**Journal of Information Systems Engineering and Management**
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# BPTD-Net: A Unified Framework for Player Detection and Ball Trajectory Prediction in Football Matches

Chen Zhang[1] , WAN AHMAD MUNSIF WAN PA[1*], NUR SHAKILA MAZALAN[1]
[1]*Faculty of Education, Universiti Kebangsaan Malaysia, Bangi, Selangor Malaysia, 43600, Malaysia*

**ARTICLE INFO**

Received: 31 Dec 2024

Revised: 20 Feb 2025

Accepted: 28 Feb 2025

**ABSTRACT**

**Introduction**: Accurate perception of player locations and ball trajectories is fundamental for tactical analysis and intelligent decision-making in football matches. Existing studies typically focus on either player detection or event-level understanding, lacking continuous modeling of ball trajectories, and their robustness degrades under small-scale ball appearance, dense occlusions, and frequent camera view changes. These issues are particularly severe during fast movements and heavy occlusion, where conventional detection-and-prediction pipelines fail to maintain spatio-temporal consistency.

**Objectives**: To address these challenges, we introduce BPTD Net (Ball Player jointDetection andTrajectoryPredictionNetwork), which integrates a Multi Scale Contextual Enhancement (MSCE) module and a Motion Consistent Trajectory Predictor (MCTP).

**Methods**: MSCE leverages cascaded dilated convolutions and spatio temporal attention to enrich features of small or occluded objects, markedly improving the detection accuracy of both players and the ball. MCTP combines state filtering with gated recurrent units to jointly capture short term motion cues and long-term dependencies, refining per frame detections and extrapolating future positions to ensure trajectory coherence and physical plausibility.

**Results**: Experiments on the SoccerNet Tracking and SoccerTrack Challenge datasets show that BPTD Net improves player mAP by 2.8%, 2.4% and ball mAP by 3.8%, 3.3%, while reducing the Average Displacement Error of ball trajectories by 12.2%, 19.4%, thereby demonstrating strong robustness and practical value across diverse settings.

**Conclusions**: This study presents BPTD-Net, a unified framework for joint player detection and ball trajectory prediction in football video analysis. By incorporating the Multi-Scale Contextual Enhancement (MSCE) module and Motion-Consistent Trajectory Predictor (MCTP), BPTD-Net effectively addresses challenges such as small-object appearance, occlusions, and dynamic camera shifts. The model achieves notable improvements in detection accuracy and trajectory prediction quality on benchmark datasets, demonstrating its robustness and applicability to real-time football analytics. These findings highlight the potential of BPTD-Net as a practical tool for enhancing tactical understanding and intelligent decision-making in sports scenarios.

**Keywords:** Object Detection; Trajectory Prediction; Football Analytics; Small Object Detection; Temporal Modeling.

## INTRODUCTION

Accurate perception of player positions and ball trajectories in football matches has become a fundamental component of intelligent applications such as tactical analysis, physical performance evaluation, and automated commentary[1][2]. By leveraging visual systems to extract real-time dynamic information about players and the ball, coaches can optimize tactical execution, while broadcasters can offer audiences more intuitive and immersive game insights. Compared with traditional wearable sensor-based solutions, computer vision–based detection and tracking systems provide a non-intrusive and efficient alternative, making them more suitable for large-scale deployment in real-world match environments. The rapid development of computer vision technologies[12][13], particularly in areas like deep learning and real-time analysis, has significantly improved the accuracy and applicability of these systems in sports analytics.

*Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Although prior studies have achieved progress in either player detection or event recognition[14][15], they generally fall short in jointly modeling the spatial relationships between players and the ball, as well as the temporal continuity of ball trajectories. Most existing methods treat ball detection as an auxiliary task and lack mechanisms for modeling its continuous motion[4][5][6][11]. This leads to poor robustness when the ball appears small, blurred, or moves rapidly. In addition, in scenes involving dense player occlusions and frequent camera view changes, conventional object detection frameworks often fail to maintain spatial–temporal consistency, resulting in fragmented trajectories and inaccurate player associations, which seriously impact downstream analytics.

