Advances in Nonlinear Variational Inequalities
ISSN: 1092-910X
Vol 28 No. 7s (2025)

DEFI-Net: Dual-Enhanced Feature Integration for Accurate Multi-Object Tracking in Sports Analytics

Zhihao Zhang[1]，Wan Ahmad Munsif Bin Wan Pa[1*]，Nur Shakila Binti Mazalan[1]，Wenyue Liu[1]

[1]Faculty of Education, Universiti Kebangsaan Malaysia, Bangi, Selangor Malaysia, 43600, Malaysia

Article History:

Received: 15-12-2024

Revised:23-1-2025

Accepted:3-2-2025

Abstract:

Introduction: Multi-object tracking (MOT) technology has significant applications in intelligent sports event analysis, enabling precise tracking and behavior recognition of athletes without human intervention. However, current MOT systems often face two major challenges in high-density interactions and fast-changing dynamic backgrounds. The first challenge is tracking discontinuity caused by motion blur and multiple occlusions in complex backgrounds, and the second is the difficulty in distinguishing targets due to the high similarity between target and background information. To address these issues, we propose a Dual-Enhanced Feature Integration Network (DEFI-Net), which combines background separation and motion prediction strategies to achieve comprehensive improvements in tracking stability and recognition accuracy.

Methods: Firstly, to solve the issue of tracking discontinuity resulting from motion blur and multiple occlusions, we designed a Background Separation Adaptive Module (BSAM). This module leverages adaptive separation techniques to distinguish dynamic backgrounds from target areas, thus reducing background interference and ensuring continuous tracking in complex backgrounds. Secondly, to enhance recognition accuracy when target and background are highly similar, we introduced a Motion Prior Fusion Module (MPFM), which captures historical motion patterns and spatial position priors to accurately predict target positions, improving the model's ability to differentiate similar targets and track them accurately. The innovation of DEFI-Net lies in its dual enhancement through background separation and motion prior integration, enabling robust tracking performance in high-density motion and complex backgrounds.

Results: Experimental results show that DEFI-Net achieves significant performance improvements on public datasets such as LSP and SportsMOT, particularly excelling in scenarios with multiple occlusions and high background similarity.

Conclusions: DEFI-Net technology significantly improves intelligent sports event analysis by addressing challenges like tracking discontinuity and target-background similarity, ensuring stable tracking and accurate recognition even in complex environments. Its practical application enhances real-time performance analysis, tactical decision-making, and injury prevention by enabling precise athlete tracking in high-density and dynamic settings.

Keywords: Multi-object tracking; Sports event analysis; Intelligent sports monitoring

1. Introduction

Multi-objective tracking (MOT) technology[1][2] has important application value in intelligent sports event analysis. Through accurate positioning and behavioral recognition of athletes, MOT technology can provide real-time sports data support and analysis without human intervention, helping to enhance the viewing experience of sports events and improve the performance of athletes[3][4]. At the same time, MOT technology can also provide coaches and sports analysts with more detailed technical statistics and strategic advice to help them develop more scientific training programs and game strategies. In addition, motion tracking in intelligent sports events can not only promote the professional development of competitive sports, but also show a wide range of application prospects in amateur sports activities and sports health monitoring, providing data-driven support for popular sports and national health.

However, the application of MOT in sports events faces many challenges[3][5][6]. In the dynamic context of high-density interactions and rapid changes, existing MOT systems face two major technical difficulties. The first is the problem of tracking discontinuity due to motion blur and multiple occlusions. High-speed movements and complex interactive actions of athletes in sports scenes often cause blurring, while staggered movements between athletes are prone to occlusion, making it difficult for the tracking system to capture the target trajectory consistently and stably[7]. For example, in a scene of intense confrontation, part or all of the athlete's body may be occluded by other players, leading to tracking interruption and thus affecting the complete recording of the motion trajectory. Second, the high similarity between the target and the background information also brings difficulty in recognition. In sports scenarios, the visual characteristics of background elements such as the field, spectators, or other equipment may be similar to those of the athletes. Especially in highly dynamic scenarios, the mixing of the target and the background makes it easy for the system to mistakenly recognize a background object as a target, or misclassify the target as the background. This phenomenon is especially obvious during long-time tracking, as the constant changes and dynamic interference of the background can further exacerbate the model confusion and reduce the accuracy of the recognition.

In addition, existing MOT systems often struggle to effectively capture subtle features when distinguishing between multiple similar-looking athletes[8], thus limiting the ability to recognize similar targets.

