frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 06 December 2023
DOI 10.3389/fnbot.2023.1289203

OPEN ACCESS

EDITED BY
Hong Qiao,
University of Chinese Academy of Sciences, China

REVIEWED BY
Ruiheng Zhang,
Beijing Institute of Technology, China
Norio Tagawa,
Tokyo Metropolitan University, Japan

*CORRESPONDENCE
Jiangtao Luo
✉ Luojt@cqupt.edu.cn

RECEIVED 05 September 2023
ACCEPTED 13 November 2023
PUBLISHED 06 December 2023

CITATION
Zhang P and Luo J (2023) Player detection method based on scale attention and scale equalization algorithm.
Front. Neurorobot. 17:1289203.
doi: 10.3389/fnbot.2023.1289203

Player detection method based on scale attention and scale equalization algorithm

Pan Zhang[1,2] and Jiangtao Luo[1,3]*

[1]School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China, [2]Data Recovery Key Laboratory of Sichuan Province, Neijiang Normal University, Neijiang, China, [3]Electronic Information and Networking Research Institute, Chongqing University of Posts and Telecommunications, Chongqing, China

Introduction: Object detection methods for team ball games players often struggle due to their reliance on dataset scale statistics, resulting in missed detections for players with smaller bounding boxes and reduced accuracy for larger bounding boxes.

Methods: This study introduces a two-fold approach to address these challenges. Firstly, a novel multi-scale attention mechanism is proposed, aiming to reduce reliance on scale statistics by utilizing a specially created SIoU (Similar to Intersection over Union) label that explicitly represents multi-scale features. This label guides the training of multi-scale attention network modules at two granularity levels. Secondly, an integrated scale equalization algorithm within SIoU labels enhances the detection ability of multi-scale targets in imbalanced samples.

Results and discussion: Comparative experiments conducted on basketball, volleyball, and ice hockey datasets validate the proposed method. The relative optimal approach demonstrated improvements in the detection accuracy of players with smaller and larger scale bounding boxes by 11%, 7%, 15%, 8%, 9%, and 4%, respectively.

KEYWORDS

multi-scale target detection, scale attention, SIoU, scale equalization, implicit feature fusion

1 Introduction

In team sports, such as basketball, volleyball, and ice hockey, the precise detection of players serves as the fundamental basis for intelligent auxiliary analysis of player movement data, assessment of multi-player coordinated behaviors, and comprehensive team technical and tactical analysis (Lu et al., 2011, 2013; Nishikawa et al., 2017; Stein et al., 2018; Kong et al., 2020). However, in the aforementioned competition scenarios, the statistical distribution of players' bounding boxes becomes wider and unbalanced due to the diversity of shooting distances and angles, along with the continuous movement and random switching of the camera. Specially, this substantial imbalance impairs the detection and localization abilities of existing model algorithms, particularly concerning extremely small and extremely large scale bounding boxes targets. Therefore, enhancing the detection ability of multiple players in non-equilibrium scale statistical scenes has become a significant challenge in the research and improvement of numerous algorithms in the field of computer vision.

As for the improvement of traditional algorithms, the primary emphasis lies on explicit multi-scale feature acquisition and fusion. In Lu et al. (2011), the combination of Histogram of Oriented Gradients (HOG) with color information is proposed. Stein et al. (2018) suggests the fusion of color histograms with target center points.

Additionally, Santhosh and Kaarthick (2019) introduces the combination of the Deformable Parts Model (DPM) with Scale Invariant Feature Transform (SIFT) keypoints. These methods can significantly enhance the ability to extract explicit features of players through artificially designed operators. However, they exhibit more localized effectiveness and encounter difficulties in adaptively detecting targets of all scale bounding boxes.

The improvement based on deep learning models primarily leverages the universal object detection framework and its extensions (Akan and Varli, 2022; Sah and Direkoglu, 2023) to achieve the acquisition and fusion of implicit multi-scale features. As demonstrated in Nishikawa et al. (2017), the multi-branch output structure of the enhanced YOLOv3 model is directly employed to acquire and merge adjacent scale basketball player features. Building upon the addition of various scale feature detection branches, Kong et al. (2019) further integrates a spatial pyramid pooling (SPP) module, enhanced by hole convolution, into the training of the medium scale detection branch with the largest sample volume. This integration aims to enhance the complexity and precision of feature extraction and mitigate potential model overfitting or underfitting arising from sample imbalance. In Buric et al. (2019), features from non-adjacent scales were fused by integrating improved Feature Pyramid Networks (FPNs) into the backbone network, and the Fast R-CNN model was combined to enhance the detection effectiveness of multi-scale football players. Simultaneously, incorporating an attention mechanism into the backbone network for multi-scale feature extraction and fusion is also a prevalent approach. In line with this, both Komorowski et al. (2020) and Hurault et al. (2020) utilize attention mechanisms to enhance the detection capability of football players. In He (2022), attention mechanism was combined with a encoder-decoder model to obtain and fuse multi-scale features through encoding and decoding, achieving the detection of multiple types of multi-scale players. However, the naturally formed player detection dataset still exhibits an imbalance in the distribution of scales, resulting in a significant number of omissions in the detection of players with small sacle bounding boxes and inaccurate positioning of players with large scale bounding boxes in the aforementioned improved algorithms.

In response to the above issues, and inspired by techniques from partial feature fusion (Zhang et al., 2022) and data processing (Ding et al., 2023), this article proposes a multi-scale attention mechanism that weakly relies on the scale statistical distribution features of the dataset and a scale equalization algorithm. These methods combine the strong implicit feature extraction ability of deep learning models with the local enhancement characteristics of traditional operators describing explicit features, thereby further improving the accuracy of multi-scale player detection. The main innovations and contributions of this article include: (1) The proposal introduces the Similar to Intersection over Union (SIoU) label to represent explicit feature information of multi-scale targets. Based on this label, relevant network modules are constructed to generate coarse-grained scale attention feature planes that aid in multi-scale target detection. (2) An algorithm combining non Supervised learning and interval estimation using the statistical distribution information of the coarse-grained scale attention feature plane is proposed, so as to form a fine-grained scale attention with higher concentration. (3) We presents a scale equalization algorithm that is attached to the SIoU label and integrated into the training of the scale attention generation module. The algorithm aims to address the issue of network overfitting during training, which arises from the presence of a significant volume of samples with identical scale targets. Additionally, it mitigates the training error caused by the imbalance in the scale distribution of players' bounding boxes in ball team competitions.

2 The principle of SIoU label

The Intersection over Union (IoU) (Yu et al., 2016) is a metric commonly employed in object detection tasks to assess algorithm performance. It is defined as the ratio between the intersection and union of the predicted field of view bounding box and the target's actual bounding box. This article formulates equation (2) using equation (1) to compute the SIoU (Similar to Intersection over Union) label. The SIoU label represents the ratio of intersection and union between the predicted field of view bounding box and the actual bounding box of the target in the output feature plane of the observed field of view. It calculates this value while continuously shifting the center position (x, y) of the predicted field of view bounding box. $S_{\text{target}}(k)$ denotes the true bounding box of the k-th target, and $S_{\text{kernel}}(x,y,z)$ represents the predicted bounding box of the z-th observation field when the output feature plane is centered at point (x, y). The SIoU values that can be generated through systematic variation of the size of the predicted field of view bounding box and the target's actual bounding box are illustrated in Figures 1, 2. This numerical characteristic of change exhibits similarity to the credibility of the human visual system when observing multi-scale targets across different fields of view, thus providing an explicit expression of multi-scale characteristics.

$$\text{IoU} = \frac{S_{\text{overlap}}}{S_{\text{union}}} \tag{1}$$

$$\text{SIoU}(x,y,z,k) = \frac{S_{\text{overlap}}(x,y,z,k)}{S_{\text{union}}(x,y,z,k)} = \frac{S_{\text{target}}(k) \cap S_{\text{kernel}}(x,y,z)}{S_{\text{target}}(k) \cup S_{\text{kernel}}(x,y,z)} \tag{2}$$

Figure 2 displays a representative statistical distribution of SIoU values, obtained through a typical single point quantization calculation, applied to targets of various scales using four corresponding equivalent prediction field of view boundary boxes. The typical single point quantization value refers to the SIoU value calculated when the predicted field of view bounding box aligns precisely with the center position of the target's actual bounding box. This serves as an illustrative example of certain feature points in Figure 1.

In Figure 2, the distinct line types represent different predicted branches η to which the target belongs. The calculation of these branches is determined by equation (3), where $\eta_{\max}$ denotes the upper limit of the number of predicted branches in the model. In equation (3), $\ell_{\text{target}}(k)$denotes the edge length of the k-th target, which is computed following equation (4). Likewise, $\ell^{z}_{\text{kernel}}(x_{\text{center}}, y_{\text{center}})$ signifies the edge length of the z-th basic predicted field of view bounding box, calculated based on equation

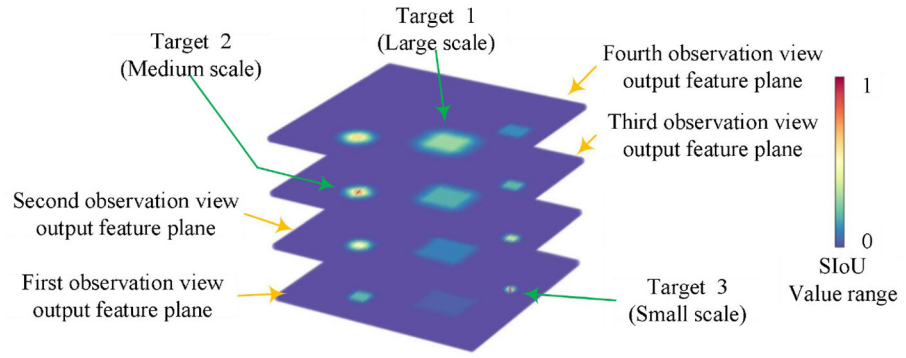

FIGURE 1
The visualization of SIoU features across distinct scales for various targets.