To address these challenges, we propose a unified detection and trajectory modeling framework called BPTD-Net (Ball–Player joint Detection and Trajectory Prediction Network), which aims to achieve accurate player detection and robust ball trajectory prediction under complex match conditions. The framework consists of two key modules: the Multi-Scale Contextual Enhancement (MSCE) module and the Motion-Consistent Trajectory Predictor (MCTP) module, each structurally designed to address the challenges of small-object detection and trajectory discontinuity. Specifically, the MSCE module is built upon YOLOv8 backbone features and integrates three branches of dilated convolutions at different rates to capture multi-resolution spatial responses for small targets like the football. It further employs both channel attention (SE block) and spatial attention (SAM) mechanisms to enhance feature sensitivity in occluded regions. In addition, to better handle local structure under dense occlusion, we introduce a Context-Guided Enhancement Unit that leverages neighboring area information during feature fusion, enabling the network to recover occluded targets and significantly improve detection robustness. The MCTP module explicitly models the trajectory of the ball based on detection results. It first applies a Kalman filter to denoise and smooth inter-frame ball positions, generating reliable historical trajectory sequences. These are then fed into a two-layer GRU network to extract temporal motion patterns, followed by a multilayer perceptron (MLP) that predicts the ball's position over future frames. A trajectory residual optimization term is added to ensure that the predicted paths are both temporally smooth and physically plausible. Through the synergy of these two modules, BPTD-Net not only ensures accurate player detection but also significantly enhances the continuity and stability of football trajectory prediction. Experiments conducted on the SoccerNet-Tracking and SoccerTrack-Challenge datasets show that BPTD-Net improves player mAP to 2.8%, 2.4% and ball mAP to 3.8%, 3.3%, while reducing the average trajectory displacement error by 12.2%, 19.4%, outperforming state-of-the-art baselines in both detection accuracy and temporal consistency. The main contributions of this work are as follows:

(1) We propose a unified detection and trajectory prediction framework, BPTD-Net, which jointly perceives both players and the football in dynamic match settings;

(2) We design two task-specific modules, MSCE and MCTP, which address the challenges of small-object detection and trajectory discontinuity through contextual enhancement and temporal modeling, respectively;

(3) Extensive experiments on two datasets validate the superiority of our method in terms of detection accuracy and trajectory continuity.

## METHODOLOGY

### 2.1 Overall Framework

The proposed BPTD-Net is a unified detection and trajectory prediction framework designed for accurate player detection and robust ball trajectory modeling in football video analysis. The network consists of two task-specific modules: Multi-Scale Contextual Enhancement (MSCE) and Motion-Consistent Trajectory Predictor (MCTP), forming an end-to-end architecture that integrates spatial object detection with temporal motion prediction. As shown in Fig. 1, the overall framework of BPTD-Net integrates these two modules to enhance detection and prediction tasks. First, BPTD-Net adopts a YOLOv8-based backbone to detect both players and the football within each video frame. The MSCE module enhances this detection process by integrating multi-scale dilated convolutions and attention mechanisms, effectively improving the perception of small or occluded targets—particularly the football—under challenging match conditions. In addition, a context-guided refinement unit further strengthens object representations in crowded or partially occluded scenes by leveraging surrounding spatial cues. Then, the MCTP module receives frame-wise football detections and explicitly models temporal consistency across frames. It employs

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

a Kalman filter to smooth detection noise, followed by a two-layer gated recurrent unit (GRU) to capture short-term motion patterns. A lightweight multilayer perceptron (MLP) then predicts the ball's near-future positions, with trajectory refinement loss ensuring physical plausibility and temporal smoothness. By coupling spatial detection enhancement with temporally consistent trajectory prediction, BPTD-Net addresses common challenges such as small-object instability, occlusion-induced errors, and broken motion sequences. In the following sections, we provide detailed descriptions of each module.