To address the aforementioned challenges, we propose a Dual-Enhanced Feature Integration Network (DEFI-Net), which enhances model tracking stability and target recognition accuracy through a combination of background separation and motion priors. Specifically, to tackle motion blur and multiple occlusion problems in complex backgrounds, we design a Background Separation Adaptive Module (BSAM). The BSAM utilizes a spatiotemporal feature extraction network to distinguish dynamic backgrounds from target regions in the input

video. By effectively separating the background and target areas, the BSAM significantly reduces the impact of background interference, ensuring the model's capability to continuously track targets in complex backgrounds. Moreover, the BSAM adapts the separation strategy dynamically, allowing the system to flexibly espond to background changes, thereby improving tracking stability for athletes. Additionally, to enhance recognition accuracy between highly similar targets and backgrounds, we introduce the Motion Prior Fusion Module (MPFM). The MPFM captures prior information on historical motion trajectories and spatial locations to build an understanding of target movement patterns, allowing for accurate prediction of potential positions when targets are occluded or influenced by background interference. Specifically, the MPFM models historical motion and spatial features to capture the movement trends of athletes, using prior fusion to reduce misidentification with similar backgrounds. The introduction of this module significantly improves the model's ability to distinguish similar targets in complex scenarios. In DEFI-Net, the BSAM and MPFM modules work collaboratively in sequence: the BSAM first reduces background interference, providing cleaner target features for subsequent motion prior fusion, while the MPFM further improves target recognition accuracy using historical motion information and location priors. Through this dual-enhancement mechanism, DEFI-Net achieves robust tracking and precise target recognition in high-density movement and complex background scenes. Experimental results show that DEFI-Net significantly outperforms existing methods on the LSP and SportsMOT datasets, particularly in scenarios with multiple occlusions and high background similarity, where DEFI-Net demonstrates stronger continuous tracking and target differentiation capabilities. The innovations of this paper are as follows:

(1) We propose DEFI-Net, a novel framework combining background separation and motion prior integration to address key challenges in multi-object tracking, including tracking discontinuity and high target-background similarity in dynamic sports settings. This dual-enhancement approach enables DEFI-Net to achieve robust tracking and precise target recognition.

(2) To mitigate tracking interruptions from motion blur and occlusions, we design BSAM, which adaptively distinguishes dynamic background elements from target regions. This module reduces background interference, allowing for continuous and stable tracking even in complex, high-density motion scenes.

(3) MPFM integrates historical motion patterns and spatial position priors to enhance target recognition when background similarity is high. By leveraging prior motion information, MPFM accurately predicts target positions and reduces misidentification, effectively differentiating similar-looking targets in dynamic environments.

(4) Experimental results on the LSP and SportsMOT datasets show that DEFI-Net improves tracking accuracy by 1.1% on LSP and achieves substantial gains on SportsMOT.

2. Related Work

2.1 Multi-Object Tracking

Video-based multi-object tracking (MOT) remains a highly active research area within computer vision. Several studies incorporate CNN-based detectors for the tracking phase[9] while others use global optimization techniques. For example, joint segmentation and tracking of multiple objects are explored in[10], and in[11], both full-body and head detectors are combined to enhance performance. In[12], Convolutional Neural Networks (CNNs) are integrated with a Temporal-Flow-Fields method to improve tracking. An alternative line of tracking methods focuses on Discriminant Correlation Filters (DCF), striking a balance between accuracy and computational efficiency. These methods typically involve an initial feature extraction phase followed by correlation filtering. Early implementations utilized hand-crafted features like HoG, while more recent models employ deep learning features from pretrained networks. End-to-end feature learning for tracking has shown further improvements in performance[13]. Additionally, there is a growing trend towards unsupervised training in deep learning-based tracking methods[14][15].

2.2 Multi-Object Tracking in Sports

Monitoring player movements in team sports has gained significant attention, not only for automating game statistics recording but also to provide sports analysts with comprehensive insights through video-based scene analysis. Unlike pedestriantracking, multi-object tracking (MOT) in sports is considerably more complex. This complexity arises from factors such as the players' rapid and unpredictable movements, the visual similarity among teammates, and frequent occlusions due to the high intensity of the games. Most recent approaches to sports MOT adopt a tracking-by-detection framework, often incorporating re-identification networks to produce embedding features for associating players across frames. For instance, Vats et al.[16] integrate team classification with player identification to enhance tracking accuracy in hockey. Similarly, Yang et al.[17] and Maglo et al.[18] demonstrate that localizing the field along with players in football can lead to more precise tracking outcomes. In basketball, Sangüesa et al.[19] utilize human pose and action recognition as embedding features to improve tracking performance. Additionally, Huang et al.[20] combine OC-SORT[21] with appearance-based post-processing for tracking across various sports, including basketball, volleyball, and football.

3. Method

In this section, we detail the architecture and operational mechanisms of the proposed Dual-Enhanced Feature Integration Network (DEFI-Net) for multi-object tracking (MOT) in intelligent sports event analysis. DEFI-Net is designed to address the challenges of high-density interactions and rapidly changing dynamic backgrounds by integrating two primary modules: the Background Separation Adaptive Module (BSAM) and the Motion Prior Fusion Module (MPFM). These modules collaboratively enhance tracking stability and recognition accuracy in complex, dynamic environments.

3.1 Overall Architecture of DEFI-Net

DEFI-Net is structured as an end-to-end framework that integrates feature extraction, background separation, motion prediction, and target association processes to achieve robust multi-object tracking. The architecture consists of the following main components:

(1) Feature Extraction Backbone: Utilizes a Convolutional Neural Network (CNN) to extract high-level features from input video frames.

(2) Background Separation Adaptive Module (BSAM): Separates dynamic backgrounds from target regions to mitigate tracking discontinuities.

(3) Motion Prior Fusion Module (MPFM): Incorporates historical motion patterns and spatial position priors to enhance target recognition accuracy.

