IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 6, 2020, accepted October 20, 2020, date of publication October 26, 2020, date of current version November 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3033580

Detection of Ice Hockey Players and Teams via a Two-Phase Cascaded CNN Model

TIANXIAO GUO[1], KUAN TAO[2], QINGRUI HU[1], AND YANFEI SHEN[2]
[1]School of Sports Science, Beijing Sport University, Beijing 100084, China
[2]School of Sports Engineering, Beijing Sport University, Beijing 100084, China

Corresponding authors: Kuan Tao (taokuan@bsu.edu.cn) and Yanfei Shen (syf@bsu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 72071018 and Grant 11901037, in part by the National Key Research and Development Program of China under Grant 2018YFC2000600, and in part by the Fundamental Research Funds for the Central Universities, Beijing Sport University, China, under Grant 2020042.

ABSTRACT The accurate detection of ice hockey players and teams during a game is crucial to the tracking of individual players on the rink and team tactical decision making and is therefore becoming an important task for coaches and other analysts. However, hockey is a fluid sport due to its complex situation and the frequent substitutions by both teams, resulting in the players taking various postures during a game. Few player detection models from basketball and soccer take these characteristics into account, especially for team detection without prior annotations. Here, a two-phase cascaded convolutional neural network (CNN) model is designed for the detection of individual ice hockey players, and the jersey color of the detected players is extracted to further identify team affiliations. Our model filters most of the disturbing information, such as the audience and sideline advertising bars, in Phase I and refines the detection of the targeted players in Phase II, resulting in an accurate detection with a precision of 98.75% and a recall of 94.11% for individual players and an average accuracy of 93.05% for team classification with a self-built dataset of collected images from the 2018 Winter Olympics. The results for the regular season games of the 2019-2020 National Hockey League (NHL) covering all 31 teams are also presented to show the robustness of our model. Compared to state-of-the-art approaches, our player detection model achieves the highest accuracy with the self-built dataset.

INDEX TERMS Player detection, team detection, player tracking data, ice hockey.

I. INTRODUCTION

Ice hockey is a popular team sport in North America and Northern Europe; it is described as a fluid sport [1], [2], with players frequently substituting on and off the rink without timeouts. Although hockey games are fascinating to watch, the use of analytical approaches to assess player performance is still at an early age due to the games' low scores [2] and complex dynamics [3], [4]. Evaluating the performance of individual players and their contribution to the overall performance of the team [5], [6] is a major challenge in the field of sports analysis. Several metrics have been proposed for performance analysis in different team sports, e.g., ''Expected-Point-Value'' in basketball [7], [8] and ''Expected-Goal-Value'' in soccer [9] and American football [10]. In professional ice hockey leagues, such as the National Hockey League (NHL) in North America, winning the final championship is the greatest honor and goal of all players and teams. As a result, a number of natural concerns arise, such as how to assemble a winning team with players maximizing their capacity and how to design the most effective tactics after comparing different offensive and defensive formats. The key factor to answer these questions is to take advantage of enormous data.

As ice hockey is considered spatiotemporally complicated, the most valuable data are trajectory tracking data, which encode vital information on the actions and intentions of the players [3] and could be analyzed in multiple ways, such as visualization of player trajectories [11], heat map analyses [12], event recognition [4], [12], [13] and performance assessments [14], [15]. Many models use deep learning diagrams to analyze player and team dynamics based on trajectory data in different team sports. Le *et al.* [7] utilized deep imitation learning to generate alternative strategies for defensive teams

The associate editor coordinating the review of this manuscript and approving it for publication was Naveed Akhtar.

in soccer. Another group carried out similar work in the same sport [16]. Miller and Bornn [17] analyzed National Basketball Association (NBA) team strategies through probabilistic theme modeling that captured the structure of player trajectories. Wang *et al.* and Mehrasa *et al.* [3], [18] used a convolutional neural network (CNN) to classify offensive plays in basketball games, while Tian *et al.* [19] distinguished defensive patterns through a number of machine learning models based on team trajectory data.

One of the most essential steps in collecting trajectory data is the detection of the targeted players. Although deep learning approaches [20]–[26] have been widely applied in the detection of objects, the detection of players is more difficult because of complex game dynamics and the sparse distribution of players, as seen from broadcasted videos. A variety of player detection studies have been introduced via nonintrusive methods. Lara *et al.* [27] used two calibrated cameras to capture the location of individual tennis players to assist in an auxiliary training medium. Lu *et al.* [28] and Parisot and De Vleeschouwer [29] achieved player detection from broadcasted images via a single calibrated camera.

However, despite numerous studies on trajectory tracking models, few have focused on ice hockey. The main reason is because traditional models fail to recover several common characteristic features of ice hockey games, such as severe occlusions and the large number of physical confrontations between players [30]–[32]. In addition, due to their high movement speeds and abrupt direction changes, ice hockey players always produce body positions with stretched aspect ratios. These features challenge the efficiency and accuracy of the detection of individual players and teams, which is regarded as the most vital component of analyzing trajectory data.

In this paper, a two-phase cascaded CNN model is proposed for individual ice hockey player detection during ice hockey games. Phase I of the cascaded architecture roughly detects the targeted players by filtering most of the disturbing information, such as the audience and sideline advertising bars, while Phase II incorporates detailed information such as overlapping areas (occlusions) of body position caused by individual player movements and the uniform colors of the two different teams to further refine the results derived from the outputs of Phase I. An image collection from the 2018 Winter Olympics was constructed and divided into training (4048 samples) and testing (1341 samples) datasets. Then, the distribution of the aspect ratios of all players was calculated from the training data to derive a suitable bounding box using a deep learning framework to resolve the challenging situation when players exhibit a variety of postures. Following player detection, the regions containing the uniforms of the detected players are cropped, and the features of the distribution of the uniform color are represented through five color channels that are preliminarily divided according to the statistics of the uniform color features to recognize the team affiliation. Since prior knowledge of the uniform color features is obtained, the proposed method is able to classify the team affiliations without the need for extra annotations in the construction of the dataset.

The proposed method achieves high accuracy and recall for both individual player and team detection with the testing dataset. Meanwhile, the detection results for images from the 2019-2020 NHL regular season games covering all 31 teams also validate the robustness of our model. A comparison with several state-of-the-art object detection methods validates the effectiveness of our model in the detection of ice hockey players. Thus, our proposed two-phase cascaded CNN model is particularly designed to detect individual players and teams in ice hockey games, with the goal of tracking personal trajectory data to evaluate player performance and recognizing team offensive and defensive patterns to aid in decisive tactical decision making.

II. RELATED WORK

In this section, work related to the proposed cascaded CNN model for ice hockey player detection will be presented, including that on player detection and cascade-structured object detection.