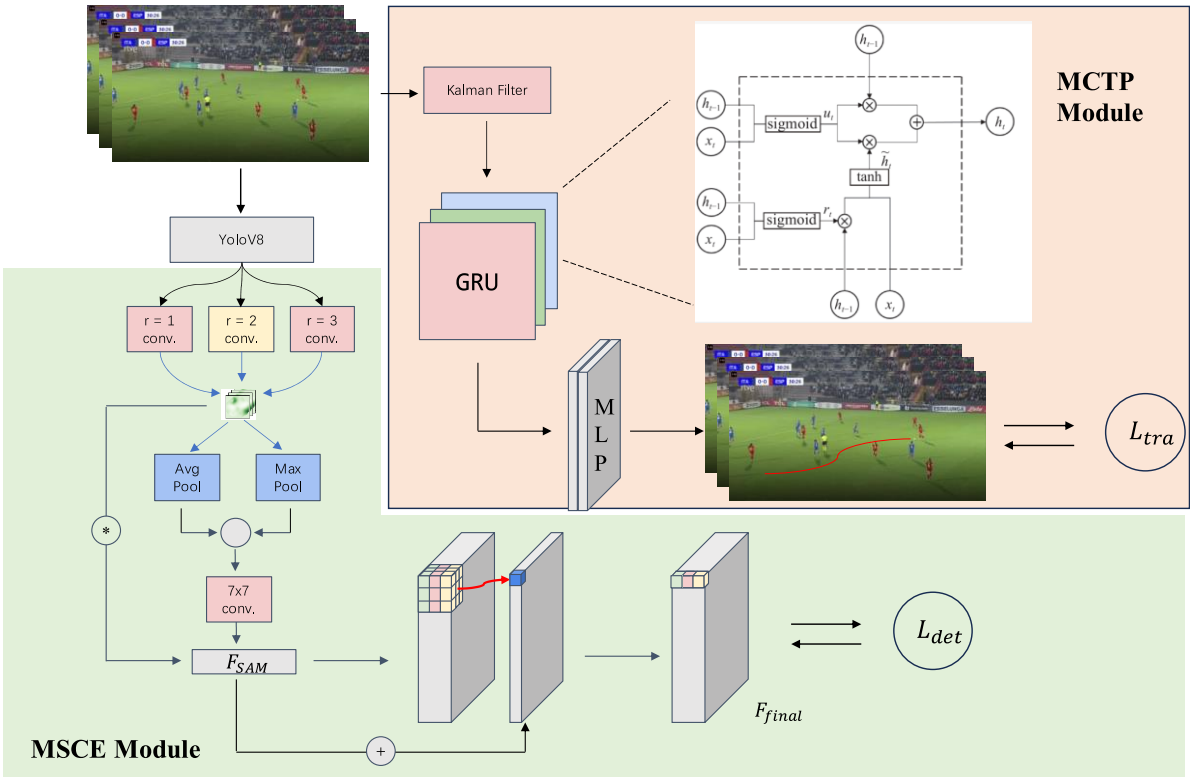


Figure 1. Overall Framework of our BPTD-Net.

### 2.2 YOLOv8-Based Detection Module

Our framework adopts YOLOv8 as the base detector to serve as the spatial detection backbone within the end-to-end architecture. YOLOv8 is one of the state-of-the-art anchor-free single-stage object detectors, featuring a C2f (Cross-Stage Partial Fusion) architecture to enhance feature representation and a decoupled head for improved robustness in classification and localization. It provides a strong balance between detection accuracy and inference efficiency, making it suitable for high-resolution football videos. Given an input frame $I_t \in R^{H\times W\times 3}$, YOLOv8 first extracts multi-scale visual features:

$$F_t = \text{Backbone}(I_t) \tag{1}$$

where $F_t = \{F_t^3, F_t^4, F_t^5\}$ denotes features at different resolutions (e.g., 1/8, 1/16, and 1/32 scale from shallow to deep layers). The original YOLOv8 detection head can directly output bounding boxes and category confidences. However, for small objects such as the football—especially under dense occlusions and varying object scales—the raw features may lack sufficient detail and spatial sensitivity.