(4) Association and Tracking Head: Associates detected targets across frames using the enhanced features from BSAM and MPFM to maintain consistent tracking identities.The data flow through DEFI-Net is illustrated in **Figure 1**.

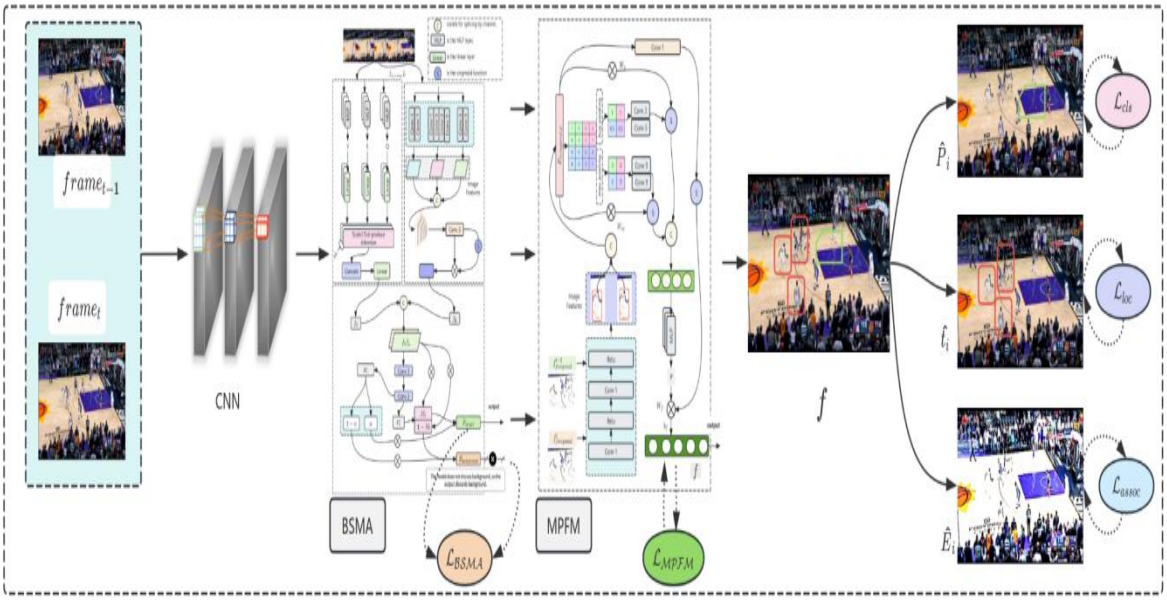


Figure 1 Overall architecture of DEFI-Net

3.2 Background Separation Adaptive Module (BSAM)

BSAM addresses tracking discontinuities caused by motion blur and multiple occlusions in complex backgrounds by adaptively separating dynamic background elements from target regions. This separation reduces background interference and ensures continuous tracking, shown in

3.2.1 Spatiotemporal Feature Extraction

BSAM leverages a spatiotemporal feature extraction network to capture both spatial and temporal information from consecutive video frames. Let I_t denote the input frame at time t, and I_{t-1} the preceding frame. The spatiotemporal features F_t are extracted as follows:

$$F_t = CNN_{spatiotemporal}(I_t, I_{t-1}) \qquad (1)$$

3.2.2 Adaptive Background Separation

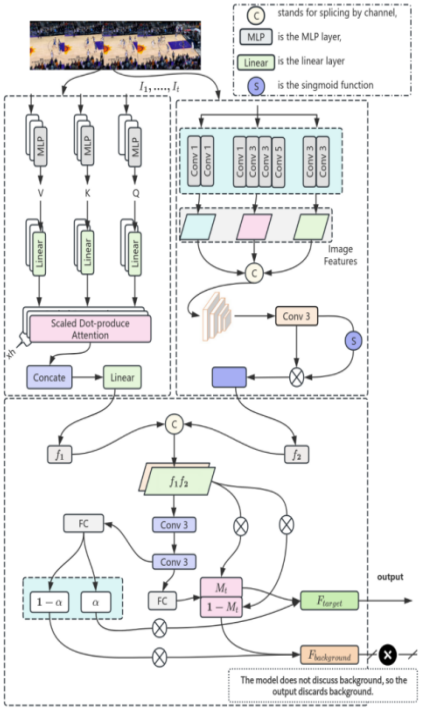
The core functionality of BSAM lies in distinguishing dynamic backgrounds from target regions through an adaptive separation mechanism. The separation process is mathematically formulated as:

$$M_t = \sigma(W_b \cdot F_t + b_b) \qquad (2)$$

$$F_{target,t} = F_t \odot M_t \qquad (3)$$

$$F_{background,t} = F_t \odot (1\text{-}M_t) \qquad (4)$$

Where M_t is the separation mask at time , computed using a sigmoid activation function σ . W_b and b_b are learnable weights and biases for the background separation layer. $\odot$ denotes element-wise multiplication. $F_{target,t}$ and $F_{background,t}$ are the separated target and background feature maps, respectively.