A. PLAYER DETECTION

Player detection has gained great popularity as a specialized form of object detection in the sports field in recent years. The earliest methods used to achieve player detection were mostly based on single-feature extraction and background subtraction [33], [34]. For example, Liu *et al.* [33] combined background subtraction based on the dominant color with a Haar-like feature detector to obtain the locations of soccer players. This kind of approach is commonly effective for specific scenarios but vulnerable to complex backgrounds, such as those in basketball and ice hockey. By contrast, pedestrian detection algorithms applied to player detection tasks performed more robustly for a variety of complicated background noises and usually included feature selection and classification [35]–[37]. Mackowiak *et al.* [36] applied a histogram of oriented gradients (HOG)-based detector for feature extraction and a support vector machine (SVM) for classification to detect football players. Nevertheless, pedestrian detection-based method performed poorly in the rotation and stretching of objects. Following the substantial development of deep learning technology for object detection [38], CNN models have been applied to player detection tasks because of their strong feature extraction ability. Mrazova and Hrincar [39] proposed a CNN-based method to detect players from broadcasted video streams and achieved high precision with robustness to image transformation to some extent. Sorano *et al.* [40] utilized YOLOv3 [41] to detect football players and balls on the ground to accomplish further event recognition. CNN-based methods have numerous advantages for feature extraction, and cascaded CNNs are among the most effective algorithms for detecting players.

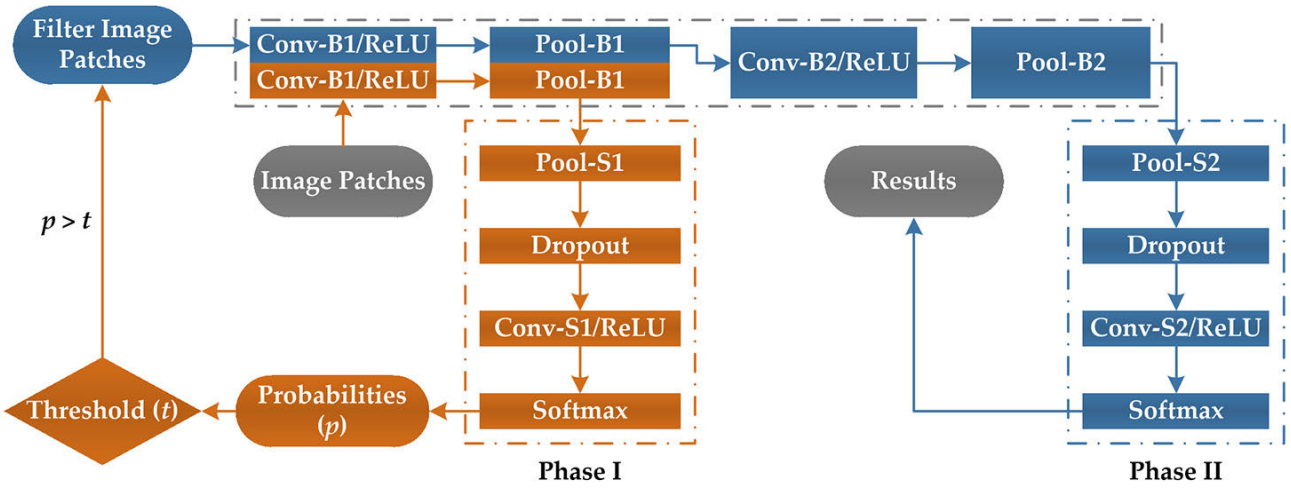


FIGURE 1. Architecture of the two-phase cascaded CNN model. The backbone network along with Phase I and II classification branches are encircled by a gray, red and blue dotted line, respectively. The suffixes "B1", "B2" and "S1", "S2" mean the convolution and pooling layer of Phase I (B1 and S1) or Phase II (B2 and S2) in both backbone network and classification phases, respectively.

B. CASCADE-STRUCTURED OBJECT DETECTION

Cascade architectures have been widely applied to multiple computer vision tasks [42]–[45]. For example, Gao *et al.* [43] proposed a cascaded boundary regression model to achieve temporal action recognition, and Xie *et al.* [45] proposed a cascaded scene classification method with a hybrid image representation that performed commendably in scene recognition and domain adaptation. Moreover, in the field of object detection, especially for one-class detection purposes, the hierarchical property of cascaded structures is beneficial to filtering out vast numbers of background regions. One classical approach [46] was to construct a cascaded detector with Haar-like features and Adaboost [47] for feature selection. Based on this pioneering work, Zhang *et al.* [48] presented the Multi-block Local Binary Pattern (MB-LBP) features to replace the Haar-like features for more effective feature extraction, which was inspiring in that better performance could be achieved by producing more discriminative representations and utilizing a cascaded architecture. Consequently, Li *et al.* [49] proposed a cascaded CNN model by integrating multiresolution stages to remove a majority of false positive samples in earlier stages. Compared with the abovementioned cascaded detector that utilizes assigned features, the cascaded CNN method integrates a CNN module for feature extraction to exclude redundant regions, enabling it to describe the features of objects accurately and robustly. Note that the cascaded CNN model was proposed above to perform one-class, multiobject, and variable-scale face detection, similar to the player detection task to some extent. For player detection, Lu *et al.* [28] applied a cascaded CNN model to basketball and soccer games to extract the players' spatial location information. Considering the complexity of ice hockey games, a cascaded CNN model was specifically designed in this paper to obtain the positions of ice hockey players.

III. MODEL FOR THE DETECTION OF ICE HOCKEY PLAYERS AND TEAM AFFILIATIONS

A. TWO-PHASE CASCADED CNN MODEL FOR ICE HOCKEY PLAYER DETECTION

Figure 1 shows the schematic diagram of our proposed two-phase cascaded neural network topology, which is categorized as two phases that branch from the backbone network: Phase I in red and Phase II in blue. With learning based on AlexNet [38] a rectified linear unit (ReLU) activation function and Dropout regularization [50], a very light architecture with a parameter memory of only 20.88 KB was designed, which is lightweight compared with the AlexNet model (the parameter memory of the AlexNet model with for an input image of size 224×224 is 238 MB). As shown in Figure 1, the two classification phases are trained separately. (1) To train Phase I, annotated player patches are inputted to the backbone network and processed with a convolution layer, Conv-B1, consisting of 16 filters of size 3×3, which is then followed by a ReLU activation layer to speed up the convergence of the learning rate and prevent the optimized function from becoming trapped in saddle points or local minima. Furthermore, a pooling layer, Pool-B1, of size 3×3 is used to subsample the output feature map from Conv-B1/ReLU to ensure the extraction of the most informative local features based on max-pooling criteria. Subsequently, the feature maps from the backbone network are compressed by layer Pool-S1 through average pooling, which could further extract characteristic features and smooth the feature maps. To avoid overfitting, Pool-S1 is subsequently followed by another layer, Dropout, that randomly drops units from the neural network. Moreover, a convolution layer, Conv-S1, equipped with a ReLU activation function is applied to produce 1×1 feature maps, which are then processed by a Softmax layer to show the likelihood that input patches will be recognized as a player. (2) The training in Phase II follows almost the same procedures as

Phase I, differing only in their inputs. Both targeted players and false positive samples are misclassified as layers in the results of Phase I; the latter are further trained as negative training samples in Phase II to develop the ability to recognize more difficult samples. Therefore, Phase II is particularly capable of classifying confusing background areas of the ground-truth players.