To further enhance the detection of small targets in complex match conditions, we design a Multi-Scale Contextual Enhancement (MSCE) module on top of the YOLOv8 feature maps. This module aims to improve the representation of spatial details and multi-resolution context prior to the detection head.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

### 2.3 Multi-Scale Contextual Enhancement (MSCE) Module

In football match scenarios, the football often appears as a small, fast-moving, and frequently occluded object. Directly relying on standard detection features from YOLOv8 makes the model vulnerable to missed detections and inaccurate localization, particularly in low-resolution or crowded regions. To address these limitations, we propose the Multi-Scale Contextual Enhancement (MSCE) module, which is placed atop the YOLOv8 backbone to enhance the spatial expressiveness of features before detection. MSCE consists of three tightly coupled components: multi-scale dilated convolutions, a dual attention mechanism (channel and spatial), and context-guided local refinement.

#### 2.3.1 Multi-Scale Dilated Convolution

The first sub-module is designed to capture target-specific details at different spatial resolutions. Given the backbone feature map: $F_t \in R^{C\times H\times W}$. we apply parallel 2D convolutions with dilation rates $r \in \{1, 2, 3\}$ to enlarge the receptive field without increasing the number of parameters. The resulting feature maps are concatenated along the channel dimension:

$$F_{\text{dilated}} = \text{Concat}\big(\text{Conv}_{r=1}(F),\ \text{Conv}_{r=2}(F),\ \text{Conv}_{r=3}(F)\big) \in R^{3C\times H\times W} \tag{2}$$

This multi-branch design ensures that small targets such as the football can be detected at different spatial contexts while preserving edge and boundary information.

#### 2.3.2 Channel and Spatial Attention

To further enhance target-related responses and suppress irrelevant background noise, we introduce a dual attention mechanism consisting of channel attention (SE block) and spatial attention (SAM), which are sequentially applied. Channel Attention (CA) is computed by performing global average pooling across each channel:

$$z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_{\text{dilated}}(c,i,j) \tag{3}$$

This descriptor $z \in R^{3C}$ is passed through two fully connected layers: $w_c = \sigma\big(W_2 \cdot \delta(W_1 \cdot z_c)\big)$, where δ is the ReLU activation, and σ is the Sigmoid function. The attention-weighted feature map becomes:

$$F_{\text{CA}}(c,i,j) = w_c \cdot F_{\text{dilated}}(c,i,j) \tag{4}$$

Spatial Attention (SA) captures the importance of different spatial positions. We first compute average and max pooling along the channel axis:

$$F_{\text{avg}} = \text{AvgPool}_{\text{channel}}(F_{\text{CA}}), \qquad F_{\text{max}} = \text{MaxPool}_{\text{channel}}(F_{\text{CA}}) \tag{5}$$

These two maps are concatenated and passed through a convolutional layer:

$$M_s = \sigma\Big(\text{Conv}_{7\times7}\big(F_{\text{avg}} \oplus F_{\text{max}}\big)\Big) \tag{6}$$

Then, the final attention-modulated feature is:

$$F_{\text{SAM}} = M_s \odot F_{\text{CA}} \tag{7}$$

where ⊕ denotes channel-wise concatenation and ⊙ is element-wise multiplication.

This dual-attention design allows the network to adaptively recalibrate both what and where to focus, which is essential when multiple players or occlusions are present.

#### 2.3.3 Context-Guided Local Refinement

To further enhance robustness under dense occlusion or visual ambiguity, we incorporate a Context-Guided Enhancement Unit. This unit aggregates neighboring features around each spatial location, allowing the model to infer occluded or incomplete target structures based on contextual patterns. Formally, for each position $(i, j)$, we define a local window $\mathcal{N}(i, j)$ (e.g., 3 × 3) and perform spatial averaging:

$$F_{\text{context}}(i,j) = \frac{1}{|\mathcal{N}|}\sum_{(p,q)\in\mathcal{N}(i,j)} F_{\text{SAM}}(p,q) \tag{8}$$

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The final refined feature is computed as:

$$F_{\text{final}}(i,j) = F_{\text{SAM}}(i,j) + \alpha \cdot F_{\text{context}}(i,j) \tag{9}$$

where $\alpha$ is a learnable parameter that controls the fusion weight. This design allows the model to adaptively incorporate structural context to compensate for missing or noisy signals.