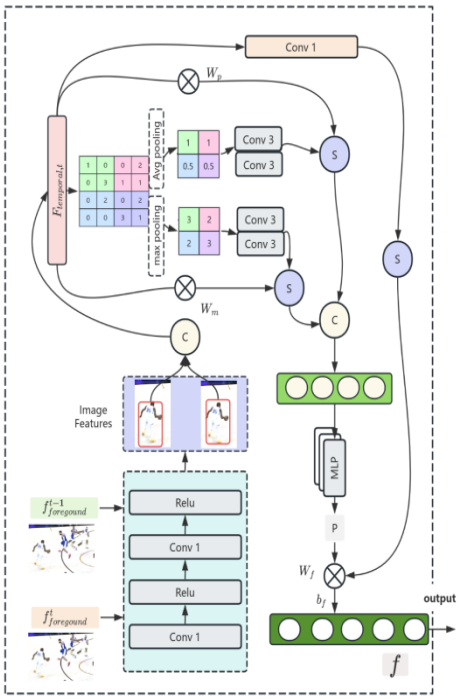


Figure 2 Background Separation Adaptive Module

Figure 3 Motion Prior Fusion Module

3.2.3 Dynamic Adaptation

To handle dynamic changes in the background, BSAM incorporates a dynamic adaptation mechanism that updates the separation strategy based on recent feature variations. This is achieved through a temporal smoothing approach:

$$M_t = \alpha M_{t-1} + (1-\alpha)\sigma(W_b \cdot F_t + b_b) \quad (5)$$

Where α is a smoothing factor ($0 < \alpha < 1$).

3.3 Motion Prior Fusion Module (MPFM)

MPFM enhances recognition accuracy by integrating historical motion patterns and spatial position priors. Instead of using a separate LSTM network, MPFM employs a temporal convolutional approach to capture spatiotemporal dependencies within the network architecture, shown in **Error! Reference source not found.**.

3.3.1 Temporal Feature Integration

MPFM utilizes temporal convolutions to encode historical motion information directly within the network. Let $F_{target,t}$ be the target feature map from BSAM at time t , and $F_{\text{target,t-n}}$ represent the feature maps from the previous N frames. The temporal integration is performed as:

$$F_{temporal,t} = Convtemporal\ (\{F_{target,t-n}\}_n^N = 1) \quad (6)$$

Where $Convtemporal$ denotes a temporal convolutional layer that aggregates information across N frames.

3.3.2 Spatial Position Priors

Spatial position priors capture the typical spatial distribution and movement patterns of targets. These priors are integrated by learning a position embedding based on target locations (xt, yt):

$$P_t = PositionEmbedding(xt,yt) \quad (7)$$

3.3.3 Motion Prediction

MPFM predicts potential target positions using the integrated temporal features and spatial priors. The predicted position P_t is given by:

$$P_t = W_m \cdot F_{temporal,t} + W_p \cdot P_t + b_m \quad (8)$$

Where W_m and W_p are learnable weight matrices, and bm is a bias term.

3.3.4 Feature Fusion

The predicted positions are fused with the current target features to enhance discrimination:

$$F_{enhanced\ t} = Concat(F_{target,t}, P_t \cdot W_f + b_f) \quad (9)$$

Where W_f and b_f are learnable parameters for the fusion layer.

3.4 Integration of BSAM and MPFM in DEFI-Net

DEFI-Net integrates BSAM and MPFM in a sequential manner to leverage their complementary strengths. The integration process is as follows:

(1) Background Separation: The input frame I_t is processed by the feature extraction backbone to obtain F_t. BSAM then separates F_t into $F_{target,t}$ and $F_{background,t}$.

(2) Motion Prior Fusion: $F_{target,t}$ is fed into MPFM along with historical motion data to obtain $F_{enhanced,t}$.

(3) Tracking and Association: The enhanced features $F_{enhanced,t}$ are used for target detection and association across frames, maintaining consistent tracking identities.

3.5 Loss Functions

To effectively train DEFI-Net, we employ a comprehensive loss function framework that optimizes both background separation and motion prediction, ensuring robust multi-object tracking performance. The loss functions are designed to address the specific challenges tackled by the Background Separation Adaptive Module (BSAM) and the Motion Prior Fusion Module (MPFM).

3.5.1 Background Separation Loss

The BSAM is trained using a binary cross-entropy loss to accurately distinguish between target and background regions. This loss compares the predicted separation mask M_t with the ground truth labels y, encouraging the model to minimize misclassification of pixels.

$$L_{BSAM} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(M_{t,i}) + (1-y_i)\log(1-M_{t,i})] \qquad (10)$$

Where N is the number of pixels, y_i is the ground truth label for pixel (1 for target, 0 for background), M_t, i is the predicted mask value for pixel at time t.

3.5.2 Motion Prediction Loss

The MPFM utilizes a Mean Squared Error (MSE) loss to minimize the difference between the predicted target positions P_t and the actual ground truth positions P_t. This ensures accurate motion prediction and enhances the model's ability to track targets reliably.

$$L_{MPFM} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{P}_{t,i} - P_{t,i}\right|^2 \qquad (11)$$

Where $P_{t,i}$ is the ground truth position of target t at time t, $\hat{P}_t$, i is the predicted position of target i at time t.

3.5.3 Detection Loss

DEFI-Net employs a detection component responsible for identifying and localizing objects within each frame. The detection loss comprises two parts: classification loss and localization loss.

$$L_{\det ect} = L_{cls} + L_{loc} \qquad (12)$$

Where L_{cls} is the classification loss, typically implemented as cross-entropy loss, which measures the accuracy of object category predictions. L_{loc} is the localization loss, commonly implemented as Smooth L1 loss,which measures the accuracy of bounding box predictions.

