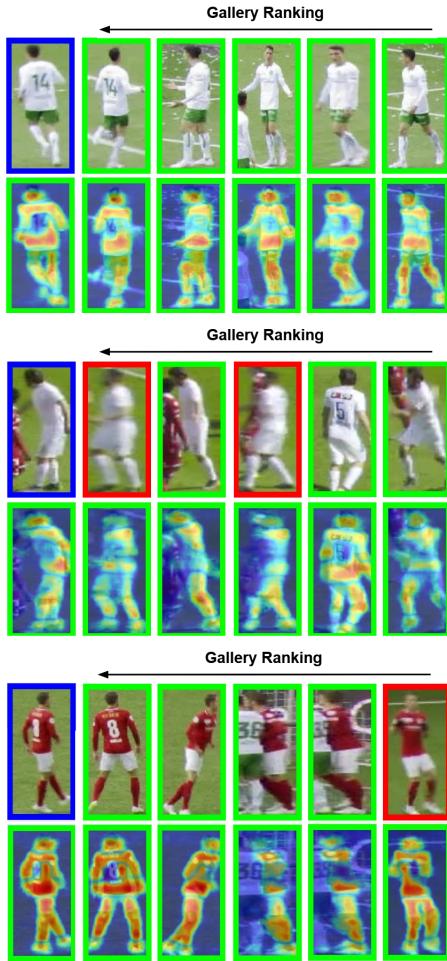


Figure 5: Visualization of three images in the query set and their top-5 retrieved images from the gallery set of proposed ReID dataset, along with foreground masks. The blue color represents the query sample, while the red color indicates a wrongly retrieved image and the green color indicates a correctly retrieved image from the gallery.

body parts heatmaps while excluding the background. Figure 4 also shows three samples of query set and their corresponding top 5 retrieved gallery images from the proposed ReID dataset, along with their foreground attention maps. The red color indicates that the retrieved image has a different identity than the query image. This visualization demonstrates that in occluded scenarios, PRTreID can properly extract the attention maps of the target person and retrieve the correct images.

Figure 3 shows qualitative results of the learned embedding space, that is reduced to 2 dimensions using the t-SNE algorithm. As shown in the figure, the embeddings of different teams are much better distributed in the embedding space when the model is trained with all three objectives than with just the re-identification objective. In summary, these qualitative results demonstrate that

using multi-task learning to train the model leads to richer and more discriminative representations.

4.7 Computational Efficiency Discussion

For the readers that are concerned about the computational efficiency of our PRTreID model: our model has two main components: a backbone and a ReID head comprising an attention mechanism for part-based feature extraction and a classifier to perform role classification. Since the ReID head induces a very small computational overhead, the overall inference time of PRTreID will be almost the same as the backbone itself. In our work, we use a HRNet-W32 for its high-resolution feature maps that are beneficial to ReID, but any other smaller backbone can be used depending on the speed requirement of the downstream application.

5 CONCLUSION AND FUTURE WORK

In this work, a novel person representation method was proposed for sports videos. A body part-based person re-identification model was utilized for extracting person embeddings. To extract more meaningful embeddings, two new objectives were added to the re-identification model for team affiliation and role classification purposes. Experimental results showed that jointly training the re-id model for re-identification, team affiliation, and role classification led to rich representations, which were then used in tracking persons as a downstream task of re-identification. Additionally, a post-processing step was proposed using part-based features to further improve the tracking results. The experimental results showed the superiority of the proposed method over the existing state-of-the-art methods. In this study, our evaluation was limited to the SoccerNet tracking dataset, as we couldn't find any other open-source datasets providing annotations for re-identification, team affiliation, and role classification all at once. In future work, we plan to integrate a Jersey Number Recognition head into our sports person representation model. This decision is motivated by the challenges in team sports like ice hockey, where identical kits covering a large portion of the player's body can compromise player re-identification.

ACKNOWLEDGMENTS

The authors would like to acknowledge the High Performance Center (HPC) of Sharif University of Technology for providing the computational resources for this project. The authors would also like to thank Sportradar AG for sponsoring this work and for its commitment to the scientific research, making it possible for us to make significant contributions to the sports technology field.

REFERENCES

- [1] Reza Afrouzian, Hadi Seyedarabi, and Shohreh Kasaei. 2016. Pose estimation of soccer players using multiple uncalibrated cameras. *Multimedia Tools and Applications* 75 (2016), 6809–6827.
- [2] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. 2022. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651* (2022).
- [3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. 2016. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*. IEEE, 3464–3468.
- [4] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. 2023. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9686–9696.