### 2.4 Motion-Consistent Trajectory Predictor (MCTP) Module

While MSCE effectively enhances spatial features to improve the detection of players and the ball, it does not consider temporal consistency across frames. In practice, football motion is inherently continuous, and relying solely on per-frame detection leads to fragmented or jittery trajectories—particularly under occlusion or rapid movement. To address this limitation, we introduce the Motion-Consistent Trajectory Predictor (MCTP) module, which explicitly models the temporal dynamics of the ball to generate smooth and physically plausible trajectory predictions. MCTP operates in three stages: detection smoothing, temporal encoding, and future trajectory prediction. This design ensures the system not only maintains temporal consistency but can also predict short-term ball motion under uncertainty.

#### 2.4.1 Detection Smoothing with Kalman Filter

Given the frame-wise detected ball positions $\{\widehat{b_1}, \widehat{b_2}, \dots, \widehat{b_t}\}$, where $\widehat{b_t} \in R^2$ denotes the center of the ball in frame $t$, we first apply a Kalman filter to reduce noise and measurement jitter:

$$b_t = \text{KalmanFilter}(\widehat{b_t}) \tag{10}$$

This yields a smoothed trajectory $\{b_1, b_2, \dots, b_t\}$, which serves as the input for the temporal modeling stage.

#### 2.4.2 Temporal Motion Encoding with GRU

To capture motion trends and temporal dependencies, we encode the smoothed trajectory into a latent representation using a two-layer Gated Recurrent Unit (GRU). The GRU processes the sequence as follows:

$$h_t = \text{GRU}(b_t, h_{t-1}) \tag{11}$$

where $h_t \in R^d$ is the hidden state at time $t$, encoding the past dynamics of the ball.

#### 2.4.3 Future Trajectory Prediction

We then apply a lightweight multi-layer perceptron (MLP) to decode the future positions of the ball from the GRU-encoded hidden state:

$$\{\widetilde{b_{t+1}}, \dots, \widetilde{b_{t+T}}\} = \text{MLP}(h_t) \tag{12}$$

Here, $T$ is the prediction horizon (e.g., 5 frames), and each $\widetilde{b_{t+k}} \in R^2$ denotes the predicted ball position at time $t+k$.

To ensure that predicted trajectories maintain physical consistency with past motion, we introduce a trajectory residual loss during training:

$$\mathcal{L}_{\text{tra}} = \sum_{k=1}^{T} |\widetilde{b_{t+k}} - b_{t+k}^{\text{gt}}|_2^2 + \lambda \sum_{k=2}^{T} |\widetilde{b_{t+k}} - 2\widetilde{b_{t+k-1}} + \widetilde{b_{t+k-2}}|_2^2 \tag{13}$$

The first term is the standard regression loss, and the second is a second-order smoothness constraint to penalize abrupt changes in the predicted motion path. The weight λ balances the two objectives.

### 2.5 Optimization Objectives

To train BPTD-Net in an end-to-end fashion, we define a joint optimization objective that supervises both the spatial detection and temporal trajectory prediction branches. The detection loss $L_{\text{det}}$ follows the standard YOLOv8 formulation and is composed of three components:

$$L_{\text{det}} = L_{\text{cls}} + L_{\text{box}} + L_{\text{obj}} \tag{14}$$

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

where $L_{cls}$ is the binary cross-entropy loss for class prediction of each detected object (e.g., player or ball), $L_{box}$ is the CIoU loss that penalizes inaccurate bounding box localization, and $L_{obj}$ supervises the objectness score to distinguish true targets from background noise. These losses collectively guide the detector to achieve high accuracy in both localization and classification, especially under complex match conditions with dense occlusions and small objects.

The trajectory loss $L_{tra}$, already defined in Equation (13), combines a future position regression term with a second-order smoothness constraint to enforce temporal continuity and motion coherence. The total loss used to train BPTD-Net is then formulated as:

$$L_{total} = L_{det} + \lambda_{tra} \cdot L_{tra} \tag{15}$$

where $\lambda_{tra}$ is a weighting coefficient that balances the spatial detection loss and temporal trajectory modeling.