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(P_i) + (1-y_i)\log(1-P_i)] \qquad (13)$$

$$L_{loc} = \frac{1}{N}\sum_{i=1}^{N} SmoothL1(t_i - \hat{t}_i) \qquad (14)$$

Where P_i is the predicted probability for class i. t_i and $\hat{t}_i$ are the ground truth and predicted bounding box coordinates for target i, respectively.

3.5.4 Association Loss

The association loss ensures consistent tracking of object identities across frames. This is crucial for maintaining the continuity of object trajectories, especially in scenarios with occlusions and high object density. The association loss can be implemented using an identity embedding loss, which penalizes discrepancies between the identity embeddings of the same object across different frames.

$$L_{assoc} = \frac{1}{N}\sum_{i=1}^{N}\left|E_i - \hat{E}_i\right|^2 \qquad (15)$$

Where E_i is the ground truth identity embedding for target i. E_i is the predicted identity embedding for target i.

3.5.5 Overall Tracking Loss

The overall tracking loss integrates the detection loss $L_{\det ect}$, association loss L_{assoc}, along with the BSAM and MPFM losses. Weighting coefficients λ_1 and λ_2 are introduced to balance the contributions of each loss component during optimization.

$$L_{total} = L_{\det ect} + L_{assoc} + \lambda_1 L_{BSAM} + \lambda_2 L_{MPFM} \qquad (16)$$

This combined loss function framework ensures that DEFI-Net simultaneously learns to effectively separate backgrounds, accurately predict target motions, detect objects precisely, and maintain consistent tracking identities across frames. By balancing these components, DEFI-Net achieves enhanced tracking stability and recognition accuracy in complex and dynamic sports environments.

4. Experiments

4.1 Dataset Description

The SportsMOT[7] dataset is an advanced,multi-person video dataset specifically created to support research in spatio-temporal action detection. It includes a diverse set of 66 detailed action categories spanning four distinct types of sports, offering a rich source of labeled data with 37,701 action instances and 902,000 bounding boxes derived from 3,200 video clips. The annotations provide precise spatio-temporal localization for each action at a frame rate of 25 frames per second, capturing the intricacies of dynamic sports environments. This high-density labeling not only accommodates the complexity of real-world sports scenarios, where multiple individuals engage in various actions concurrently, but also enables in-depth analysis of challenging multi-object tracking and action recognition tasks.

The Leeds Sports Pose (LSP)[22] dataset provides a valuable resource for research on human pose estimation, particularly in sports contexts. Comprising 2,000 images, each carefully annotated with detailed human body landmarks, this dataset captures athletes engaged in a range of sports activities, showcasing complex and dynamic body poses. Each image includes precise annotations for key joint positions, allowing for an in-depth analysis of human movement and posture during athletic performance.

4.2 Implementation Details

We implemented DEFI-Net using the PyTorch framework, initializing the ResNet-50 backbone with ImageNet pre-trained weights. The model was first trained on the LSP dataset for 50 epochs with an initial learning rate of 1e-4, which was reduced to 1e-5 at the 30th epoch. Training then proceeded on the SportsMOT dataset for an additional 50 epochs using our Dual-Enhanced Feature Integration approach. The input image size was set to 1280×720 pixels. We utilized the Adam optimizer with an initial learning rate of 0.001, dynamically adjusting it during training to enhance convergence. To prevent overfitting, Batch Normalization and Dropout (rate=0.5) were applied, and Early Stopping was employed if the validation loss did not improve for 10 consecutive epochs. For the SportsMOT dataset, we followed established guidelines and used frame-mAP and video-mAP to measure action detection performance. For video-mAP, we applied the 3D IoU, calculated as the temporal IoU of two tracks multiplied by the average IoU of the overlapping frames. The evaluation thresholds were set at 0.5 for frame-mAP and at 0.2 and 0.5 for video-mAP. Regarding the LSP dataset, our chosen evaluation metric was the Percentage of Correct Keypoints (PCK), with an estimate deemed accurate if it was within 5% of the total image size (PCK@0.05).

4.3 Results

The experimental results on the Leeds Sports Pose (LSP) dataset, shown in **Table 1**, highlight DEFI-Net’s superior performance in keypoint detection compared to state-of-the-art methods. DEFI-Net achieves the highest PCK@0.05 scores for all keypoints, including shoulder (70.7%), elbow (85.8%), wrist (84.4%), hip (86.3%), knee (86.2%), and ankle (85.2%), with an average of 83.1%. This surpasses the nearest competitor, UDAPE, by 1.1%. The enhanced performance of DEFI-Net is primarily due to its Background Separation Adaptive Module (BSAM), which effectively reduces background interference, and the Motion Prior Fusion Module (MPFM), which improves target recognition accuracy by leveraging

historical motion and spatial information. These innovations enable DEFI-Net to maintain robust tracking and precise recognition in scenarios with high-density interactions and complex backgrounds.

Table 1 PCK@0.05 on LSP.