In both classification branches, a cross-entropy loss function that is commonly used in other deep learning frameworks [3], [28], [51], [52] is applied to attune the weights of the proposed model. The training set is denoted as $S = \left\{\left(x_{i,j}, y_{i,j}\right)\right\}, 1 \leq i \leq N, 1 \leq j \leq K$ with $x_{i,j} \in \mathbb{R}^d$ representing the feature map of the *ith* sample at the *jth* cascaded phases and $y_{i,j} \in \{0, 1\}$ standing for the binary label accordingly. The probability of predicting a positive sample is:

$$p_i^p = p_i\left(y_i = 1 \left| x_i, w\right.\right) = \prod_{j=1}^{K} p_{i,j}\left(y_{i,j} = 1 \left| x_{i,j}, w\right.\right), \quad (1)$$

where w are the weights of the model. Likewise, the probability of predicting a negative sample is:

$$p_i^n = p_i\left(y_i = 0 \left| x_i, w\right.\right) = 1 - \prod_{j=1}^{K} p_{i,j}\left(y_{i,j} = 1 \left| x_{i,j}, w\right.\right), \quad (2)$$

Therefore, the loss function is defined as:

$$L_P(w) = -\sum_{i=1}^{N}\left[y_i \log\left(p_i^p\right) + (1 - y_i)\log\left(p_i^n\right)\right]. \quad (3)$$

The proposed cascaded CNN model is trained with this loss function $L_p(w)$ using the stochastic gradient descent algorithm, which yields the lowest probability of predicting false positive samples, and targeted players or nonplayer elements are accurately classified. Based on the size of our dataset, the weights are updated over 37000 iterations, and the learning rate is set to 0.001.

One of the distinct features of our two-phase cascaded CNN model is that the two separate classification phases are optimized as a unified block. The goal of Phase I is to recognize the ground truth (usually annotated in the training set) as positive samples and randomly select other elements as negative samples; thus, the rough outputs consist of both player and nonplayer elements, while the false positive samples are labeled negative samples during the training of Phase II.

The layered design is of great advantage for simplifying the architecture over a conventional CNN and for adjusting the design of each layer by observing the performance hierarchically. The first phase eliminates most nonplayer elements and confirms the detection of players faithfully, and the second phase focuses on complicated false positive samples, such as the background of audience members and sideline cameras.

B. PARAMETER SETTINGS

Since ice hockey players exhibit different postures depending on real-time team decisions, the bounding box for targeted detection should be specifically adjusted in response to the overall statistics of the players in the rink from the perspectives of spatial scale and aspect ratio. Therefore, in addition to model parameters, we determined two crucial physical parameters related to the postures of the hockey players from various games by analyzing the training set.

1) PARAMETER 1: THE SIZE OF THE INPUT IMAGE PATCHES

When setting the parameters, the size of the input image patches is prioritized to obtain a suitable size for input image patches, as the number of layers in the neural network will increase exponentially for a large input image size, while small patches are insufficient for extracting discriminative features. Note that the aspect ratios of the ice hockey players are more widely distributed than those of the players of other popular team sports, such as basketball and soccer. During ice hockey game broadcasts, the photographic distance of the cameras is much closer to the ground than in soccer games, which results in a variety of size differences for the different players. Furthermore, due to the high player moving speeds and the number of physical confrontations during ice hockey games, a larger range of aspect ratios are presented than those in broadcasted basketball games. The sizes of all players in the training set were calculated, showing that the majority of the players' aspect ratios was equal to 1.65 (height/width). Given the comprehensive design of the architecture, the size of the input patches was determined to be 42×25.

2) PARAMETER 2 ZOOMING SCALES OF THE ORIGINAL IMAGE IN THE TESTING PROCEDURE

An adjusted sliding window strategy was implemented to classify the image patches in the testing procedure. Due to the filming angles and the locations of the broadcasting cameras and the players' locations in the rink, the sizes of the players vary from image to image. For instance, players in the neutral zone are larger than those in the defending or attacking zones. As mentioned in *Parameter 1*, the size of the input image patches was determined to be 42×25 according to the statistics of the aspect ratio. To detect players with diverse pixel sizes, the input image patches should be cropped from the original images by appropriate zooming scales to ensure that the sizes of cropped players are approximately 42×25. Based on the statistics of the player pixel sizes in our training set, the zooming scales for the original images were determined to be 30%, 40% and 50%, which covered most of the player sizes. The confidence thresholds of the first phase for both the training and testing procedures were set to 0.85 to ensure a recall rate of 98% in order to exclude samples with particular appearances from the training in the second phase, which would be beneficial for the generalization of the proposed model. Considering the diversity of players' appearances, the confidence threshold of Phase II was determined to be 0.6 to recognize players whose appearances were very different from those of the training samples. Non-maximum suppression (NMS) is used after the detection portion of Phase I so

TABLE 1. Reference criteria for the HSL values of the pixels.

Filter Criteria	Classification Criteria		
	Color channels	Hue	Lightness
Lightness ∈ (0.1, 0.85)	Yellow	(30, 90)	-
	Green	(90, 150)	-
Saturation ∈ (0.2, 1.0)	Blue	(150, 270)	(0.1, 0.5)
	White	(210, 330)	(0.5, 0.85)
	Red	(0, 30), (330, 360)	-

FIGURE 2. Procedure of team classification. The uniform regions were cropped from the detected players that contained discriminate color feature of uniforms, from which the contribution of essential color were extracted to determine the team affiliations.

that Phase II needs only to further recognize very small areas of samples.

C. TEAM CLASSIFICATION BASED ON THE UNIFORM COLOR FEATURE OF THE PREDICTIONS

To adaptively classify the teams to which the players belong without requiring any additional dataset annotations, a simple but effective inference-based method is proposed based on the outcome of the prediction. The most obvious feature for team classification is the uniform color, which is notably different between the home and away teams. Therefore, five essential color channels for the ice hockey uniforms are integrated by understanding the color features of the samples and by calculating the color distribution of the region corresponding to the uniform for the five pixel-by-pixel color channels to determine the team affiliation.

After the targeted players are detected, images with detected players are processed to extract the features of the color distribution. Figure 2 shows the team classification procedure. First, the center areas of the predictions are cropped to represent the discriminatory uniform regions, and then the hue, saturation and lightness (HSL) values of the uniform regions are extracted pixel by pixel. Second, all the pixels in a uniform region are allocated into five essential color channels (green, yellow, blue, red and white) according to the HSL reference criteria, which are obtained by summarizing the color features of the ice hockey player samples. The reference criteria are shown in Table 1, where the lightness and saturation values are on a scale of 0 to 1 and the hue value is in the range of 0 to 360. Finally, the team affiliations are determined by the channels that contain the maximum proportions of pixels.

Let $S = \{(c_k, q_{c_k})\}, 1 \leq k \leq 5$ denote the intervals of the color distribution of a uniform region. $c_k = \{yellow, green, blue, red, white\}$ is the color of the kth channel. $q_{c_k} \in N^*$ is the number of pixels in the kth channel. According to the statistics on uniform color from 64 countries of the International Ice Hockey Federation (IIHF), 87.5% of the home team uniforms are white, which, compared with the other colors, is vulnerable to misclassification from the presence of a pattern on the uniform. To reduce the influence of different uniform designs on white uniforms, the weight of the white channel is increased by a weighted term λ. The final result for team membership T is defined as:

$$T = \begin{cases} white, & \lambda q_{white} \geq q_{\max} \\ c_{\max}, & \text{otherwise} \end{cases} \tag{4}$$

where q_{white} is the number of pixels in the white channel, $q_{\max}$ is the maximum number of pixels in the five channels, $c_{\max}$ is the corresponding color of the channel with $q_{\max}$, and λ is the weighted term for the pixels in the white channel. If q_{white} is equal to or greater than $q_{\max}/\lambda$, the final result will be white; otherwise, it will be $c_{\max}$. In our numerical experiments, λ is set to 0.7.