- [5] Fan Chen and Christophe De Vleeschouwer. 2011. Formulating team-sport video summarization as a resource allocation problem. *IEEE Transactions on Circuits and Systems for Video Technology* 21, 2 (2011), 193–205.
- [6] Jianhui Chen, Hoang M Le, Peter Carr, Yisong Yue, and James J Little. 2016. Learning online smooth predictors for realtime camera planning using recurrent decision trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4688–4696.
- [7] Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. 2018. A Bottom-Up Approach Based on Semantics for the Interpretation of the Main Camera Stream in Soccer Games. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2018), 1846–184609. <https://api.semanticscholar.org/CorpusID:52027492>
- [8] 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 (2022). <https://api.semanticscholar.org/CorpusID:249894209>
- [9] Anthony Cioppa, Adrien Deliège, Maxime Istasse, Christophe De Vleeschouwer, and Marc Van Droogenbroeck. 2019. ARTHuS: Adaptive real-time human segmentation in sports through online distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.
- [10] Anthony Cioppa, Silvio Giancola, Adrien Deliège, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. 2022. SoccerNet-Tracking: Multiple Object Tracking Dataset and Benchmark in Soccer Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3491–3502.
- [11] Abdulrahman Darwish and Tallal El-Shabrway. 2022. STE: Spatio-Temporal Encoder for Action Spotting in Soccer Videos. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*. 87–92.
- [12] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. 2023. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia* (2023).
- [13] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. 2020. Pose-guided visible part matching for occluded person ReID. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 11741–11749. <https://doi.org/10.1109/CVPR42600.2020.01176> arXiv:2004.00230
- [14] Seyed Abolfazl Ghasemzadeh, Gabriel Van Zandycke, Maxime Istasse, Niels Sayze, Amirafshar Moshtaghpour, and Christophe De Vleeschouwer. 2021. Deep-SportLab: a Unified Framework for Ball Detection, Player Instance Segmentation and Pose Estimation in Team Sports Scenes. *ArXiv* abs/2112.00627 (2021). <https://api.semanticscholar.org/CorpusID:244773081>
- [15] Silvio Giancola, Anthony Cioppa, Adrien Deliège, Floriane Magera, Vladimir Somers, Le Kang, Xin Zhou, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, et al. 2022. SoccerNet 2022 challenges results. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*. 75–86.
- [16] Xin He. 2022. Application of deep learning in video target tracking of soccer players. *Soft Computing* 26, 20 (2022), 10971–10979.
- [17] Jan Held, Anthony Cioppa, Silvio Giancola, Abdullah Hamdi, Bernard Ghanem, and Marc Van Droogenbroeck. 2023. VARS: Video Assistant Referee System for Automated Soccer Decision Making from Multiple Views. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2023), 5086–5097. <https://api.semanticscholar.org/CorpusID:258048692>
- [18] Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12–14, 2015. Proceedings* 3. Springer, 84–92.
- [19] Samuel Hurault, Coloma Ballester, and Gloria Haro. 2020. Self-supervised small soccer player detection and tracking. In *Proceedings of the 3rd international workshop on multimedia content analysis in sports*. 9–18.
- [20] Maxime Istasse, Julien Moreau, and Christophe De Vleeschouwer. 2019. Associative embedding for team discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.
- [21] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmén, Mustafa E Kamasak, and Mubarak Shah. 2018. Human semantic parsing for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1062–1071.
- [22] Amit Kumar KC, Laurent Jacques, and Christophe De Vleeschouwer. 2016. Discriminative and efficient label propagation on complementary graphs for multi-object tracking. *IEEE transactions on pattern analysis and machine intelligence* 39, 1 (2016), 61–74.
- [23] Maria Koshkina, Hemanth Pidaparthi, and James H Elder. 2021. Contrastive learning for sports video: Unsupervised player classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4528–4536.
- [24] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. 2021. Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association. *IEEE Transactions on Intelligent Transportation Systems* 23, 8 (2021), 13498–13511.
- [25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.
- [26] Hengyue Liu and Bir Bhanu. 2019. Pose-guided R-CNN for jersey number recognition in sports. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.
- [27] Wei-Lwun Lu, Jo-Anne Ting, James J Little, and Kevin P Murphy. 2013. Learning to track and identify players from broadcast sports videos. *IEEE transactions on pattern analysis and machine intelligence* 35, 7 (2013), 1704–1716.
- [28] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 0–0.
- [29] Gerard Maggiolini, Adnan Ahmad, Jinkun Cao, and Kris Kitani. 2023. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. *arXiv preprint arXiv:2302.11813* (2023).
- [30] Adrien Maglo, Astrid Orcesi, and Quoc-Cuong Pham. 2022. Efficient tracking of team sport players with few game-specific annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3461–3471.
- [31] Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. 2023. SoccerNet-Caption: Dense Video Captioning for Soccer Broadcasts Commentaries. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2023), 5074–5085. <https://api.semanticscholar.org/CorpusID:258049025>
- [32] Nima Najafzadeh, Mehran Fotouhi, and Shohreh Kasaei. 2015. Multiple soccer players tracking. In *2015 The international symposium on artificial intelligence and signal processing (AISP)*. IEEE, 310–315.
- [33] Atom Scott, Ikuma Uchida, Masaki Onishi, Yoshinari Kameda, Kazuhiro Fukui, and Keisuke Fujii. 2022. SoccerTrack: A dataset and tracking algorithm for soccer with fish-eye and drone videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3569–3579.
- [34] Arda Senocak, Tae-Hyun Oh, Junsik Kim, and In So Kweon. 2018. Part-based player identification using deep convolutional representation and multi-scale pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1732–1739.
- [35] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. 2023. Body part-based representation learning for occluded person Re-Identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1613–1623.
- [36] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. 2018. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*. 402–419.
- [37] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5693–5703.
- [38] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2017. Beyond Part Models: Person Retrieval with Refined Part Pooling (and a Strong Convolutional Baseline). *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11208 LNCS (nov 2017), 501–518. arXiv:1711.09349 <http://arxiv.org/abs/1711.09349>
- [39] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*. 480–496.
- [40] Mostafa Tavassoli Pour, Mahmood Karimian, and Shohreh Kasaei. 2013. Event detection and summarization in soccer videos using bayesian network and copula. *IEEE Transactions on circuits and systems for video technology* 24, 2 (2013), 291–304.
- [41] Rajkumar Theagarajan and Bir Bhanu. 2020. An automated system for generating tactical performance statistics for individual soccer players from videos. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 2 (2020), 632–646.
- [42] Graham Thomas, Rikke Gade, Thomas B Moeslund, Peter Carr, and Adrian Hilton. 2017. Computer vision for sports: Current applications and research topics. *Computer Vision and Image Understanding* 159 (2017), 3–18.
- [43] Xiaofeng Tong, Jia Liu, Tao Wang, and Yimin Zhang. 2011. Automatic player labeling, tracking and field registration and trajectory mapping in broadcast soccer video. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2, 2 (2011), 1–32.
- [44] Ikuoma Uchida, Atom Scott, Hidehiko Shishido, and Yoshinari Kameda. 2021. Automated Offside Detection by Spatio-Temporal Analysis of Football Videos. In *Proceedings of the 4th International Workshop on Multimedia Content Analysis in Sports*. 17–24.
- [45] Renaud Vandeghen, Anthony Cioppa, and Marc Van Droogenbroeck. 2022. Semi-Supervised Training to Improve Player and Ball Detection in Soccer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2022), 3480–3489. <https://api.semanticscholar.org/CorpusID:248178107>
- [46] Kanav Vats, Mehrnaz Fani, David A Clausi, and John Zelek. 2021. Multi-task learning for jersey number recognition in ice hockey. In *Proceedings of the 4th International Workshop on Multimedia Content Analysis in Sports*. 11–15.
- [47] Kanav Vats, William McNally, Pascale Walters, David A Clausi, and John S Zelek. 2022. Ice hockey player identification via transformers and weakly supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1–8.