Method	Sld	Elb	Wrist	Hip	Knee	Ankle	Avg
DAN[23]	52.2	62.9	58.9	71.0	68.1	65.1	**63.0**
DD[24]	28.4	65.9	56.8	75.0	74.3	73.9	**62.4**
RegDA[25]	62.7	76.7	71.1	81.0	80.3	75.3	**74.6**
CCSSL[26]	36.8	66.3	63.9	59.6	67.3	70.4	**60.7**
UDAPE[27]	69.2	84.9	83.3	85.5	84.7	84.3	**82.0**
DEFI- Net	**70.7**	**85.8**	**84.4**	**86.3**	**86.2**	**85.2**	**83.1**

The experimental results presented in **Table 2** demonstrate that DEFI-Net achieves superior performance across multiple multi-object tracking metrics compared to existing state-of-the-art methods. DEFI-Net attains the highest scores in IDF1 (73.3), AssA (61.3),MOTA (96.5), DetA (88.4), and LocA (93.3), surpassing competitors such as TransTrack and ByteTrack. Although TransTrack records a slightly higher HOTA score (68.9) compared to DEFI-Net's 64.3, DEFI-Net still outperforms other methods like ByteTrack (62.8) and CenterTrack (62.7) in this metric. The enhanced performance of DEFI-Net is primarily due to its Dual-Enhanced Feature Integration Network, which effectively reduces background interference through the Background Separation Adaptive Module (BSAM) and improves target recognition accuracy with the Motion Prior Fusion Module (MPFM). These innovations enable DEFI-Net to maintain robust tracking and high recognition accuracy in complex, high-density scenarios, validating its effectiveness and superiority in multi-object tracking tasks. We performed hyperparameter experiments on the LSP dataset to explore how varying these parameters impacts our experimental outcomes. The sensitivity analysis result sare depicted in the **Figure 4**. As illustrated, our method maintains consistent accuracy across a range of hyperparameter values, indicating a lack of sensitivity to these changes. This robustness highlights our model's reliability and its ability to sustain high performance without extensive hyperparameter fine-tuning.

Table 2 Tracking performance SportsMOT.

Method	HOTA	IDF1	AssA	MOTA	DetA	LocA
CenterTrack[28]	62.7	60.0	48.0	90.8	82.1	90.8
FairMOT[29]	49.3	53.5	34.7	86.5	70.2	83.9
QDTrack[30]	60.4	62.3	47.2	90.1	77.5	88.0

TransTrack[31]	68.9	71.5	57.5	92.6	82.7	91.0
GTR[32]	54.5	55.8	45.9	67.9	64.8	89.0
ByteTrack[4]	62.8	69.8	51.2	94.1	77.1	85.6
DEFI-Net	**64.3**	**73.3**	**61.3**	**96.5**	**88.4**	**93.3**

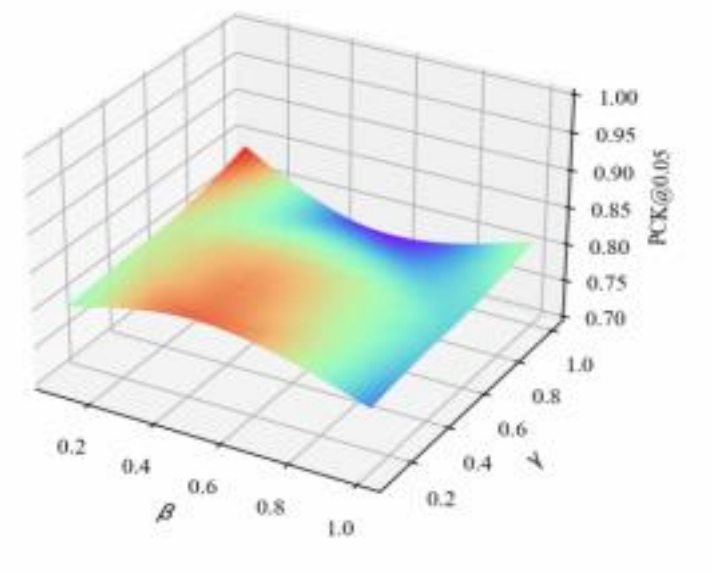


Figure 4 Hyperparametric sensitivity analysis in LSP dataset

The ablation study results in Table 3illustrate the individual and combined contributions of the Background Separation Adaptive Module (BSAM) and the Motion Prior Fusion Module (MPFM) to DEFI-Net's performance. The baseline model without BSAM and MPFM achieves a PCK@0.05 score of 61.3. When BSAM is incorporated alone (configuration b),the score increases substantially to 78.4, demonstrating its effectiveness in reducing background interference and enhancing tracking continuity. Similarly, adding only MPFM (configuration c) raises the score to 80.2, highlighting its role in improving target recognition accuracy by leveraging historical motion and spatial information. When both BSAM and MPFM are integrated in DEFI-Net, the PCK@0.05 score further improves to 83.1, showcasing the synergistic effect of the dual-enhancement strategy. This combined approach significantly enhances tracking stability and recognition precision, validating the effectiveness of both modules in addressing the challenges of high-density interactions and complex backgrounds in multi-object tracking.

Table 3 Ablation study results in our proposed method.