IV. DATASET CONSTRUCTION FROM THE 2018 WINTER OLYMPICS

Our dataset consisted of six broadcasted ice hockey games from the 2018 Winter Olympics in PyeongChang, including five men's and one women's games, namely, Russia (OAR)

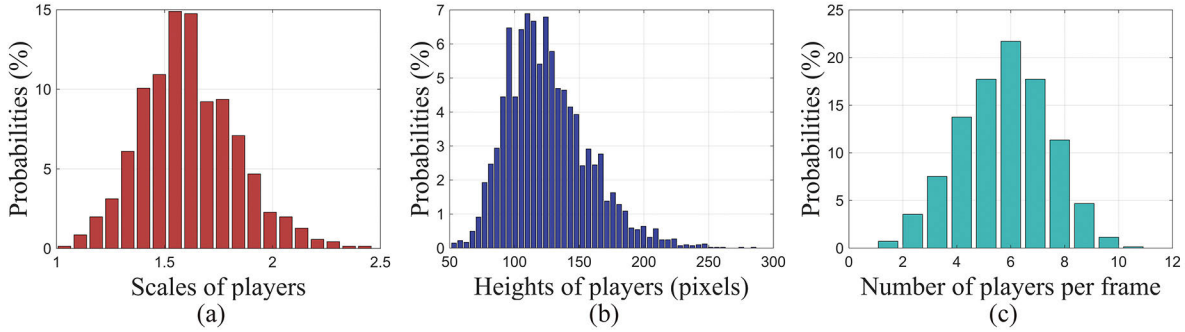


FIGURE 3. Statistics of dataset. (a) The distribution about the scales of players. Scales are defined as the ratio of height and width. (b) The distribution about the heights of players with each number calculated as pixel values per player. (c) The total number of players per image.

vs Germany (GER), Canada (CAN) vs GER, Czech Republic (CZE) vs OAR, CAN vs The United States (USA), Finland (FIN) vs OAR and Switzerland (SUI) vs Japan (JPN). These videos were recorded by official pan-tilt-zoom broadcast cameras. Most of the highlights and playback scenes were manually removed from our dataset, and some of the close shots from cameras were also excluded to ensure that the players of interest are well positioned in the video. Images with a resolution of 1280 × 720 were extracted from these game videos, where the locations of the players were manually annotated in the form of $\{x_{bbox}, y_{bbox}, h_{bbox}, w_{bbox}\}$, where x_{bbox} and y_{bbox} are the x-coordinate and y-coordinate, respectively, of the left-top point of the bounding box, h_{bbox} and w_{bbox} are the pixel values of the height and width, respectively, of the bounding box. In addition, several statistical analyses were conducted to visualize our dataset, as shown in Figure 3.

Figures 3a and 3b show that both the aspect ratios and heights are widely distributed, with nearly half of the aspect ratios close to 1.65 and the height of the players ranging from 80 pixels to 150 pixels. Additionally, the total number of players in each image can be approximated by a Gaussian distribution with a mean value of six (Figure 3c). Note that the main purpose of our proposed model is to detect individual players and teams, which indicates that team uniform color should be considered when segmenting the whole dataset. Hence, the images from four games in which the team uniform colors greatly differ from each other are divided into a training set (705 images with 4048 annotated players in total) and a testing set (212 images with 1341 annotated players in total). The parameters of the self-built dataset from the 2018 Winter Olympics are summarized in Table 2.

TABLE 2. Parameters of the self-built dataset from the 2018 Winter Olympics.

Parameters	Values
Size of dataset (images/objects)	917/5389
Resolution (height × width)	1280 × 720
Player aspect ratio	1.65 ± 0.22
Player height (pixels)	125 ± 31
Number of players per frame	5.74 ± 1.83
Number of uniform pairs	4

V. EXPERIMENTAL RESULTS

In this section, numerical experiments on the self-built dataset from the 2018 Winter Olympics in PyeongChang were conducted to validate our proposed two-phase cascaded CNN model for the detection of individual players and teams. Furthermore, the proposed model was applied to the broadcasted videos of the 2019-20 NHL regular season games covering all 31 teams (source from https://v.qq.com/) to verify the robustness of the model. After that, our method was compared with a baseline and several state-of-the-art object detection algorithms. For a statistical analysis of all the experimental results, an intersection over union (IOU) threshold of 0.3 was chosen. Different from conventional object detection (0.7 for Faster R-CNN [53], 0.5 for SSD [54] and YOLOv3 [41]), ice hockey players moving at high speed result in seriously stretched aspect ratios and blurred limbs, which makes the predicted boxes unable to cover the entirety of the player regions the way that the ground-truth boxes do. Thus, a lower IOU threshold was selected to ensure the integrity of the statistics.

A. DETECTION OF INDIVIDUAL PLAYERS AND TEAM AFFILIATIONS FOR THE SELF-BUILT DATASET FROM THE 2018 WINTER OLYMPICS

In silico experiments were performed with our model to detect individual players from four games, namely, GRE vs OAR, CZE vs OAR, FIN vs OAR, and CAN vs USA. The first row in Figure 4 represents Phase I of the detection algorithm, with green bounding boxes indicating the outputs from the first branch of the classification. As the figure shows, misclassifications occur when the player images are in a complex context, such as mixed with the audience (for example, see FIN vs OAR and CAN vs USA), or when the backgrounds are complicated; e.g., stripes appearing on the sideline advertising bars could easily be recognized as player jersey numbers (for example, see GER vs OAR and CZE vs OAR). However, these misclassifications disappear in Phase II because,

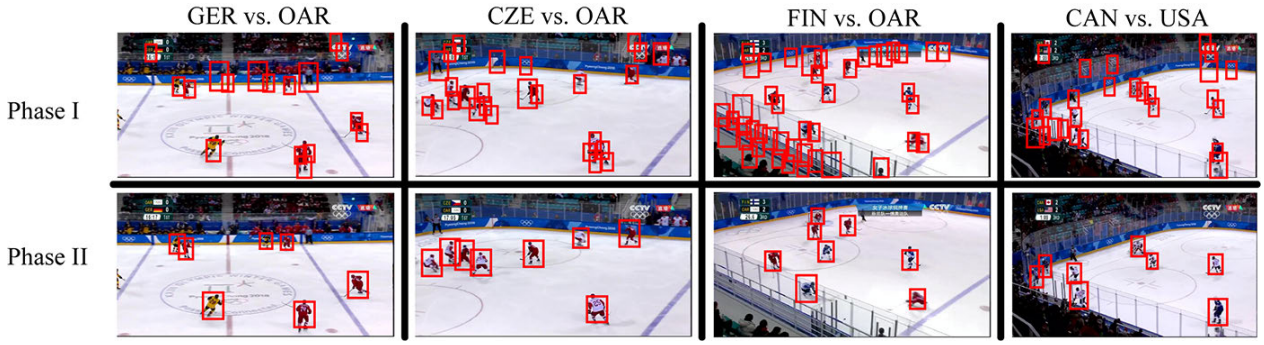


FIGURE 4. Performance of player detection in two classification phases.