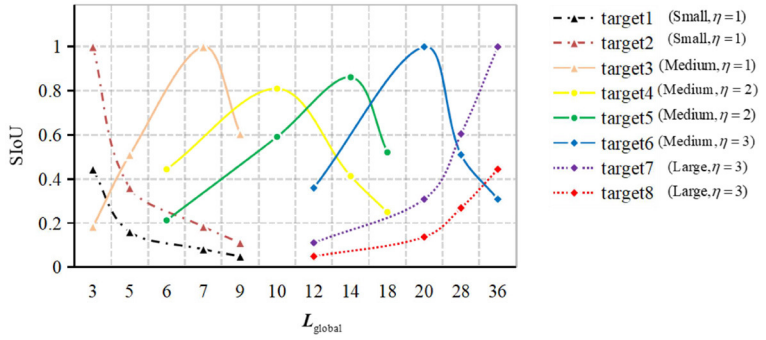

FIGURE 2
Numerical distribution of SIoU values for typical scale targets.

(5). The set L_{global}, comprising the edge lengths of all globally equivalent predicted field of view bounding boxes in the figure, is derived following equation (6).

$$\eta = \min(\max(\log_2 \frac{\ell_{\text{target}}(k)}{\max(\{\ell^z_{\text{kernel}}(x_{\text{center}}, y_{\text{center}})\})} + 2, 1), \eta_{\max}) \quad (3)$$

$$\ell_{\text{target}}(k) = \sqrt{S_{\text{target}}(k)} \quad (4)$$

$$\ell^z_{\text{kernel}}(x_{\text{center}}, y_{\text{center}}) = \sqrt{S_{\text{kernel}}(x_{\text{center}}, y_{\text{center}}, z)} \quad (5)$$

$$L_{global} = \{\ell^z_{\text{kernel}}(x_{\text{center}}, y_{\text{center}}) \times 2^{\eta-1}\} \quad (6)$$

The variation pattern observed in different color curves in Figure 2 indicates that the SIoU value exhibits correlation between the same target and different predicted fields of view bounding box. Moreover, it demonstrates distinguishability for targets of the same category but different scales. Among the four consecutive SIoU values obtained, those corresponding to small-scale bounding box targets exhibit relatively small values and display a decreasing trend. In contrast, the SIoU values for medium-scale bounding box targets are relatively larger, with an initial increase followed by a subsequent decrease. For large-scale bounding box targets, the SIoU values are relatively small and demonstrate an upward trend. These trends primarily emphasize the relative relationships among SIoU values, rather than the absolute values themselves.

Figure 3 presents the statistical distribution of all corresponding SIoU values computed for equivalent target bounding box sizes ranging from 3 × 3 to 54 × 54. These calculations are performed when the observation view output feature planes of the three prediction branches are set to 56 × 56, 28 × 28, and 14 × 14, respectively. The SIoU values are categorized into two groups based on the size of the predicted view bounding box and the actual target bounding box. As depicted in the Figure 3, the SIoU numerical ranges for the majority of target exhibit considerable overlap and intersections with one another. This observation suggests that employing any volume of samples and training the model to extract the four required SIoU numerical features for targets of diverse scales indirectly enhances the extraction capability of relevant SIoU values for targets of other scales. Moreover, it indicates a weak dependence of the SIoU value on the scale statistical distribution of the dataset.

When employing the SIoU value-based label to assist the depth Convolutional Neural Network in constructing a multi-scale attention plane, and under the condition where all branches share the same SIoU value, the network model can accommodate different scale targets through its multi-scale branch structure. Additionally, the predicted field of view bounding boxes at various

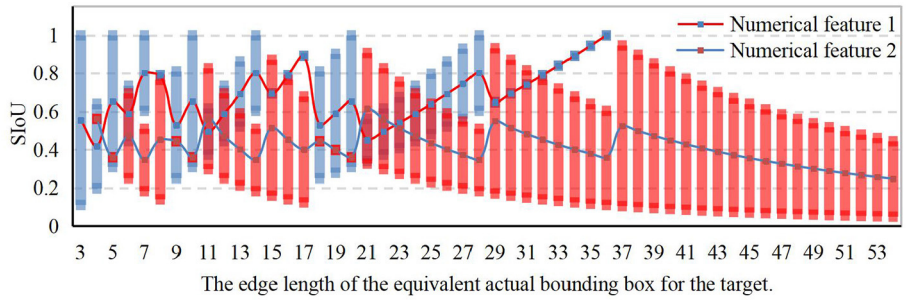

FIGURE 3
Statistical distribution of SIoU numerical ranges for all scale targets.

scales can be efficiently replaced by globally equivalent predicted field of view bounding boxes in different branches, utilizing basic convolutional kernels with size of 3 × 3, 5 × 5, 7 × 7, and 9 × 9, respectively. As depicted in Figure 4, the input basketball game image comprises a total of 6 targets, consisting of 2 large-size targets, 2 medium-size targets, and 2 small-size targets. Following the aforementioned guidelines, the scale attention for targets A and B is assigned to the small-size branch 3, the scale attention for targets C and D is assigned to the medium-size branch 2, and the scale attention for targets E and F is allocated to the large-size branch 1.

3 Proposed method

Based on the SIoU label, we initially construct a network module to extract multi-dimensional distribution features. It utilizes coarse granularity scale attention formed by the explicit features of multiple scales to enhance multi-target detection with scale imbalance. Subsequently, leveraging the distinctive characteristic of a single target type in team sports, the K-medoids algorithm is enhanced by incorporating player bounding box information and statistical features, resulting in a fine-grained scale attention optimization algorithm. Finally, the proposed scale equalization algorithm is integrated with the SIoU label to jointly facilitate the training of the network model incorporating multi-scale attention.

3.1 Network module for SIoU feature extraction

This article introduces a network module named MdSNet (Multidimensional SIoU Net) designed to extract multi-dimensional SIoU features generated by multi-scale targets through the application of multi-scale convolution kernels. As depicted in Figure 5A, MdSNet comprises three main components: a planar scale attention processor, a stereoscopic scale attention processor, and a scale attention fine-tuning structure. Their corresponding training loss functions are denoted as loss1, loss2, and loss3, respectively. Simultaneously, we illustrate the relationship between the MdSNet module and traditional object detection and localization models in Figure 5B. Ultimately, the module outputs a fine-grained scale attention feature plane.

The planar scale attention processor incorporates multi-scale convolutional kernels and sigmoid functions. The four sizes of convolutional kernels generate four planar scale attention feature maps for all corresponding targets in their respective scale branches. The resulting feature maps are then concatenated to form a multi-channel structure. The stereoscopic scale attention processor is composed of a 3D convolutional kernel and sigmoid functions. It takes multi-channel planar scale attention concatenation maps as input, producing coarse-grained scale attention planes, and predicting the number of potential targets within the planes. The scale attention fine-tuning structure comprises a statistical feature extraction process and a codec, ultimately yielding a fine-grained scale attention plane.

3.2 Process for coarse-grained attention generation

The planar scale attention processor and the stereoscopic scale attention processor collectively constitute the pivotal components of the SIoU multi-dimensional distribution feature extraction network module. The training process commences sequentially, considering both the sample volume of the dataset and the structure of the network model. Firstly, the planar scale attention processor is trained, and the data labels during training are generated based on equation (2). Figure 6A is a conventional feature map, while Figure 6B is a single channel feature map obtained using a fixed size convolution kernel. Figures 6C, D illustrate the predicted data and label data, respectively. At this stage, the loss function loss1 is constructed based on the L2 norm, which is the mean square error function, and the optimizer used is the Stochastic Gradient Descent (SGD) algorithm. The main objective of this training process is to discriminate the various SIoU numerical information generated by different scale bounding box targets under the influence of the same size convolutional kernel. The emphasis lies in obtaining the absolute distribution of SIoU features in the plane space, as expressed by each output channel feature map. Secondly, the stereoscopic scale attention processor is trained to improve the capability of extracting multi-dimensional SIoU features, with a particular emphasis on capturing the relative