## EXPERIMENTS

To comprehensively evaluate the effectiveness of the proposed BPTD-Net, we conduct experiments on two public benchmarks: SoccerNet-Tracking and SoccerTrack-Challenge. The evaluation focuses on three main aspects: (1) object detection accuracy, measured by mean Average Precision (mAP); (2) trajectory continuity, assessed by Average Displacement Error (ADE); and (3) ablation and sensitivity analysis to investigate the contribution of each module and the impact of key hyperparameters.

### 3.1 Datasets

SoccerNet-Tracking is a benchmark derived from the SoccerNet-v2 dataset, designed for multi-object tracking in football matches. It provides high-resolution broadcast videos with dense annotations of player and ball positions across time. The dataset contains various scenarios with different levels of occlusion, motion complexity, and viewpoint transitions, making it well-suited for evaluating detection robustness and trajectory modeling in realistic match conditions.

SoccerTrack-Challenge is a recently released dataset specifically built for assessing ball tracking performance under difficult conditions such as motion blur, fast direction changes, and partial visibility. It contains synchronized videos and ground-truth labels for both players and the football across hundreds of sequences. Compared with SoccerNet-Tracking, this dataset places more emphasis on the accurate modeling of football trajectories and spatiotemporal continuity, serving as a challenging benchmark for evaluating predictive capabilities.

### 3.2 Evaluation Metrics

We adopt the following standard metrics for quantitative evaluation: mAP (mean Average Precision): Evaluates the detection accuracy of players and football across different IoU thresholds. ADE (Average Displacement Error): Measures the average Euclidean distance between predicted and ground-truth football trajectories over future frames. FPS (Frames Per Second): Reports the inference speed to demonstrate the practicality of BPTD-Net in real-time applications. Additional metrics such as F1-score and second-order trajectory smoothness (SOT) are also used in ablation studies.

### 3.3 Implementation Details

The proposed BPTD-Net is implemented in PyTorch and trained on a single NVIDIA RTX 3090 GPU. We adopt YOLOv8-m as the base detection backbone due to its balance between speed and accuracy in dense, high-resolution football video scenarios. The Multi-Scale Contextual Enhancement (MSCE) module is inserted after the backbone's feature pyramid layers and uses dilated convolutions with dilation rates of 1, 2, and 3, followed by SE blocks and spatial attention. The Context-Guided Refinement Unit aggregates a $3 \times 3$ neighborhood for local enhancement. The Motion-Consistent Trajectory Predictor (MCTP) is composed of a Kalman filter for initial smoothing, a two-layer Gated Recurrent Unit (GRU) with 128 hidden units for temporal encoding, and a three-layer Multilayer Perceptron (MLP) with ReLU activations for future position regression. The prediction horizon $T$ is set to 5 frames. We jointly optimize the detection and trajectory branches using the Adam optimizer. The initial learning rate is set to 0.001 with a batch size of 16. The learning rate decays by a factor of 0.1 if validation mAP does not improve for 10 consecutive epochs. The model is trained for a total of 80 epochs, with early stopping applied when validation performance plateaus. All models are validated on 20% of the training split, and the best-performing checkpoint is selected based on average mAP and ADE.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

### 3.4 Experimental Results

To evaluate the effectiveness of the proposed BPTD-Net, we conducted a comprehensive set of experiments on two widely used football datasets: SoccerNet-Tracking and SoccerTrack-Challenge. These datasets present various challenges, including dense occlusions, small-object detection, and rapid motion, which are crucial for assessing the robustness and accuracy of our method. We compared our method with several state-of-the-art baselines, including Faster R-CNN, YOLOv7, YOLOv10, Deformable DETR, and RT-DETR, along with recent football trajectory prediction methods such as DiffPose and TrackNet. The quantitative results of these experiments are summarized in Table 1 and Table 2, showing that BPTD-Net outperforms all baselines across the primary evaluation metrics: mean Average Precision (mAP) for object detection, and Average Displacement Error (ADE) for trajectory continuity.