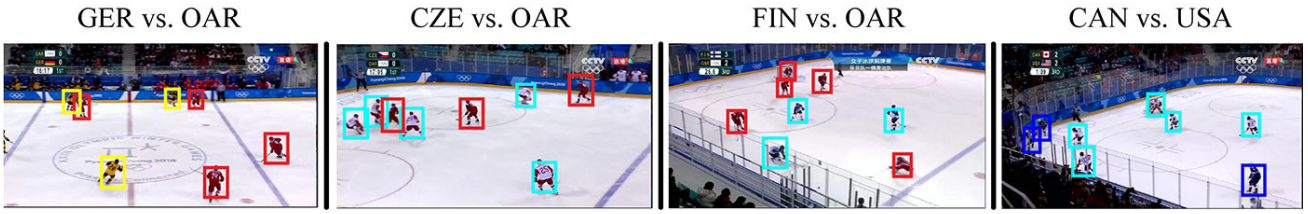


FIGURE 5. Detection of teams. The colored bounding boxes indicate the inferred teams suggested by the model.

TABLE 3. Performance in the detection of individual players.

Game	Correct	False	Missing	Precision	Recall	F-score
GER vs. OAR	278	3	22	0.9893	0.9266	0.9569
CZE vs. OAR	349	3	15	0.9915	0.9588	0.9749
CAN vs. USA	442	5	27	0.9888	0.9424	0.9650
FIN vs. OAR	193	5	15	0.9747	0.9279	0.9507

as the true positive samples, the targeted players have larger confidence scores than the false positive samples. Hence, the second branch of classification accurately selects all false positive samples.

The performance of our two-phase cascaded CNN model was evaluated according to precision, recall and F-score, which are shown in Table 3. Across all the selected games, the average precision and recall reach 0.9875 and 0.9411, respectively. Since our model is capable of detecting players in blurred images and recognizing players with various body positions, the model was thus able to achieve a high recall in the testing set. The F-scores for all four games exceed 95%, demonstrating the robustness and effectiveness of our individual player detection.

The validation of our model for team detection is further verified by exploring team classification based on the color information from the jerseys. Table 4 summarizes the evaluation of the team classification, where GT stands for the ground truth and INF stands for the inferred teams predicted by our model, comprising team A (TA) and team B (TB). OTHER means that the predicted results belong to a third category, for instance, a referee or misclassified teams. The results in Table 3 indicate that our approach attains an average accuracy of 93.05% for team detection. Figure 5 presents several team classification examples for the four different uniform components.

B. DETECTION OF INDIVIDUAL PLAYERS AND TEAM AFFILIATIONS FOR THE SELF-BUILT DATASET FROM 2019-2020 NHL REGULAR SEASON GAMES

The NHL is the most popular ice hockey league in North America and includes 31 teams, with each team scheduled for 82 games during the regular season. Due to the high impact and high-level competence of the NHL, the vast amount of data from the games are worth exploring to develop winning tactics and improve player performance. Therefore, another image set was built that contains the newest 39 games from the 2019-20 season that cover all 31 teams in the league to verify the robustness of the two-phase cascaded CNN model. Since individual players and teams are well detected according to Figure 4 and Figure 5, the main purpose here is to test the ability of our method to detect more challenging samples.

TABLE 4. Performance in team classification.

Games	GT	INF		
		TA (%)	TB (%)	OTHER (%)
GER (TA) vs. OAR (TB)	TA	95.83	4.17	0
	TB	0	100	0
CZE (TA) vs. OAR (TB)	TA	74.18	25.82	0
	TB	0	100	0
CAN (TA) vs. USA (TB)	TA	97.85	1.29	0.86
	TB	4.76	95.24	0
FIN (TA) vs. OAR (TB)	TA	80.46	0	19.54
	TB	0	100	0

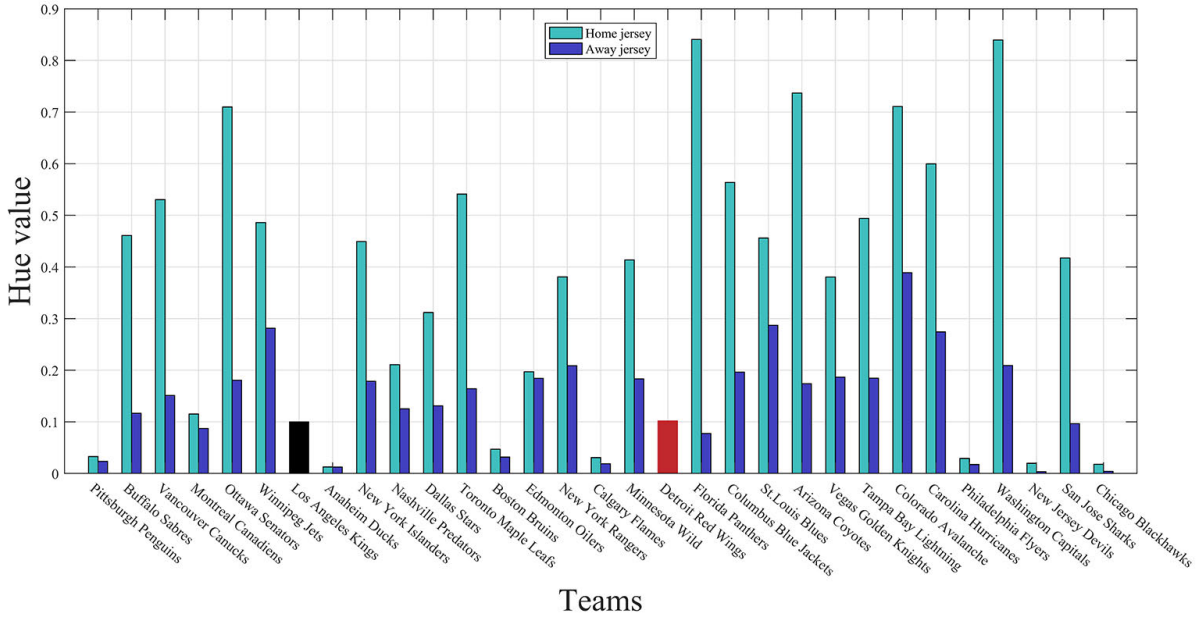


FIGURE 6. Hue value statistics of 31 NHL teams' uniforms.

To understand the diversity of player appearances from the 31 NHL teams, first, the color components of the home and away uniforms of all NHL teams were calculated (source from nhluniforms.com), and the results are shown in Figure 6. Then, the mean hue (the attribute of color that enables an observer to classify it as red, green, blue, purple, etc., and excludes white, black, and shades of gray) values of the different uniforms was extracted and normalized to a range of 0 to 1 to preliminarily describe the major color component, or essential color, of the uniform. Nonetheless, note that the uniforms of the Los Angeles Kings and Detroit Red Wings are exceptions since the hue values of their essential colors equal zero (the major color components are black for Kings and red for Red Wings); thus, the corresponding bars are colored in black and red, respectively, in Figure 6.