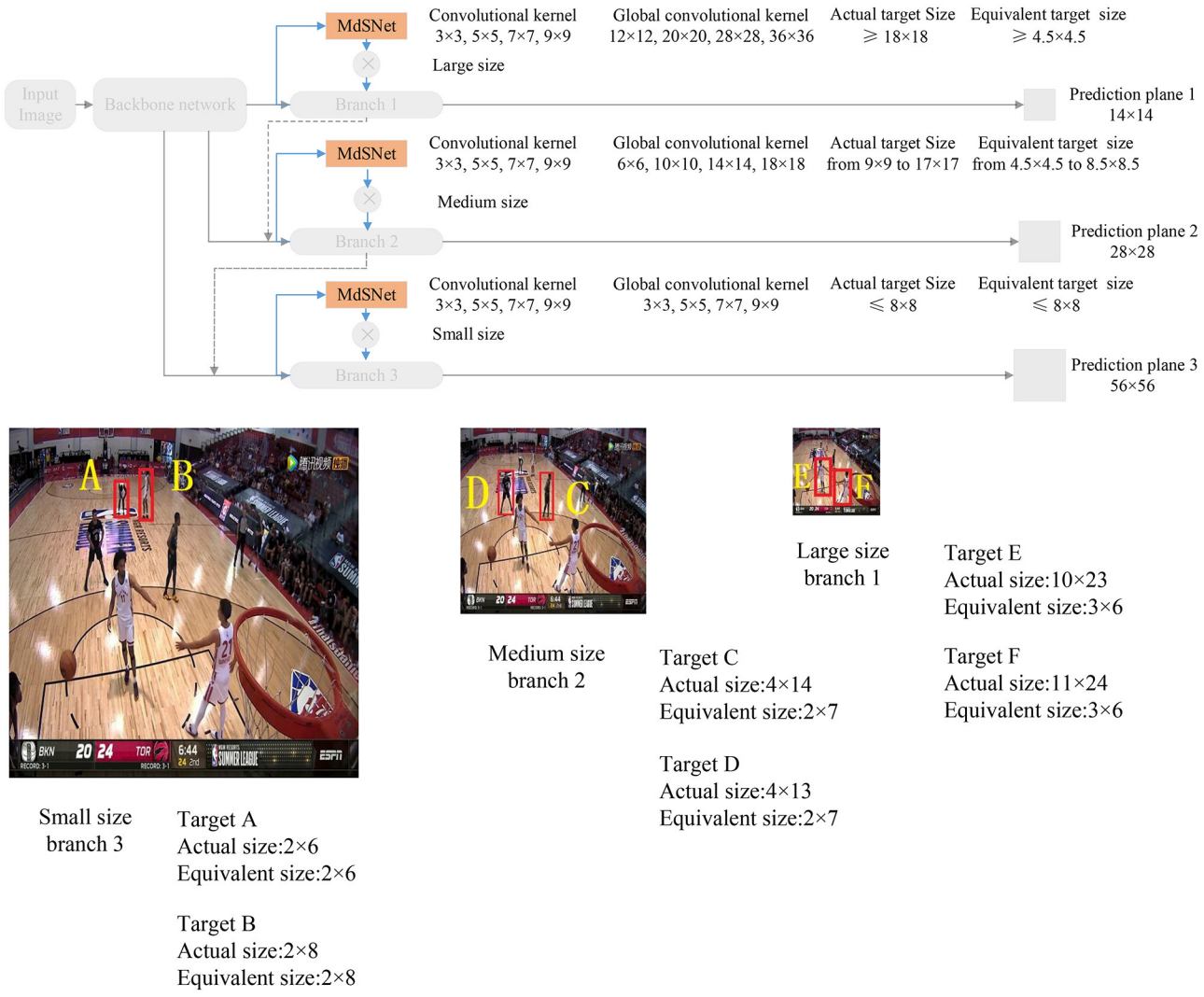

FIGURE 4
Guidelines for selecting target belonging branches.

relationships between the SIoU values of each channel within the input multi-channel planar scale attention feature map. Data labels required for training are generated using a normal distribution, where the statistical distribution of each target is set as a normal distribution with parameter (μ, σ), serving to approximate the coarse-grained scale feature range. The specific values of this parameter can be determined through experimental evaluations. This is shown in Figure 6E. At this stage, the loss function loss2 is constructed based on the L2 norm.

3.3 Process for fine-grained attention generation

Scale attention fine-tuning structure employs real data to compensate for the subjectivity of the SIoU label in this study, and it aims to optimize the coarse-grained scale attention features produced by the MdSNet network module. This structure executes Algorithm 1, initially employing the enhanced K-medoids algorithm in conjunction with the number of targets predicted by the previous processor in the feature map to compute the center position of each target on the coarse-grained scale attention feature plane. Subsequently, the orientations of all targets are sorted using the Manhattan distance. Finally, through training with a codec and statistical interval estimation method, the confidence interval derived from real data guides the module to generate the best-matched confidence interval, achieving fine-tuning of scale attention.

The essence of the K-medoids algorithm improvement resides in the distance calculation method between the associated feature points, as illustrated in equation (7). φ and τ are obtained based on equations (8) and (9), respectively, where $(x_0, x_1)(y_0, y_1)$ represents the coordinate information of the two points, and f_{size} denotes the size of the current feature plane. Considering that competitive game images are resized to a standard size of 448×448 before being fed into the network model, the bounding boxes of players exhibit

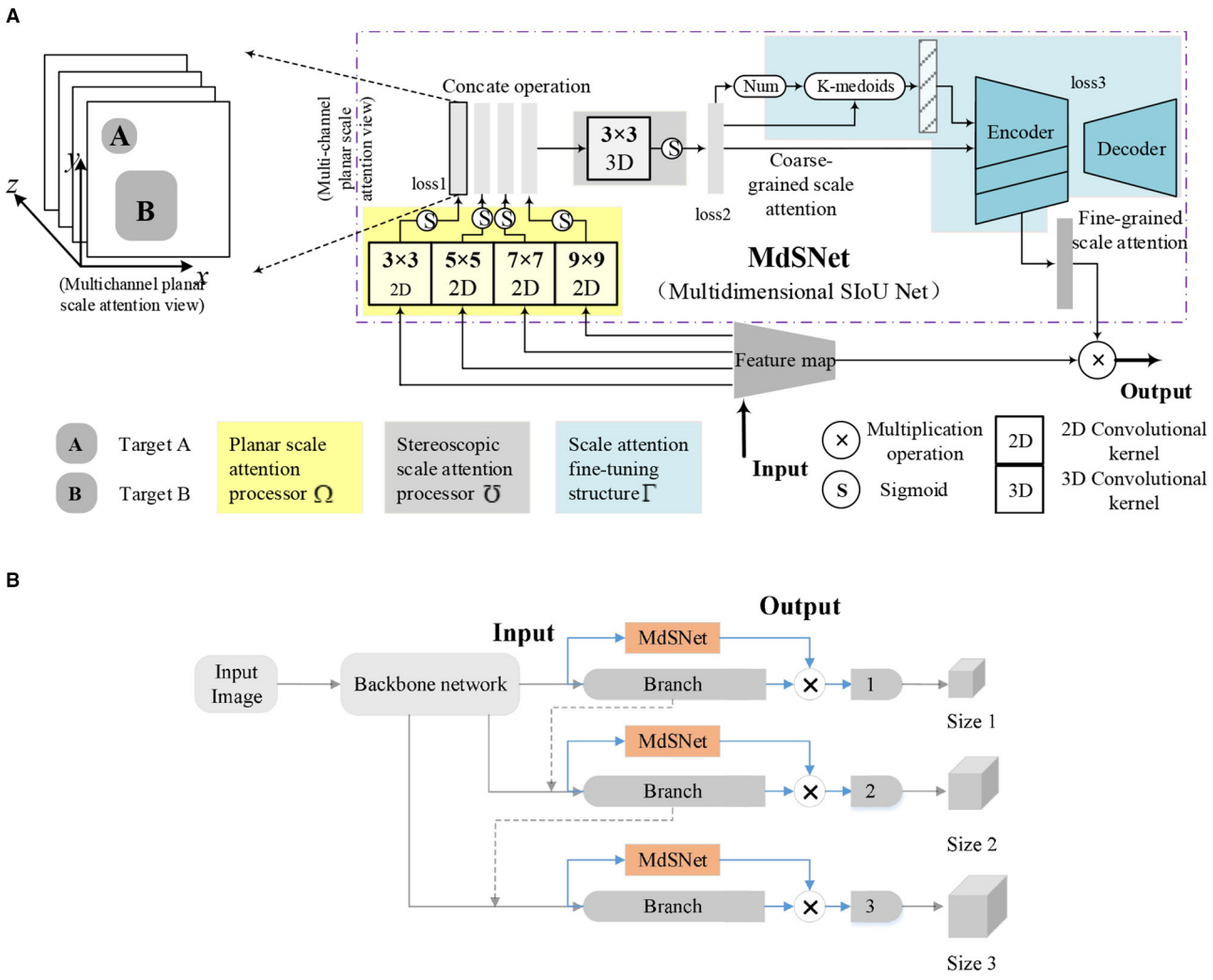

FIGURE 5
Schematic diagram of the MdSNet and its relationship with target detection model. **(A)** Schematic diagram of the MdSNet. **(B)** The relationship between MdSNet and a typical target detection model.

evident aspect ratio characteristics. Consequently, for distance calculation, an ellipse with a major-to-minor axes ratio of τ is constructed, and the ratio τ is adjusted based on the statistical distribution characteristics of the player's bounding box width and height.