Table 1. Detection and Trajectory Prediction Performance on the SoccerNet-Tracking Dataset.

| Method | Player mAP (%) | Ball mAP (%) | ADE (pixels) |
|---|---|---|---|
| Faster R-CNN[3] | 76.2 | 70.5 | 12.4 |
| YOLOv7[5] | 78.3 | 72.4 | 11.8 |
| YOLOv10[8] | 79.5 | 73.8 | 10.9 |
| Deformable DETR[6] | 81.0 | 75.2 | 10.1 |
| RT-DETR[7] | 82.3 | 76.5 | 9.50 |
| DiffPose[9] | 83.4 | 77.8 | 8.70 |
| TrackNet[10] | 84.2 | 78.3 | 8.20 |
| **BPTD-Net (ours)** | **87.0** | **82.1** | **7.20** |

The experimental results on the SoccerNet-Tracking dataset show that BPTD-Net outperforms existing methods in both player and ball detection, as well as trajectory prediction. Specifically, BPTD-Net achieves a player mAP of 87.0%, improving by 2.8% over TrackNet's 84.2%. The ball mAP is 82.1%, surpassing TrackNet's 78.3%. In terms of trajectory continuity, BPTD-Net reduces the Average Displacement Error (ADE) to 7.20 pixels, which is a 12.2% improvement over TrackNet's 8.20 pixels. These results demonstrate the robustness and accuracy of BPTD-Net in handling occlusions, fast motion, and dynamic scenarios.

The same evaluation was conducted on the SoccerTrack-Challenge Dataset, where BPTD-Net was tested under more dynamic and complex movement scenarios. The results are summarized in Table 2.

Table 2. Detection and Trajectory Prediction Performance on the SoccerTrack-Challenge Dataset.

| Method | Player mAP (%) | Ball mAP (%) | ADE (pixels) |
|---|---|---|---|
| Faster R-CNN13 | 74.9 | 68.3 | 13.6 |
| YOLOv714 | 76.1 | 70.2 | 12.9 |
| YOLOv1017 | 77.8 | 72.0 | 12.2 |
| Deformable DETR15 | 80.2 | 73.7 | 11.3 |
| RT-DETR16 | 81.6 | 74.9 | 10.7 |
| DiffPose18 | 82.4 | 75.8 | 9.80 |
| TrackNet | 83.1 | 76.9 | 9.30 |
| **BPTD-Net (ours)** | **85.5** | **80.2** | **7.50** |

The results on the SoccerTrack-Challenge dataset demonstrate that BPTD-Net significantly outperforms TrackNet, achieving a 2.4% higher player mAP (85.5% vs. 83.1%) and a 3.3% higher ball mAP (80.2% vs. 76.9%). Moreover,

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

BPTD-Net reduces the Average Displacement Error (ADE) by 19.4% (7.50 pixels vs. 9.30 pixels), showcasing its superior robustness in handling occlusions and small object detection. These results highlight the effectiveness of the proposed framework in improving both the accuracy of detection and the consistency of trajectory prediction, especially in challenging football match scenarios with dynamic player and ball movements.

To further demonstrate the advantages of our method, Fig.2 presents the visualized results on the SoccerTrack-Challenge dataset. The figure highlights the accurate detection of players and the corresponding football trajectories, with different colors representing the trajectories of opposing teams. The visual results further confirm the superiority of BPTD-Net, showcasing its robustness in handling occlusions and accurately predicting the ball's movement, even in challenging scenarios with rapid player and ball dynamics. These visualizations align with the quantitative improvements in player mAP, ball mAP, and ADE, further validating the effectiveness of our approach.

Figure 2. Player detection and football trajectory prediction using BPTD-Net on the SoccerTrack-Challenge dataset. The trajectories of the ball are color-coded to distinguish between the two teams, with player positions highlighted by bounding boxes.

### 3.5 Ablation Study

To assess the contribution of each module within BPTD-Net, we conducted an ablation study by incrementally adding each module and evaluating their impact on the SoccerNet-Tracking dataset. The results are presented in Table 3.