In addition, 12 sample images were selected in which all 31 teams appeared, and a video of approximately 10 seconds, consisting of 287 frames, was used to observe if our method was still effective. Figure 7 shows the detection result for 24 sample images where most players were accurately classified into their correct teams. However, some samples were misclassified because the illumination conditions in the stadiums and the positions of the cameras are different for each team. Additionally, the tempo of an NHL games is quicker than that of one of the Winter Olympic Games, resulting in severe overlap between players and large numbers of physical collisions, which worsen the detection accuracy for both individual players and team classification. In general, the proposed method performs well in detecting individual players and determining the team to which they belong within the NHL image set, which motivated us to apply our method to a continuous image sequence to validate the potential for tracking players. A broadcasted video of the Pittsburgh Penguins vs the Anaheim Ducks in the 2019-2020 season is illustrated as an example (see supplementary mate-

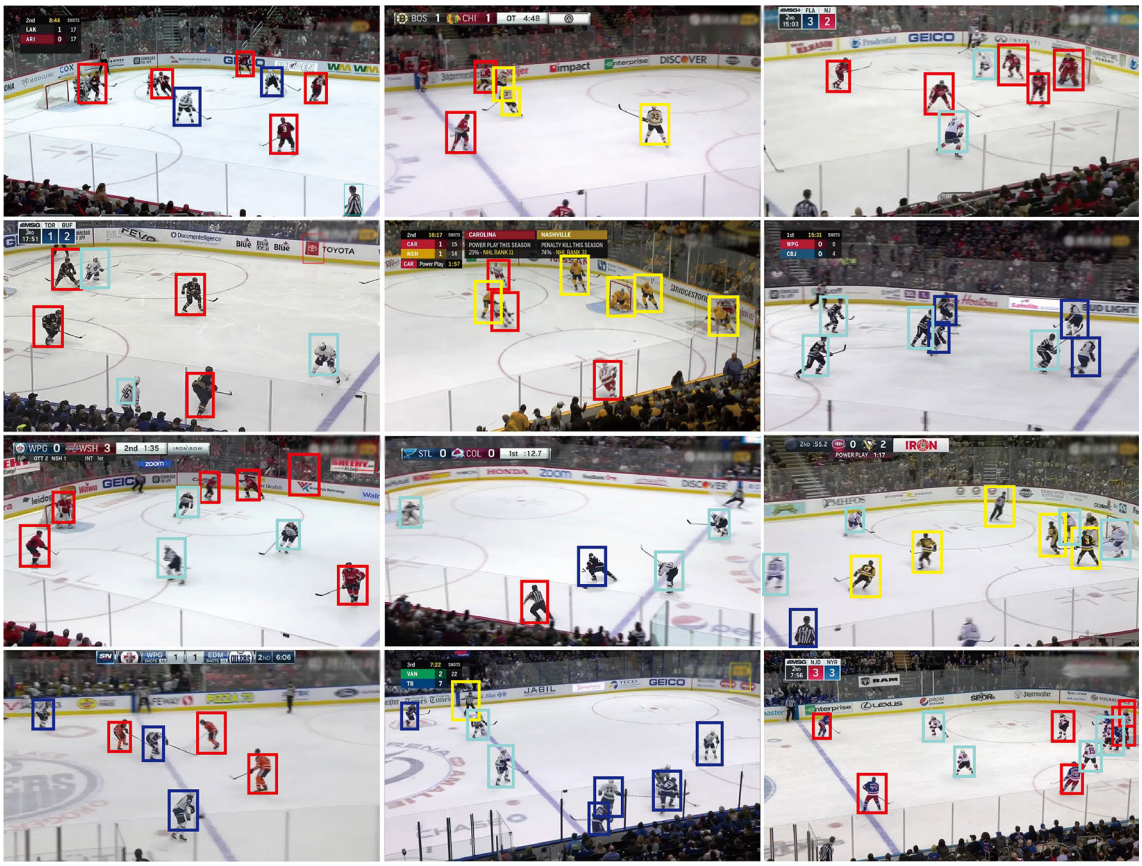


FIGURE 7. Detection of individual players and teams simultaneously from NHL matches.

rial). This video exhibits a power play from the Pittsburgh Penguins that ends with a goal, which is a characteristic offensive pattern that occurs in most ice hockey games. This video is therefore useful for coaches to further devise tactical decisions that weigh the pros and cons of the positions of the players in the offensive team and of the defensive patterns of the opposing team. The results for this continuous image sequence show that our method is able to form a solid foundation for extracting ice hockey player tracking data.

C. COMPARISON EXPERIMENTS FOR THE SELF-BUILT DATASET FROM THE 2018 WINTER OLYMPICS

The size of the input image patches (Parameter 1 in 2.2) was adjusted based on the statistics of the dataset to make our model suitable to detect ice hockey players. To evaluate the benefit of the adjustment of Parameter 1, our method was compared with a baseline model with an input size of 40×18, which is closer to the aspect ratio of pedestrians and soccer and basketball players. The results are presented in Table 5. The precision of our method shows significant improvements over the baseline, confirming that adjusting the size of the input patches is beneficial for extracting the features of the players, especially for our shallow architecture. The recall is slightly lower in our method, as the parameter was adjusted to match the average size of player samples, which to some extent weakens the generalizability for uncommon sizes. Overall, the adjustment of Parameter 1 improves the performance of our method over the baseline in terms of both the F-score and the area under the curve (AUC) value. In addition, several state-of-the-art object detection methods were compared with our method: YOLOv3 [41], Faster R-CNN [53] (vgg16, res101) and CornerNet [55]. YOLOv3 is an efficient one-stage method for object detection, Faster R-CNN is a CNN-based method and CornerNet is an anchor-free object detection algorithm.

Experiments were conducted using the self-built Winter Olympics ice hockey dataset, and the results are shown in Table 6, based on calculations with a confidence score threshold of 0.6. Our method achieves the best overall performance with the highest F-score and outperforms all of the state-of-the-art methods in terms of precision. The recall is much better than that of YOLOv3 and close to the highest value, achieved by the Faster R-CNN-vgg16 method. For the self-built dataset, YOLOv3 performs very well in terms of precision, and the Faster R-CNN methods can expertly recall more players. Our method yields competitive results in both precision and recall, and it outperforms YOLOv3 and the Faster R-CNN methods overall. The performance from the

TABLE 5. Comparison of performances between our method and the baseline.

Methods	Precision	Recall	F-score	AUC
Baseline	0.8043	**0.9798**	0.8834	0.8919
Our method	**0.9875**	0.9411	**0.9637**	**0.9566**

TABLE 6. Comparison of the performances of our method and state-of-the-art algorithms.

Methods	Precision	Recall rate	F-score
YOLOv3	0.9821	0.9023	0.9405
Faster R-CNN-vgg16	0.9515	**0.9657**	0.9585
Faster R-CNN-res101	0.9606	0.9619	0.9613
CornerNet	0.7608	0.6282	0.6881
Our method	**0.9875**	0.9411	**0.9637**

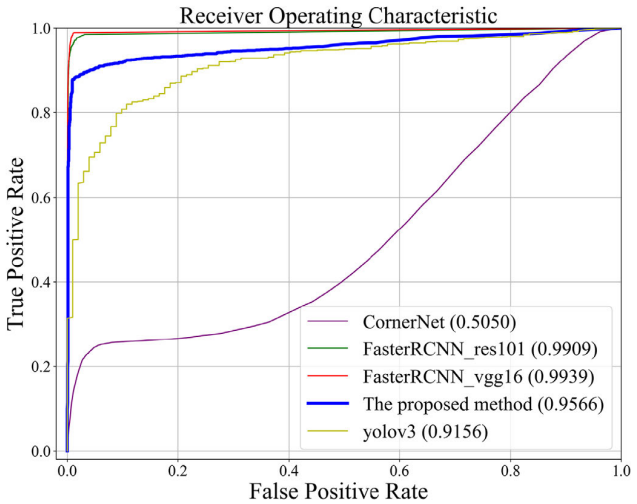


FIGURE 8. ROC curves of the methods. The AUC values are presented in the legend.