$$L_{xy} = \left(\frac{x_0^2}{\varphi} + \frac{x_1^2}{\tau \cdot \varphi} - \frac{y_0^2}{\varphi} - \frac{y_1^2}{\tau \cdot \varphi}\right)^2 \tag{7}$$

$$\varphi = \left(x_0 + \tan\left(\frac{1}{f_{\text{size}}}\right) \cdot x_1\right)^2 \tag{8}$$

$$\tau = \frac{x_1^2}{\varphi - x_0^2} \tag{9}$$

The specific process is depicted in Figure 7. When implementing a codec, the confidence interval within the corresponding bounding box range serves as both the decoder and encoder. The confidence region range is determined using equation (10), where $\overline{A}$ represents the sample mean and $\overline{B}$ represents the interval width. $\overline{B}$ can be computed using equation (11), where $\overline{S}$ represents the square root of the sample variance, and n is the number of sample points. For coarse-grained scale attention planes, once the bounding box information for each target is established, it can be assumed that its scale features follow a normal distribution. Although the true mean and variance of the corresponding statistical distribution are unknown, confidence data within a certain bounding box can be used as a sampling sample to calculate its sample mean $\overline{A}$ and sample variance $\overline{S}$. Consequently, the confidence interval for the statistical mean μ at a confidence level 1-α can be computed. The decoder obtains the necessary bounding box information from the real data labels of the target. Given the fixed scale size of each branch adapted by the MdSNet network module, the encoder acquires the boundary box information from the boundary boxes obtained after multiple length and width expansions or contractions of each scale branch. By utilizing the feature information from the encoder with the narrowest confidence interval range (i.e., the encoder feature with the most concentrated scale feature data), along with the real label information from the decoder, the loss function is solved in accordance with equation (12), this corresponds to the loss3 in

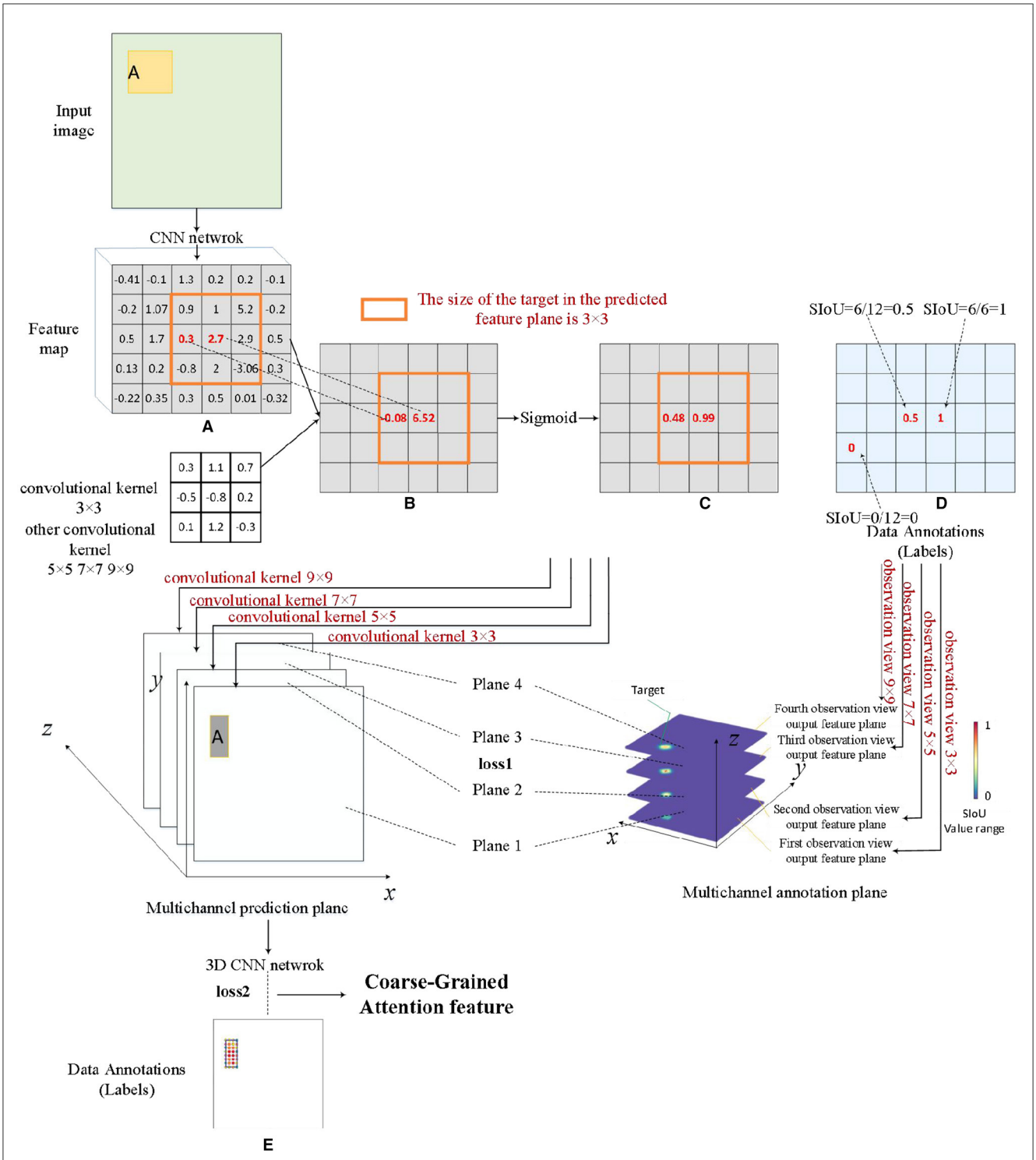

FIGURE 6
Coarse-grained attention feature generation process. **(A)** Feature map, **(B)** single channel feature map, **(C)** feature map after sigmoid function, **(D)** the intermediate process of data annotations operation, and **(E)** data annotations.

the figure.

$$(\overline{A} - \overline{B}, \overline{A} + \overline{B}) \tag{10}$$

$$\overline{B} = \frac{\overline{S}}{\sqrt{n}} t_{\alpha/2}\,(n-1) \tag{11}$$

$$\text{loss} = \sqrt{(\overline{A_r} - \overline{A_p})^2} + \sqrt{(\overline{B_r} - \overline{B_p})^2} \tag{12}$$

3.4 Scale equalization algorithm

The scale equalization algorithm equalizes image scale statistics that approximate a normal distribution. It achieves this by indirectly using the scaling factor $\gamma_{h,w}^{i,j}$, without directly altering the sample bounding box size in the dataset. The algorithm's purpose is to reduce missed detections of relatively

Input: Θ, coarse-grained scale feature plane; $\mathbb{N}$, the predicted number of targets in the feature map; Scale adjustment encoder quantity K and corresponding branch target basic size w_{Anchor}.
Output: Fine-grained scale feature plane.
1: Center of target on Θ :
$C_p = [(cr\_x_1, cr\_y_1), \cdots, (cr\_x_{\mathbb{N}}, cr\_y_{\mathbb{N}})]_{1\times\mathbb{N}} \leftarrow$ Improved K-medoids and $\mathbb{N}$.
2: Use Manhattan distance function $f_{\text{mhd}}(\cdot)$ to sort the orientation of the target:
$[(cp\_x_1, cp\_y_1), \cdots, (cp\_x_{\mathbb{N}}, cp\_y_{\mathbb{N}})]_{1\times\mathbb{N}} = f_{\text{mhd}}(C_p)$,
$C_p = [(cr\_x_1, cr\_y_1), \cdots, (cr\_x_{\mathbb{N}}, cr\_y_{\mathbb{N}})]_{1\times\mathbb{N}}$, C_p predicted target.
$[(ct\_x_1, ct\_y_1), \cdots, (ct\_x_{\mathbb{N}}, ct\_y_{\mathbb{N}})]_{1\times\mathbb{N}} = f_{\text{mhd}}(C_t)$,
$C_t = [(ct\_x_1, ct\_y_1), \cdots, (ct\_x_{\mathbb{N}}, ct\_y_{\mathbb{N}})]_{1\times\mathbb{N}}$, C_t real target.
3: **for** $\lambda \in [1, \mathbb{N}]$ **do**
Regional sample mean $\overline{A}_r^{\lambda}$, Regional sample variance related variable $\overline{B}_r^{\lambda}$
4: **for** $\kappa \in [0, K]$ **do**
5: $p_\kappa = w_{\text{Anchor}} \pm \kappa * \Delta$, perform κ times of scaling.
6: Predict regional sample mean $\overline{A}_{p_\kappa}^{\lambda}$.
7: Predict regional sample variance related variable $\overline{B}_{p_\kappa}^{\lambda}$.
8: $\overline{B}_p^{\lambda} \leftarrow \min(\overline{B}_p^{\lambda}, \overline{B}_{p_\kappa}^{\lambda})$. smallest region sample variance related variable.
9: **end for**
10: loss $\leftarrow$ loss + loss_λ, $\text{loss}_\lambda \leftarrow \text{loss} = \sqrt{(\overline{A}_r - \overline{A}_p)^2} + \sqrt{(\overline{B}_r - \overline{B}_p)^2}$.
11: **end for**

Algorithm 1. Fine-grained scale attention optimization algorithm.

small-scale bounding box targets within the dataset. Drawing inspiration from the image grayscale value equalization algorithm (Acharya and Kumar, 2021), we transform the probability density functions of the image height statistic h and width statistic w, following equations (13) and (14) respectively, to derive new statistics ϕ and ψ. Since h and w are independent of each other, ϕ and ψ are also independent, as indicated by their joint probability density as shown in equation (15).

$$\phi : H(h) = \int_0^h f(h)dh \tag{13}$$

$$\psi : H(w) = \int_0^w f(w)dw \tag{14}$$

$$f(w,h) = f(\phi) \cdot f(\psi) \tag{15}$$

Since both ϕ and ψ follow a uniform distribution after transformation, $f(w,h) = 1$ also adheres to a uniform distribution probability density on $0 \leq w \leq 1$ and $0 \leq h \leq 1$. As a result, the statistical information of the non-balanced scale quantity in the dataset can be effectively balanced. The scaling factor $\gamma_{h,w}^{i,j}$, obtained through the equalization algorithm, and the SIoU label designed in this paper can be multiplied and fused following equation (16). The parameters m_h^i and n_h^i represents the quantity values of the i-th level of height statistics for targets in the source dataset before and after the execution of the algorithm, respectively, while m_w^j and n_w^j represent the quantity values of the j-th level of width statistic for targets in the source dataset before and after algorithm execution. The fundamental principle of this scale equalization lies in the utilization of scaling factors to introduce perturbations during the training process, particularly for targets with a large volume of specific scales, with the aim of mitigating overfitting.