- Pattern Recognition*. 3451–3460.
- [48] Kanav Vats, Pascale Walters, Mehrnaz Fani, David A Clausi, and John S Zelek. 2023. Player tracking and identification in ice hockey. *Expert Systems with Applications* 213 (2023), 119250.
- [49] HongXia Wang, Xiang Chen, and Chun Liu. 2021. Pose-guided part matching network via shrinking and reweighting for occluded person re-identification. *Image and Vision Computing* 111 (2021), 104186.
- [50] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*. IEEE, 3645–3649.
- [51] Yunjin Xu, LiaoYing Zhao, and Feiwei Qin. 2021. Dual attention-based method for occluded person re-identification. *Knowledge-Based Systems* 212 (2021), 106554.
- [52] Fan Yang, Shigeyuki Odashima, Shoichi Masui, and Shan Jiang. 2023. Hard to track objects with irregular motions and similar appearances? make it easier by buffering the matching space. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 4799–4808.
- [53] Yukun Yang, Ruiheng Zhang, Wanneng Wu, Yu Peng, and Min Xu. 2021. Multi-camera sports players 3d localization with identification reasoning. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 4497–4504.
- [54] Gabriel Van Zandycke, Vladimir Somers, Maxime Istasse, Carlo Del Don, and Davide Zambrano. 2022. DeepSportRadar-v1: Computer Vision Dataset for Sports Understanding with High Quality Annotations. *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports* (2022). <https://api.semanticscholar.org/CorpusID:251623078>
- [55] Ruiheng Zhang, Lingxiang Wu, Yukun Yang, Wanneng Wu, Yueqiang Chen, and Min Xu. 2020. Multi-camera multi-player tracking with deep player identification in sports video. *Pattern Recognition* 102 (2020), 107260.
- [56] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. 2022. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*. Springer, 1–21.
- [57] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. 2021. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision* 129 (2021), 3069–3087.
- [58] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*. 1116–1124.