Component	BSAM	MPFM	PCK@0.05
(a)			61.3
(b)	√		78.4
(c)		√	80.2

DEFI-Net	√	√	**83.1**

Our visualization results on SportsMOT are shown in **Figure 5**. The visualizations of DEFI-Net applied to basketball, volleyball, and soccer demonstrate its high accuracy and robustness in tracking multiple athletes across diverse and dynamic sports environments. The model effectively handles fast-paced actions, frequent occlusions, and complex interactions, consistently maintaining precise detection and tracking performance, which validates its applicability for intelligent sports monitoring systems.

Figure 5 Visualization of DEFI-Net multi-object tracking performance in basketball, volleyball, and soccer.

The visualizations demonstrate the impact of the Background Separation Adaptive Module (BSAM) on feature separation in DEFI-Net. In the "After Separation" scatter plot, target (red) and background (blue) features are distinctly separated, highlighting BSAM's effectiveness in minimizing background interference. This clear distinction allows the model to focus more accurately on relevant target information, thereby enhancing detection accuracy and tracking robustness. The improved feature separation confirms BSAM's crucial role in elevating DEFI-Net's overall performance in complex multi-object tracking scenarios with high-density interactions and dynamic backgrounds **Figure 6**.

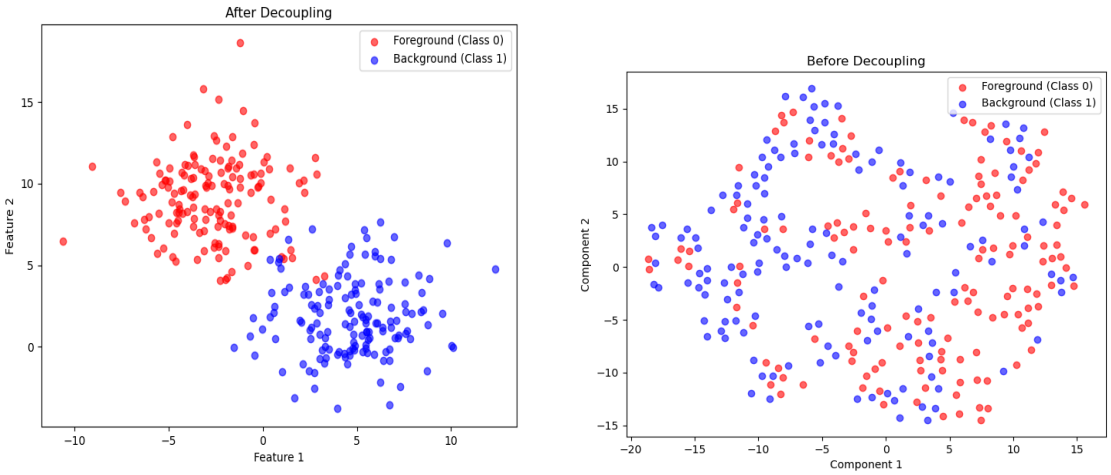


Figure 6 Visualization of features before and after decoupling.

5. conclusion

In this paper, we introduced the Dual-Enhanced Feature Integration Network (DEFI-Net) to address critical challenges in multi-object tracking for intelligent sports event analysis, specifically targeting issues of tracking discontinuity and high target-background similarity in dynamic sports settings. By combining a Background Separation Adaptive Module (BSAM) with a Motion Prior Fusion Module (MPFM), DEFI-Net achieves enhanced tracking stability and recognition accuracy. BSAM effectively mitigates background interference by adaptively separating dynamic backgrounds from target regions, while MPFM leverages historical motion and spatial priors to accurately predict target locations in scenarios with high background similarity. Experimental results on public datasets, including LSP and SportsMOT, demonstrate that DEFI-Net consistently outperforms existing approaches, showing superior robustness in complex, high-density motion environments with significant improvements in tracking continuity and target differentiation.

The proposed DEFI-Net technology holds significant practical value in the context of intelligent sports event analysis. By leveraging dual-enhancement techniques—background separation and motion prediction—this system addresses key challenges in tracking athletes, such as tracking discontinuity caused by motion blur and occlusions, and the difficulty of distinguishing targets due to background similarities. These features are particularly valuable in sports scenarios where athletes often move in high-density groups or in environments with complex and dynamic backgrounds, such as team sports or crowded arenas.

In practical terms, DEFI-Net's ability to maintain stable tracking and improve recognition accuracy enhances the analysis of athlete behavior and performance in real-time. This can lead to more precise performance evaluations, tactical insights, and even injury prevention by enabling more effective monitoring of athlete movement patterns. Furthermore, it can aid coaches in making data-driven decisions, improve the accuracy of automated video highlights, and facilitate advanced analytics in both professional and amateur sports settings. By overcoming the challenges of occlusions and target-background similarity, DEFI-Net

contributes to the development of more efficient, intelligent sports tracking systems that enhance both training and competition experiences.

References

[1] Aharon, N., Orfaig, R. & Bobrovsky, B.-Z. Bot-sort: Robust associations multi-pedestrian tracking. arXiv preprint arXiv:2206.14651 (2022).

[2] Du, Y. et al. Strongsort: Make deepsort great again. IEEE Transactions on Multimed. 25, 8725‑8737 (2023).