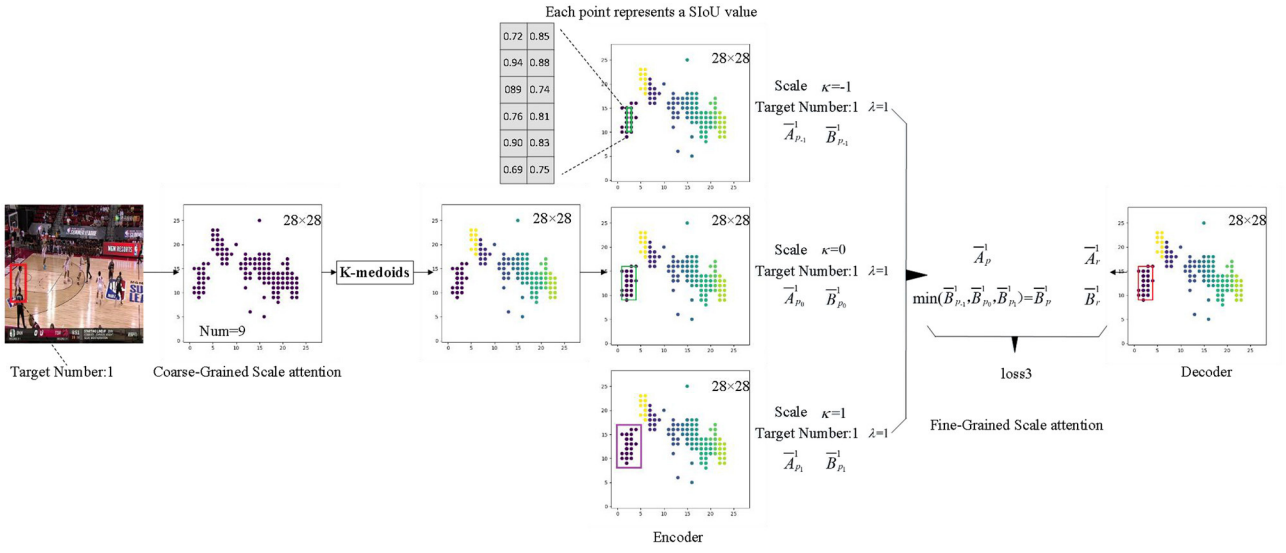

FIGURE 7
The specific generation process of fine-grained attention features.

FIGURE 8
Sample images from the dataset. **(A)** Primary scale samples. **(B)** A limited number of small-scale and large-scale samples.

$$\gamma_{h,w}^{i,j} = \frac{n_h^i}{m_h^i} \cdot \frac{n_w^j}{m_w^j} \tag{16}$$

4 Experiment

4.1 Dataset and statistical distribution analysis

Current competitive game datasets predominantly encompass medium-scale bounding box samples, as depicted in Figure 8A, for player detection, often overlooking the relatively scarce instances of both small-scale and large-scale bounding box samples, illustrated in Figure 8B.

We undertake the reconstruction of a comprehensive competitive competition dataset that encompasses targets of diverse scale bounding boxes. Sample scale equalization, based on Algorithm 2, is then implemented. The dataset comprises three distinct game scenarios: basketball, volleyball, and ice hockey. Each scenario encompasses ∼25 min of valid video sequences, each with a frame rate of 25. Extracting 5% of the image frames from the video, player information is annotated, resulting in around 15K, 13K, and 16K labels for basketball, volleyball, and ice hockey, respectively. The initial scale distribution of the dataset, depicted in Figure 9A, exhibits unevenness and approximately follows a normal distribution. Post-processing with Algorithm 2 yields the scale distribution depicted in Figure 9B, markedly enhancing overall distribution balance compared to the original dataset.

Input: Height and Width of the bounding boxes of all samples and their quantities.
Output: Scaling factor $\gamma_{h,w}^{i,j}$.
$H, W \leftarrow$ Grade height and width at certain intervals respectively.
$m_h, m_w \leftarrow H.\text{size}(), W.\text{size}()$, Count the number of lengths and widths.
$H(h_\xi) = \sum_{\chi=0}^{\varsigma} f(h_\chi) \leftarrow : H(w) = \int_0^w f(w)dw.$
$\phi_\varsigma = H(h_\varsigma) = \sum_{\chi=0}^{\varsigma} f(h_\chi) = \sum_{\chi=0}^{\varsigma} m_h^\chi / m_h$, Length grade that exists after transformation.
$H(w_\chi) = \sum_{\varepsilon=0}^{\xi} f(h_\varepsilon) \leftarrow f(w,h) = f(\phi) \cdot f(\psi)$
$\psi_\xi = H(w_\xi) = \sum_{\varepsilon=0}^{\xi} f(w_\varepsilon) = \sum_{\varepsilon=0}^{\xi} m_w^\varepsilon / m_w$, Widths grade that exists after transformation.
Restores ϕ_ς and ψ_ξ to the standard normalized grade value.
for $i \in [1, m_h]$, $i \in [1, m_w]$ **do**
  $m_h^i \leftarrow H[i], m_w^j \leftarrow W[j]$, The quantity of each grade before transformation.
  $n_h^i \leftarrow f_{\text{hisEqu}}(\phi_\varsigma, m_h, m_h^i), n_w^j \leftarrow f_{\text{hisEqu}}(\psi_\xi, m_w, m_w^j)$, Number of length and width grade after scale equalization, $f_{\text{hisEqu}}(\cdot)$ histogram equalization algorithm.
  $\gamma_{h,w}^{i,j} = \frac{n_h^i}{m_h^i} \cdot \frac{n_w^j}{m_w^j}$, Scaling factor.
end for

Algorithm 2. Sample Scale Equalization Algorithm.

4.2 Experiment on multi-scale attention generation

The process of formulating scale attention predominantly encompasses acquiring two categories of information: the coarse-grained features of multi-scale attention and the fine-grained features of multi-scale attention. In the experiment, the ResNet architecture was adopted as the backbone network, leading to the construction of three scale attention branches: large, medium, and small. The ultimate dimensions of the predicted feature planes were 56 × 56, 28 × 28, and 14 × 14, respectively. To acquire coarse-grained information of multi-scale attention features, the hyperparameters were set as follows: $\mu = 0.85$ and $\sigma = 0.15$, utilized during the generation of training labels. For the fine-grained information of multi-scale attention features, following the principles outlined in Algorithm 1, corresponding quantity fine-tuning encoders were designed for the three scale branches. The visualization outputs of the experience are depicted in Figure 10, where Figure 10A is the original image. These results illustrate that

coarse-grained scale attention, Figure 10B, effectively segregates the scale features of the target for detection and enhances its positional information. Additionally, fine-grained scale attention, Figure 10C, further refines the precision and concentration of potential target positions, building upon the foundation laid by coarse-grained scale attention. Certainly, fine-grained scale attention not only enhances detection accuracy but also results in a several-fold increase in the overall runtime of the network model. This is especially due to the improved K-medoids algorithm, which adds considerable time overhead. Therefore, the scale attention model is better suited for offline video processing, similar to the one investigated in this article.

4.3 Comprehensive experiment

This section presents three comprehensive sets of experiments concerning multi-scale player detection. The first set is ablation experiments focusing on the three fundamental processes outlined in our method, aiming to evaluate the efficacy of each process. The second set involves experiments conducted with a dataset volume of approximately 10%, serving as a preliminary validation of the proposed method's capacity to enhance target detection accuracy. In the third set of experiments, algorithmic comparisons are conducted across various dataset volumes, serving to underscore the limited influence of sample size distribution on the multi-scale attention model.

4.3.1 Ablation experiment

The experimental findings, presented in Figure 11, depict ablation experiments conducted on the three core processes encompassing coarse-grained scale attention, fine-grained scale attention, and scale equalization, as formulated in the methodology of this article.