CornerNet method is unsatisfactory in the detection of ice hockey players, as the applied bottom-up design detects a top-left corner set and a bottom-up corner set at first and then groups the corner pairs by calculating the distances between the corresponding embedding vectors, which possibly leads to poor sensitivity in recognizing similar-appearing samples from a single class.

Figure 8 shows the receiver operating characteristic (ROC) curves and AUC values for all methods. Our method (blue line) outperforms YOLOv3 and CornerNet with obvious advantages but underperforms with respect to the Faster R-CNNs. As a result of not setting the threshold of the confidence score for our method and YOLOv3, the levels of detection for low confidence scores were close to those for the true positives, which therefore might be considered positive samples, leading to a lower AUC value. On the other hand, the low-score detections of the Faster R-CNN methods were more likely to appear in the background, resulting in a better performance. However, our method is still competitive in terms of AUC value.

VI. DISCUSSION

Accurate detection of the players in ice hockey games is crucial for improving player performance and team tactical decision making and is therefore urgently needed for coaches and other analysts. In this paper, a two-phase cascaded CNN model was proposed to detect individual players labeled with team affiliations on a self-built dataset from the 2018 Winter Olympics in PyeongChang. The precision and recall for individual player detection in the ice hockey games in the Winter Olympics dataset were 98.75% and 94.11%, respectively, while the average accuracy of the team classification reached 93.05%. Our model was also applied to the dataset containing 2019-20 NHL regular season games covering all teams in the league, in which our model accomplished the concurrent detection of players and their team affiliations from image and video inputs.

Our model is designed in particular for player tracking data collection in ice hockey and is capable of dealing with some characteristic video and image features, for example, attaining high detection accuracy with fuzzy images. In addition, the main architecture of the proposed deep learning network is effective and straightforward, as the two phases in the network subtly establish different confidence scores for the targeted players and other noninformative contexts, which avoids the need to update parameters redundantly and repetitively. Nevertheless, some further improvements have to must be implemented in future studies, such as the recognition of players with extremely stretched aspect ratios and precise team classification when players are in physical contact. The current model is potentially capable of extracting players' tracking data for other group sports similar to ice hockey, such as rugby and handball. In addition, the proposed player detection model can be applied to player annotation and event recognition in broadcasted video. Additionally, it is worth noting that the trajectories of the players can be reconstructed so that the locations and movements of players can be visualized temporally (Figure 9). Advanced game statistics could be extracted by developing our current models, which will shed light on insightful analyses and predictions in ice hockey.

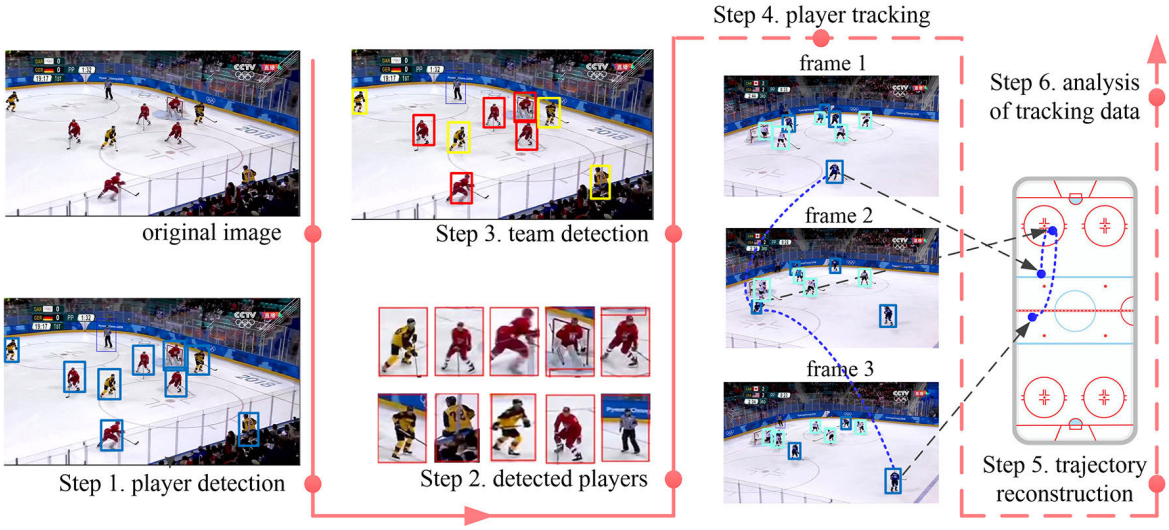


FIGURE 9. Envisioned frameworks for tracking data analysis in ice hockey. Steps on solid line are completed while Step 4 to 6 on the dashed line constitute future work.

REFERENCES

[1] J. Weissbock, *Forecasting Success in the National Hockey League using In-Game Statistics and Textual Data*. Ottawa, ON, Canada: Univ. Ottawa, 2014.

[2] M. Schuckers and J. Curro, "Total Hockey Rating (THoR): A comprehensive statistical rating of National Hockey League forwards and defensemen based upon all on-ice events," in *Proc. 7th Annu. MIT Sloan Sports Anal. Conf.*, 2013, pp. 1–10.

[3] N. Mehrasa, "Deep learning of player trajectory representations for team activity analysis," in *Proc. 11th MIT Sloan Sports Anal. Conf.*, 2018, pp. 1–8.

[4] M. R. Tora, J. Chen, and J. J. Little, "Classification of puck possession events in ice hockey," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 147–154.

[5] T. C. Y. Chan, J. A. Cho, and D. C. Novati, "Quantifying the contribution of NHL player types to team performance," *Interfaces*, vol. 42, no. 2, pp. 131–145, Apr. 2012.

[6] A. Rajšp and I. Fister, "A systematic literature review of intelligent data analysis methods for smart sport training," *Appl. Sci.*, vol. 10, no. 9, p. 3013, Apr. 2020.

[7] H. M. Le, P. Carr, Y. Yue, and P. Lucey, "Data-driven ghosting using deep imitation learning," in *Proc. MIT Sloan Sports Anal. Conf.*, Boston, MA, USA, Mar. 2017, pp. 1–15.

[8] D. Cervone, A. D'Amour, and L. Bornn, "POINTWISE: Predicting points and valuing decisions in real time with NBA optical tracking data," in *Proc. 8th MIT Sloan Sports Anal. Conf.*, Boston, MA, USA, 2014, p. 3.

[9] P. Lucey, "Quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data," in *Proc. 8th Annu. MIT Sloan Sports Anal. Conf.*, 2014, pp. 1–9.

[10] B. Macdonald, "An expected goals model for evaluating NHL teams and players," in *Proc. MIT Sloan Sports Anal. Conf.*, 2012, pp. 1–8.

[11] H. Pileggi, C. D. Stolper, J. M. Boyle, and J. T. Stasko, "Snapshot: Visualization to propel ice hockey analytics," *IEEE Trans. Vis. Computer Graphics*, vol. 18, no. 12, pp. 2819–2828, Dec. 2012.

[12] T. Imai "Play recognition using soccer tracking data based on machine learning," in *Advances in Network-Based Information Systems*. Cham, Switzerland: Springer, 2018, pp. 875–884.

[13] P. K. Santhosh and B. Kaarthick, "An automated player detection and tracking in basketball game," *Comput., Mater. Continua*, vol. 58, no. 3, pp. 625–639, 2019.