[3] Lei, M. & Wang, X. Epps: Advanced polyp segmentation via edge information injection and selective feature decoupling. arXiv preprint arXiv:2405.11846 (2024).

[4] Zhang, Y. et al. Bytetrack: Multi-object tracking by associating every detection box. In European conference on computer vision, 1‑21 (Springer, 2022).

[5] Ciaparrone, G. et al. Deep learning in video multi-object tracking: A survey. Neurocomputing 381, 61‑88 (2020).

[6] Hassan, S., Mujtaba, G., Rajput, A. & Fatima, N. Multi-object tracking: a systematic literature review. Multimed. Tools Appl. 83, 43439‑43492 (2024).

[7] Cui, Y. et al. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 9921‑9931 (2023).

[8] Wojke, N., Bewley, A. & Paulus, D. Simple online and realtime tracking with a deep association metric. In 2017 IEEE international conference on image processing (ICIP), 3645‑3649 (IEEE, 2017).

[9] Cao, J., Pang, J., Weng, X., Khirodkar, R. & Kitani, K. Observation-centric sort: Rethinking sort for robust multi-object tracking. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 9686‑9696 (2023).

[10] Ramanathan, V. et al. Detecting events and key actors in multi-person videos. In Proceedings of the IEEE conference on computer vision and pattern recognition, 3043‑3053 (2016).

[11] Milan, A., Leal-Taixé, L., Schindler, K. & Reid,I. Joint tracking and segmentation of multiple targets. In Proceedings of the IEEE conference on computer vision and pattern recognition, 5397‑5406 (2015).

[12] Henschel, R., Leal-Taixé, L., Cremers, D. & Rosenhahn, B. Fusion of head and full-body detectors for multi-object tracking. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 1428‑1437(2018).

[13] Doering, A., Iqbal, U. & Gall, J. Joint flow: Temporal flow fields for multi person tracking. arXiv preprint arXiv:1805.04596.(2018).

[14] Wang, Q., Gao, J., Xing, J., Zhang, M. & Hu, W. Dcfnet: Discriminant correlation filters network for visual tracking. arXivpreprint arXiv:1704.04057 (2017).

[15] Wang, X., Jabri, A. & Efros, A. A. Learning correspondence from the cycle-consistency of time. In Proceedings of the
IEEE/CVF conference on computer vision and pattern recognition, 2566‑2576 (2019).

[16] Huang, H.-W. et al. Iterative scale-up expansioniou and deep features association for multi-object tracking in sports. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 163‑172 (2024).

[17] Vats, K., Walters, P., Fani, M., Clausi, D. A. & Zelek, J. S. Player tracking and identification in ice hockey. Expert. Systems with applications 213, 119250 (2023).

[18] Yang, Y., Zhang, R., Wu, W., Peng, Y. & Xu, M. Multi-camera sports players 3d localization with identification reasoning.In 2020 25th International Conference on Pattern Recognition (ICPR), 4497‑4504 (IEEE, 2021).

[19] Maglo, A., Orcesi, A. & Pham, Q.-C. Efficient tracking of team sport players with few game-specific annotations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3461‑3471 (2022).

[20] Arbués-Sangüesa, A., Ballester, C. & Haro, G. Single-camera basketball tracker through pose and semantic feature fusion.arXiv preprint arXiv:1906.02042 (2019).

[21] Huang, H.-W. et al. Observation centric and central distance recovery for athlete tracking. In Proceedings of the IEEE/CVFWinter Conference on Applications of Computer Vision, 454‑460 (2023).

[22] Johnson, S. & Everingham, M. Learning effective human pose estimation from inaccurate annotation. In CVPR 2011,
1465‑1472 (IEEE, 2011).

[23] Long, M., Cao, Y., Wang, J. & Jordan, M. Learning transferable features with deep adaptation networks. In International conference on machine learning, 97‑105 (PMLR, 2015).

[24] Zhang, Y., Liu, T., Long, M. & Jordan, M. Bridging theory and algorithm for domain adaptation. In International conference on machine learning, 7404‑7413 (PMLR, 2019).

[25] Jiang, J. et al. Regressive domain adaptation for unsupervised keypoint detection. In Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition, 6780‑6789 (2021).

[26] Mu, J., Qiu, W., Hager, G. D. & Yuille, A. L. Learning from synthetic animals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12386‑12395 (2020).

[27] Kim, D., Wang, K., Saenko, K., Betke, M. & Sclaroff, S. A unified framework for domain adaptive pose estimation. In
European Conference on Computer Vision, 603‑620 (Springer, 2022).

[28] Zhou, X., Koltun, V. & Krähenbühl, P. Tracking objects as points. In European conference on computer vision, 474‑490(Springer, 2020).

[29] Zhang, Y., Wang, C., Wang, X., Zeng, W. & Liu, W. Fairmot: On the fairness of detection and re-identification in multiple object tracking. Int. journal computer vision 129, 3069‑3087 (2021).

[30] Pang, J. et al. Quasi-dense similarity learning for multiple object tracking. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 164‑173 (2021).

[31] Sun, P. et al. Transtrack: Multiple object tracking with transformer. arXiv preprint arXiv:2012.15460 (2020).

[32] Zhou, X., Yin, T., Koltun, V. & Krähenbühl, P. Global tracking transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8771‑8780 (2022)