The evaluation metrics employed in this experiment are computed according to equation (17), where TP denotes the count of correctly predicted positive player instances, FP

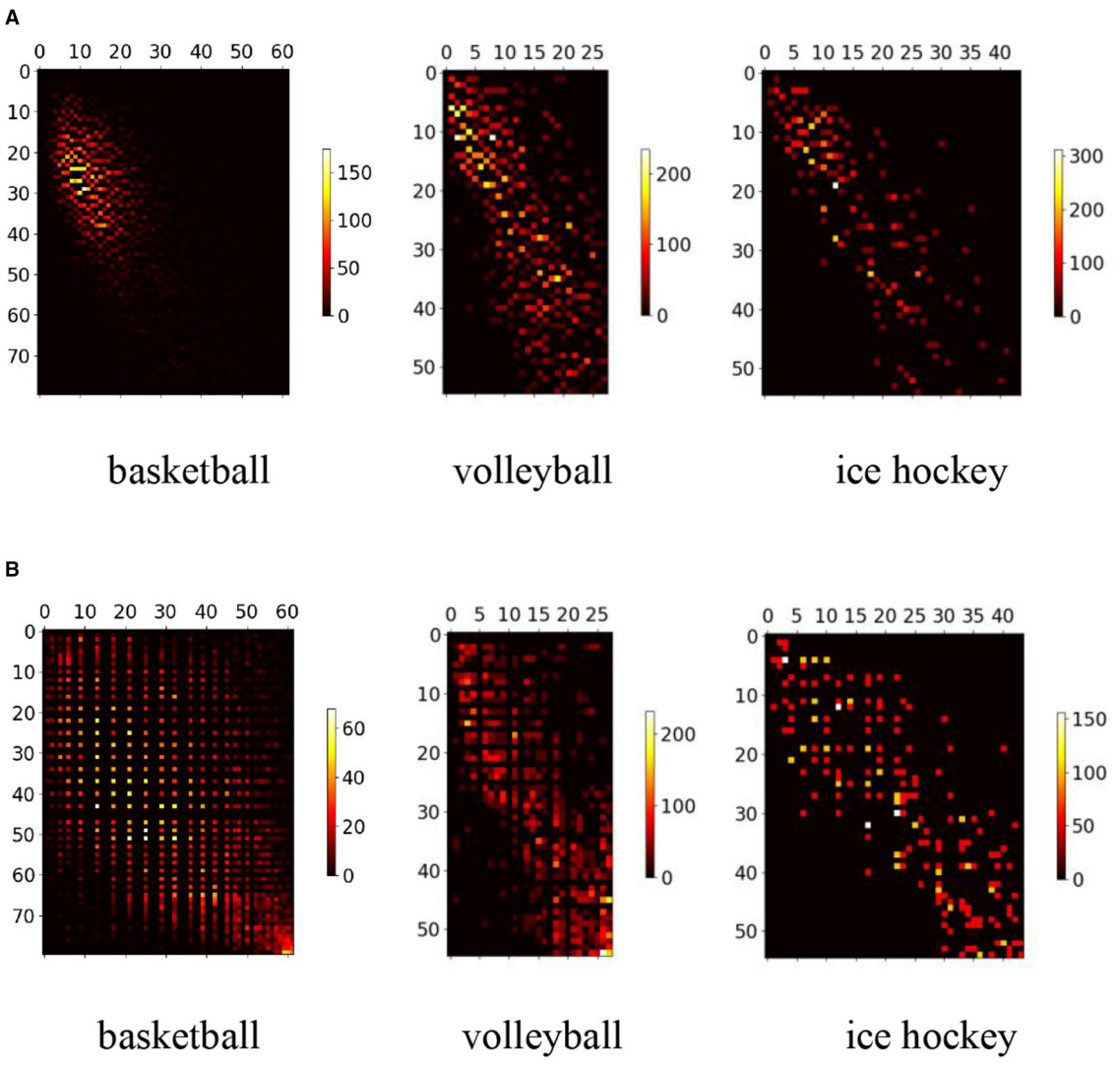

FIGURE 9
The scale distribution of the dataset. **(A)** The scale distribution of the original dataset. **(B)** The scale distribution after dataset scale equalization.

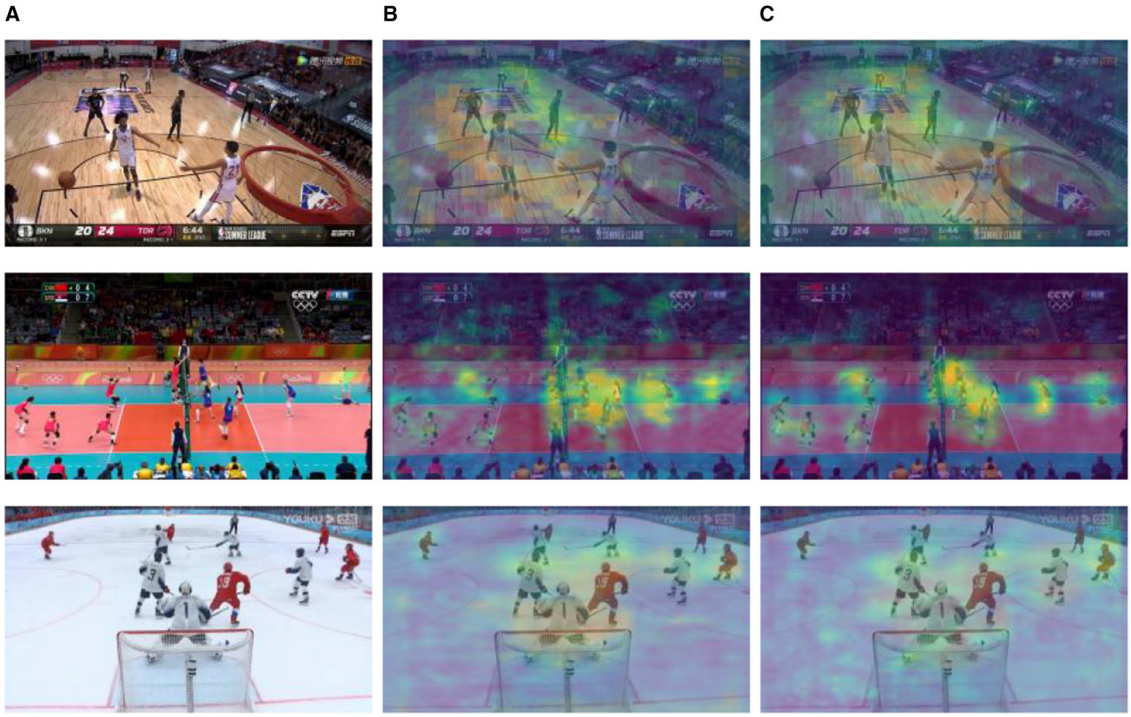

FIGURE 10
Comparison between coarse-grained and fine-grained scales attention. **(A)** Original image **(B)** Coarse-grained scale attention **(C)** Fine-grained scale attention.

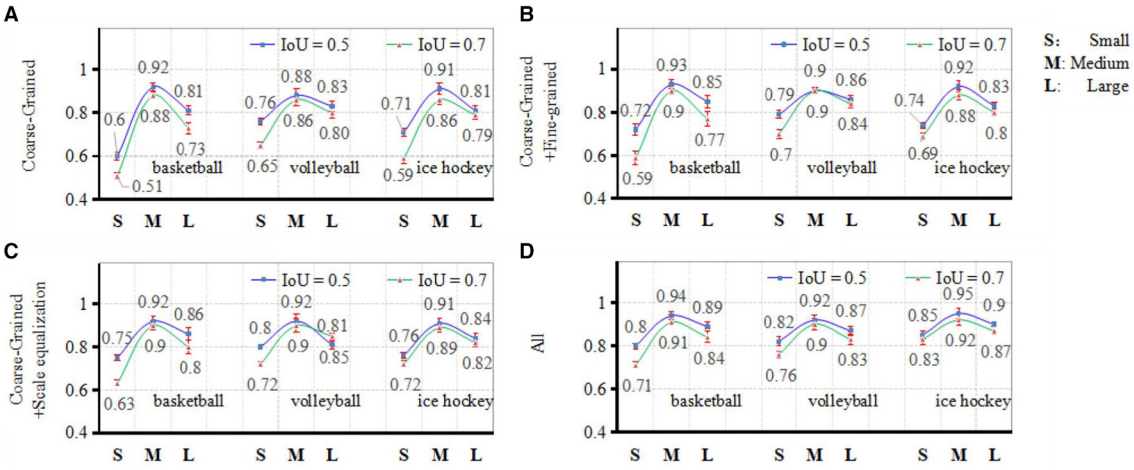

FIGURE 11
Results from ablation experiments comparison. **(A)** Coarse-grained, **(B)** coarse-grained and fine-grained, **(C)** coarse-grained and scale equalization, and **(D)** all.

signifies the count of erroneously predicted positive player instances, and FN represents the count of erroneously predicted negative player instances. In the course of the experiment, the IoU thresholds for player detection were set at 0.5 and 0.7, respectively. The accuracy of target detection was assessed across four scenarios: solely employing coarse-grained scale attention, utilizing both coarse-grained and fine-grained scale attention, incorporating coarse-grained scale attention and the scale equalization algorithm, and integrating all three core processes. Analyzing the results reveals that coarse-grained scale attention serves as the fundamental framework for achieving multi-scale object detection in ball games. Fine-grained attention functions as a secondary refinement of coarse-grained attention, showcasing more pronounced enhancements in detection outcomes particularly under higher IoU requirements. The scale equalization algorithm is particularly effective in enhancing the detection capability for maximum and minimum scale bounding box targets within smaller sample volume,