[14] O. Schulte and Z. M. Zhao Javan, "Apples-to-apples: Clustering and ranking NHL players using location information and scoring impact," in *Proc. MIT Sloan Sports Anal. Conf.*, 2017, pp. 1–14.

[15] O. Schulte, M. Khademi, S. Gholami, Z. Zhao, M. Javan, and P. Desaulniers, "A Markov game model for valuing actions, locations, and team performance in ice hockey," *Data Mining Knowl. Discovery*, vol. 31, no. 6, pp. 1735–1757, Nov. 2017.

[16] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews, "Identifying team style in soccer using formations learned from spatiotemporal tracking data," in *Proc. IEEE Int. Conf. Data Mining Workshop*, Dec. 2014, pp. 9–14.

[17] A. C. Miller and L. P. Bornn, "Possession sketches: Mapping NBA strategies," in *Proc. 11th MIT Sloan Sports Anal. Conf.*, 2017, pp. 1–12.

[18] K.-C. Wang and R. Zemel, "Classifying NBA offensive plays using neural networks," in *Proc. MIT Sloan Sports Anal. Conf.*, 2016, pp. 1–9.

[19] C. Tian, V. De Silva, M. Caine, and S. Swanson, "Use of machine learning to automate the identification of basketball strategies using whole team player tracking data," *Appl. Sci.*, vol. 10, no. 1, p. 24, Dec. 2019.

[20] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, and A. Yuille, "PCL: Proposal cluster learning for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 176–191, Jan. 2020.

[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[22] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2129–2137.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[25] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware CNN model," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1134–1142.

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[27] J. P. R. Lara, C. L. R. Vieira, M. S. Misuta, F. A. Moura, and R. M. L. D. Barros, "Validation of a video-based system for automatic tracking of tennis players," *Int. J. Perform. Anal. Sport*, vol. 18, no. 1, pp. 137–150, Jan. 2018.

[28] K. Lu, "Light cascaded convolutional neural networks for accurate player detection," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2017, pp. 1–13.

[29] P. Parisot and C. De Vleeschouwer, "Scene-specific classifier for effective and efficient team sport players detection from a single calibrated camera," *Comput. Vis. Image Understand.*, vol. 159, pp. 74–88, Jun. 2017.

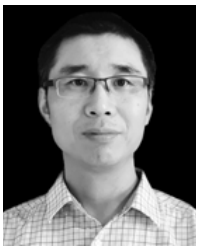
[30] N. U. Huda, K. H. Jensen, R. Gade, and T. B. Moeslund, "Estimating the number of soccer players using simulation-based occlusion handling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1824–1833.

[31] D. Nandashri and P. Smitha, "An efficient tracking of multi object visual motion using Hungarian method," *Int. J. Eng. Res.*, vol. 4, no. 4, Apr. 2015, Art. no. 017577.
[32] J. Liu, X. Tong, W. Li, T. Wang, and Y. Zhang, "Automatic player detection, labeling and tracking in broadcast soccer video," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 103–113, 2009.
[33] J. Liu, X. Tong, W. Li, T. Wang, Y. Zhang, and H. Wang, "Automatic player detection, labeling and tracking in broadcast soccer video," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 103–113, Jan. 2009.
[34] Y. Junqing and W. H. X. Yunfeng, "Camera calibration and player detection in 3D reconstruction," *J. Softw.*, vol. 19, pp. 151–160, 2008.
[35] E. Cheshire, C. Halasz, and J. K. Perin, "Player tracking and analysis of basketball plays," in *Proc. Eur. Conf. Comput. Vis.*, 2013, pp. 1–6.
[36] S. Maćkowiak, J. Konieczny, M. Kurc, and P. Maćkowiak, "Football player detection in video broadcast," in *Computer Vision and Graphics*, vol. 6375. Berlin, Germany: Springer, 2010, pp. 118–125.
[37] R. Miyamoto and T. Oki, "Soccer player detection with only color features selected using informed haar-like features," in *Proc. Adv. Concepts Intell. Vis. Syst. 17th Int. Conf. (ACIVS)*, 2016, pp. 238–249.
[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[39] I. Mrazova and M. Hrincar, "Fast and reliable detection of hockey players," *Procedia Comput. Sci.*, vol. 20, pp. 121–127, 2013.
[40] D. Sorano, F. Carrara, P. Cintia, F. Falchi, and L. Pappalardo, "Automatic pass annotation from soccer VideoStreams based on object detection and LSTM," 2020, *arXiv:2007.06475*. [Online]. Available: http://arxiv.org/abs/2007.06475
[41] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," in *Computer Vision and Pattern Recognition*. Washington, DC, USA: IEEE Computer Society, 2018.
[42] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3907–3916.
[43] J. Gao, Z. Yang, and R. Nevatia, "Cascaded boundary regression for temporal action detection," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–11.
[44] H. Qin, J. Yan, X. Li, and X. Hu, "Joint training of cascaded CNN for face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3456–3465.
[45] G.-S. Xie, X.-Y. Zhang, S. Yan, and C.-L. Liu, "Hybrid CNN and dictionary-based models for scene recognition and domain adaptation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 6, pp. 1263–1274, Jun. 2017.
[46] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2001, pp. I–I.
[47] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Proc. Conf. Learn. Theory*, 1997, vol. 55, no. 1, pp. 119–139.
[48] L. Zhang, R. Chu, S. Xiang, S. Liao, and S. Z. Li, "Face detection based on multi-block LBP representation," in *Proc. Int. Conf. Biometrics*, 2007, pp. 11–18.
[49] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5325–5334.
[50] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[51] T. Rahman, M. E. H. Chowdhury, A. Khandakar, K. R. Islam, K. F. Islam, Z. B. Mahbub, M. A. Kadir, and S. Kashem, "Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray," *Appl. Sci.*, vol. 10, no. 9, p. 3233, May 2020.
[52] I. Kandel and M. Castelli, "A novel architecture to classify histopathology images using convolutional neural networks," *Appl. Sci.*, vol. 10, no. 8, p. 2929, Apr. 2020.
[53] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[54] W. Liu, D. Anguelov, D. Erhan, and C. Szegedy, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.
[55] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.

TIANXIAO GUO received the B.Sc. degree in information engineering from Huaqiao University, Quanzhou, China, in 2017. He is currently pursuing the M.Sc. degree in sports science with Beijing Sport University, Beijing, China. His research interests include sports video understanding and analysis.

KUAN TAO received the Ph.D. degree in mathematical biology from Peking University in 2019. He is currently an Assistant Professor with the School of Sports Engineering, Beijing Sport University. His research interests include sports data analysis and cell movement.

QINGRUI HU received the B.Sc. degree from Chuzhou University, Chuzhou, China, in 2018. He is currently pursuing the M.Sc. degree with Beijing Sport University. His research interests include human pose estimation and human pose recognition.



YANFEI SHEN received the B.S. and the M.S. degrees in computer science from the Key Laboratory of Multimedia and Network Communication, Wuhan University, China, in 1999 and 2002, and the Ph.D. degree in computer applications from the University of Chinese Academy of Sciences, China, in 2014. He has served as an Associate Professor with the Institute of Computing Technology, Chinese Academy of Sciences, and the Beijing University of Posts and Telecommunications. He is currently a Professor with Beijing Sport University. His research interests include intelligent sensing technology for motion capture, computer vision for human action recognition and analysis, and performance analysis for team sports.