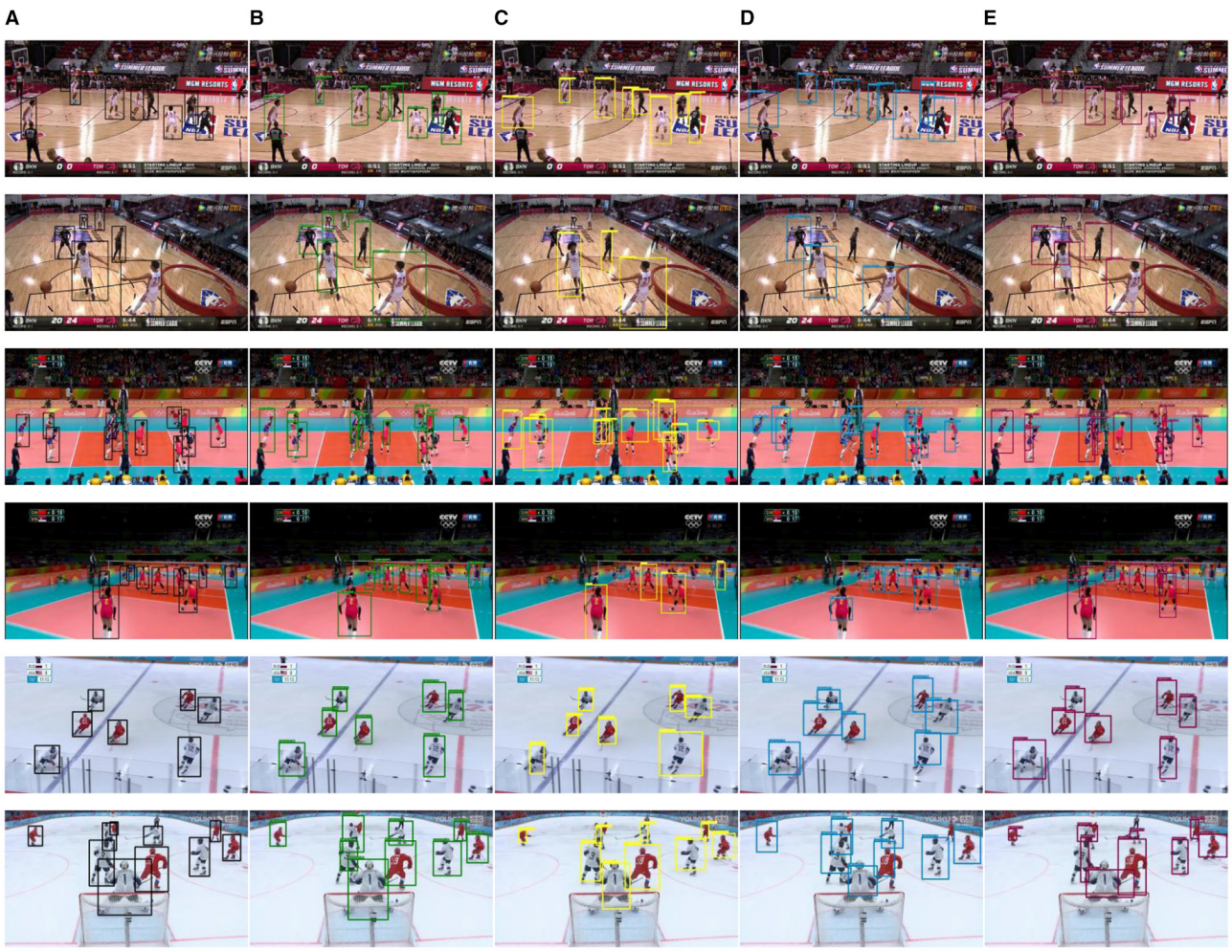

FIGURE 12
Results of player detection comparison. **(A)** Data labels. **(B)** Ours. **(C)** FPN. **(D)** PANet. **(E)** BiFPN.

TABLE 1 **Comparison of player detection normalization results for algorithms with a 10% data volume.**

Algorithm	Data volume proportion	Basketball			Volleyball			Ice hockey		
		S	M	L	S	M	L	S	M	L
YOLOv3+FPN		0.33	0.82	0.48	0.29	0.81	0.43	0.39	0.89	0.51
YOLOv3+PANet		0.38	**0.87**	**0.51**	0.31	0.79	**0.46**	0.41	0.90	0.58
YOLOv3+BiFPN		0.34	0.84	0.47	0.32	**0.82**	0.41	0.40	0.87	**0.60**
RetinaNet+AMF	10%	0.37	0.81	0.44	0.31	0.80	0.44	0.41	0.91	**0.60**
DeepPlayer		0.42	0.84	**0.51**	0.37	0.80	**0.48**	0.47	**0.91**	**0.61**
YOLOVX+ESPHead		**0.44**	0.85	**0.50**	0.32	**0.82**	**0.46**	**0.51**	0.90	0.55
YOLOv6		0.43	**0.88**	0.48	**0.40**	**0.81**	0.45	0.49	**0.92**	0.53
YOLOv3+MdSNet(Ours)		**0.57**	0.86	0.47	**0.51**	0.80	**0.48**	**0.62**	**0.91**	0.56

Bold numbers represent optimal and suboptimal data, respectively.

yielding notably improved effects compared to fine-grained scale attention.

$$ACC=\frac{TP}{TP+FP+FN} \tag{17}$$

4.3.2 Algorithm comparison experiment under low data volume

To provide an initial validation of the capability of multi-scale attention to enhance the accuracy of conventional object detection algorithms, a subset amounting to approximately 10% of

TABLE 2 Comparison of player detection normalization results for algorithms with a 30% data volume.

Algorithm	Data volume proportion	Basketball			Volleyball			Ice hockey		
		S	M	L	S	M	L	S	M	L
YOLOv3+FPN		0.36	0.89	0.52	0.30	0.87	0.46	0.44	0.91	0.60
YOLOv3+PANet		0.37	**0.91**	0.54	0.33	0.86	0.41	0.47	0.90	0.63
YOLOv3+BiFPN		0.40	0.88	0.49	0.36	**0.89**	0.49	0.43	0.93	0.61
RetinaNet+AMF	30%	0.46	0.90	0.54	0.37	**0.88**	0.52	0.50	**0.94**	0.63
DeepPlayer		**0.57**	**0.91**	0.68	0.48	0.86	**0.61**	0.59	0.92	0.64
YOLOVX+ESPHead		0.55	**0.91**	**0.70**	0.44	0.86	0.59	**0.62**	0.91	**0.69**
YOLOv6		**0.57**	0.89	0.62	**0.56**	**0.88**	**0.60**	0.61	**0.94**	0.65
YOLOv3+MdSNet(Ours)		**0.71**	0.87	**0.74**	**0.68**	0.84	**0.72**	**0.70**	0.92	**0.75**

Bold numbers represent optimal and suboptimal data, respectively.

TABLE 3 Comparison of player detection normalization results for algorithms with a 50% data volume.

Algorithm	Data volume proportion	Basketball			Volleyball			Ice hockey		
		S	M	L	S	M	L	S	M	L
YOLOv3+FPN		0.43	0.90	0.55	0.37	0.88	0.49	0.52	0.90	0.65
YOLOv3+PANet		0.41	**0.92**	0.54	0.35	**0.92**	0.47	0.55	**0.93**	0.60
YOLOv3+BiFPN		0.49	0.89	0.57	0.42	0.90	0.50	0.58	0.91	0.61
RetinaNet+AMF	50%	0.52	0.90	0.63	0.46	0.90	0.58	0.64	0.92	0.70
DeepPlayer		0.48	0.91	0.74	0.51	0.91	0.67	0.61	0.92	0.74
YOLOVX+ESPHead		0.61	0.91	**0.75**	0.54	0.91	0.69	0.68	0.92	**0.81**
YOLOv6		**0.65**	**0.92**	0.74	**0.63**	**0.93**	**0.72**	**0.73**	**0.93**	0.77
YOLOv3+MdSNet(Ours)		**0.75**	**0.93**	**0.81**	**0.70**	0.89	**0.79**	**0.78**	**0.94**	**0.84**

Bold numbers represent optimal and suboptimal data, respectively.

TABLE 4 Comparison of player detection normalization results for algorithms with a 100% data volume.

Algorithm	Data volume proportion	Basketball			Volleyball			Ice hockey		
		S	M	L	S	M	L	S	M	L
YOLOv3+FPN		0.48	0.93	0.58	0.41	0.90	0.53	0.65	0.93	0.62
YOLOv3+PANet		0.50	**0.95**	0.55	0.46	**0.93**	0.55	0.63	0.90	0.63
YOLOv3+BiFPN		0.49	0.90	0.60	0.42	0.91	0.51	0.65	**0.96**	0.67
RetinaNet+AMF	100%	0.55	0.93	0.67	0.49	**0.93**	0.62	0.68	**0.97**	0.73
DeepPlayer		0.51	0.92	**0.84**	0.44	**0.94**	0.76	0.72	**0.96**	0.77
YOLOVX+ESPHead		0.64	0.92	0.76	0.58	**0.94**	0.74	0.71	0.93	**0.86**
YOLOv6		**0.69**	0.93	**0.82**	**0.67**	0.92	**0.79**	**0.76**	0.95	0.81
YOLOv3+MdSNet(Ours)		**0.80**	**0.94**	**0.89**	**0.82**	0.92	**0.87**	**0.85**	0.95	**0.90**

Bold numbers represent optimal and suboptimal data, respectively.

the player detection dataset was extracted. Leveraging the YOLOv3 algorithm and pretraining the backbone network on the PETA dataset, comparative experimental results were obtained for the approach presented in this article, the approach augmented with the FPN (Zhao et al., 2019) module, the approach augmented with the PANet (Bochkovskiy et al., 2020) module, and the approach augmented with the BiFPN (Zhang et al., 2021) module. As illustrated in Figure 12, the images in the odd-numbered rows depict the detection results of players enclosed within medium-scale bounding boxes. Conversely, the images in the even-numbered rows encompass the detection outcomes of players enclosed by bounding boxes of maximum or minimum scale.

Analysis reveals that the algorithm proposed by us demonstrates superior detection accuracy for a limited volume subset of extremely small-scale bounding box targets. Moreover, for a relatively small volume subset of extremely large-scale bounding box targets, the IoU indices of targets detected by this algorithm are notably improved. Across the dataset, all algorithms exhibit

comparable detection capabilities for medium-scale bounding box targets. The quantitative comparison results for these observations are tabulated in Table 1.

4.3.3 Comparative experiment of algorithms across varied data volumes

Four additional comparative algorithms were introduced (Lin et al., 2017; Zhang et al., 2020; Ge et al., 2021; Li et al., 2022). Subsequent to training on comprehensive basketball, volleyball, and ice hockey datasets, the accuracy of player detection was computed at an IoU threshold of 0.5. Ultimately, for players of both very small and very large scale bounding box within the dataset, the proposed method showcased improvements of 11%, 7%, 15%, 8%, 9%, and 4%, respectively, in comparison to the optimal method. The comprehensive experimental results are illustrated in Tables 1–4, encompassing the detection quantification outcomes obtained for approximately 10%, 30%, 50%, and the 100% volume datasets, respectively. By further considering the statistical distribution information in Figure 9, it becomes evident that with an equivalent volume of data, the model augmented with both scale attention and the scale equalization algorithm exhibits distinct advantages in the detection of players at the maximum and minimum scale bounding box. This distinction is particularly pronounced in the case of basketball player detection. This observation can be attributed to the relatively limited quantity of minimum and maximum scale bounding box targets present within the basketball player detection dataset, thereby leading to a more pronounced imbalance in scale distribution. Concurrently, it is discernible that with the expansion of dataset volume, the approach delineated by us consistently refines the detection precision for maximum and minimum scale bounding box targets. Nevertheless, the missed detection probability for the other seven algorithms showcases minimal reduction. This outcome is rooted in the possibility that the scale distribution within the sampled dataset may mirror that of the complete dataset. This observation underscores the pronounced reliance of these algorithms on the scale statistical distribution attributes intrinsic to the dataset. Regrettably, they may lack the capability to rectify inaccuracies stemming from scale imbalance. In contrast, the algorithm proposed by us evinces reduced sensitivity to dataset scale balance. It demonstrates a weaker interdependence on the dataset's scale distribution characteristics when compared to the other seven algorithms.

5 Conclusion

This article initiates the concept of SIoU and meticulously scrutinizes its viability as a label for explicitly conveying multi-scale attributes. Subsequently, a network module is devised to extract the multi-dimensional distribution characteristics inherent to SIoU features, leveraging it to bolster the precision of multi-scale object detection within ball team sports. This module primarily encompasses a two-tiered granularity scale attention generation mechanism. The initial tier deploys an array of 2D convolutional kernels to derive numerous planar scale attentions, which are then merged with 3D convolutional kernels to construct spatial scale attention of 3D spatial features, culminating in the creation of a coarse-grained scale attention feature plane. The subsequent tier involves an enhanced K-medoids algorithm, coupled with interval estimation to establish a codec, thereby giving rise to a fine-grained scale attention feature plane. By harnessing label training models to ascertain the interrelated dynamics among SIoU numerical features during the extraction of attention features across multiple 2D plane scale levels and 3D spatial scale dimensions, the prominence of their absolute numerical attributes is diminished. Consequently, the process of scale attention generation becomes less predicated on the intricate scale distribution attributes within the dataset, thereby primarily mitigating the challenge of missed detections pertaining to targets of maximal and minimal scale bounding box. Furthermore, the integration of sample scale equalization algorithms into the model training procedure disrupts the overfitting tendency observed during training for specific scale bounding box targets with abundant instances. This augmentation further enhances the accuracy of multi-scale target detection, particularly for very small and very large scale bounding box players that appear less frequently. Building upon the findings of this current study, future research will place heightened emphasis on unraveling the interpretability and controllability of convolutional neural networks as a means of advancing the capabilities for multi-scale object detection.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

PZ: Writing – original draft. JL: Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Acharya, U. K., and Kumar, S. (2021). Directed searching optimized mean-exposure based sub-image histogram equalization for grayscale image enhancement. *Multimed. Tools Appl.* 80, 24005–24025. doi: 10.1007/s11042-021-10855-7

Akan, S., and Varli, S. (2022). Use of deep learning in soccer videos analysis: survey. *Multim. Syst.* 29, 897–915. doi: 10.1007/s00530-022-01027-0

Bochkovskiy, A., Wang, C., and Liao, H. (2020). YOLOv4: optimal speed and accuracy of object detection. *arXiv [Preprint]. arXiv: 2004.10934.*

Buric, M., Ivasic-Kos, M., and Pobar, M. (2019). "Player tracking in sports videos," in *The 2019 IEEE International Conference on Cloud Computing Technology and Science* (Sydney: IEEE), 334–340.

Ding, B., Zhang, R., and Xu, L. (2023). "U2D2Net: unsupervised unified image dehazing and denoising network for single hazy image enhancement," in *IEEE Transactions on Multimedia*, 1–6. doi: 10.1109/TMM.2023.3263078

Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. (2021). *YOLOX: Exceeding YOLO Series* in *2021*.

He, X. (2022). Application of deep learning in video target tracking of soccer players. *Soft Computing* 26, 10971–10979. doi: 10.1155/2022/3540642

Hurault, S., Ballester, C., and Haro, G. (2020). "Self-supervised small soccer player detection and tracking," in *The 3rd International Workshop on Multimedia Content Analysis in Sports (MMSports '20)* (New York: MMSports), 9–18.

Komorowski, J., Kurzejamski, G., and Sarwas, G. (2020). "Footandball: integrated player and ball de-tector," in *The 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020)* (Valletta: VISIGRAPP), 47–56.

Kong, L., Huang, D., Qin, J., and Wang, Y. (2019). A joint framework for athlete tracking and action recognition in sports videos. *IEEE Trans. Circuits Syst. Video Techn.* 30, 532–548. doi: 10.1109/TCSVT.2019.2893318

Kong, L., Huang, D., and Wang, Y. (2020). Long-term action dependence based hierarchical deep association for multi-athlete tracking in sports videos. *IEEE Trans. Image Proc.* 29, 7957–7969. doi: 10.1109/TIP.2020.3009034

Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., et al. (2022). YOLOv6: a single-stage object detection framework for industrial applications. *arXiv [Preprint]. arXiv: 2209.02976.* doi: 10.48550/arXiv.2209.02976

Lin, T., Goyal, P., Girshick, R., He, K., and Dollar, P. (2017). "Focal Loss for Dense Object Detection," in *The 2017 IEEE International Conference on Computer Vision (ICCV)* (Venice: IEEE), 2999–3007. doi: 10.1109/ICCV.2017.324

Lu, W., Ting, J., Little, J., and Murphy, K. (2013). Learning to track and identify players from broadcast sports videos. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1704–1716. doi: 10.1109/TPAMI.2012.242

Lu, W., Ting, J., Murphy, K., and Little, J. (2011). "Identifying players in broadcast sports videos using conditional random fields," in *The 2011 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)* (Colorado: IEEE), 3249–3256. doi: 10.1109/CVPR.2011.5995562

Nishikawa, Y., Sato, H., and Ozawa, J. (2017). "Performance evaluation of multiple sports player tracking system based on graph optimization," in *The 2017 IEEE International Conference on Big Data (Big Data)* (Boston: IEEE), 2903–2910.

Sah, M., and Direkoglu, C. (2023). Review and evaluation of player detection methods in field sports. *Multimed. Tools Appl.*82, 13141–13165. doi: 10.1007/s11042-021-11071-z

Santhosh, P., and Kaarthick, B. (2019). An automated player detection and tracking in basketball game. *Comp.Mater. Continua* 58, 625–639. doi: 10.32604/cmc.2019.05161

Stein, M., Janetzko, H., Lamprecht, A., Breitkreutz, T., Zimmermann, P., Goldlücke, B., et al. (2018). Bring it to the pitch: combining video and movement data to enhance team sport analysis. *IEEE Trans. Vis. Comput. Graph.* 24, 13–22. doi: 10.1109/TVCG.2017.2745181

Yu, J., Jiang, Y., Wang, Z., Cao, Z., and Huang, T. (2016). "UnitBox: an advanced object detection network," in *The 24th ACM International Conference on Multimedia (MM '16)* (New York: MM), 516–520.

Zhang, C., Tian, Z., Song, J., Zheng, Y., and Xu, B. (2021). "Construction worker hardhat-wearing detection based on an improved BiFPN," in *The 25th International Conference on Pattern Recognition (ICPR)* (Milan: IEEE), 8600–8607. doi: 10.1109/ICPR48806.2021.9412103

Zhang, R., Wu, L., and Yang, Y. (2020). Multi-camera multi-player tracking with deep player identification in sports video deepplyer. *Pattern Recognit.* 102, 107260. doi: 10.1016/j.patcog.2020.107260

Zhang, R., Yang, S., and Zhang, Q. (2022). Graph-based few-shot learning with transformed feature propagation and optimal class allocation. *Neurocomputing* 470, 247–256. doi: 10.1016/j.neucom.2021.10.110

Zhao, B., Zhao, B., Tang, L., Wang, W., and Wu, C. (2019). Multi-scale object detection by top-down and bottom-up feature pyramid network. *J. Syst. Eng. Electron.* 30, 1–12. doi: 10.21629/JSEE.2019.01.01