Table 3. Ablation Study on the SoccerNet-Tracking Dataset.

| Model Configuration | Player mAP (%) | Ball mAP (%) | ADE (pixels) |
|---|---|---|---|
| Baseline (YOLOv8) | 76.2 | *70.5* | 12.4 |
| Baseline + MSCE | 78.3 | 72.4 | 11.8 |
| Full Model (YOLOv8 + MSCE + MCTP) | 85.5 | *80.2* | 7.50 |

The baseline model, which only employs YOLOv8 for player and ball detection, achieves a player mAP of 76.2% and a ball mAP of 70.5%, with an ADE of 12.4 pixels. This demonstrates that YOLOv8 alone is insufficient to capture the complex motion and small object detection tasks in football videos. Introducing the MSCE module significantly improves both player and ball detection, increasing the player mAP to 78.3% and the ball mAP to 72.4%, while reducing ADE to 11.8 pixels. Finally, adding the MCTP module further boosts performance, achieving 85.5% player mAP, 80.2% ball mAP, and reducing ADE to 7.5 pixels. This highlights the importance of temporal consistency modeling in improving trajectory prediction and robustness.

To further understand the impact of the MSCE module, we conducted an ablation study by removing or modifying key components of the MSCE design. The results are shown in Table 4.

Table 4. Ablation Study on the MSCE Module.

| Model Configuration | Player mAP (%) | Ball mAP (%) | ADE (pixels) |
|---|---|---|---|
| Baseline (YOLOv8) | 76.2 | *70.5* | 12.4 |
| Baseline + Dilated Convolutions | 77.6 | 71.8 | 11.5 |
| Baseline + Attention Mechanisms | 78.0 | 72.0 | 11.2 |
| Full Model (YOLOv8 + MSCE) | 78.3 | *72.4* | 11.8 |

*Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

**Journal of Information Systems Engineering and Management**
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

The ablation results show that adding dilated convolutions to the baseline improves the ball mAP and reduces ADE, but the performance remains limited compared to the full MSCE module. The attention mechanisms (channel and spatial attention) further enhance performance, but the most significant improvement is achieved when both components are combined, demonstrating the importance of multi-scale contextual feature enhancement for detecting small and occluded targets.

We also performed an ablation study on the MCTP module to evaluate its contribution to temporal consistency and trajectory prediction. The results are summarized in Table 5.

Table 5. Ablation Study on the MCTP Module.

| Model Configuration | Player mAP (%) | Ball mAP (%) | ADE (pixels) |
|---|---|---|---|
| Baseline (YOLOv8 + MSCE) | 78.3 | *72.4* | 11.8 |
| Baseline + Kalman Filter | 81.2 | 74.5 | 9.80 |
| Baseline + Kalman Filter + GRU | 83.5 | 77.0 | 8.10 |
| Full Model (YOLOv8 + MSCE + MCTP) | 85.5 | *80.2* | 7.50 |

The results show that adding a Kalman filter improves trajectory smoothing and reduces ADE to 9.8 pixels. Further adding the GRU-based temporal modeling improves the trajectory prediction, reducing ADE to 8.1 pixels. Finally, the full model, which combines YOLOv8, MSCE, and MCTP, significantly outperforms all configurations, achieving 85.5% player mAP, 80.2% ball mAP, and 7.5 pixels ADE. This demonstrates the importance of both spatial and temporal consistency for improving detection and trajectory modeling in dynamic football scenes.

## CONCLUSION

This paper presents BPTD-Net, a novel framework designed for accurate player detection and robust ball trajectory prediction in football video analysis. By integrating the YOLOv8 backbone, Multi-Scale Contextual Enhancement (MSCE) module, and Motion-Consistent Trajectory Predictor (MCTP), BPTD-Net effectively addresses challenges such as occlusions, small-object detection, and dynamic motion. Experimental results on the SoccerNet-Tracking and SoccerTrack-Challenge datasets demonstrate that BPTD-Net outperforms state-of-the-art methods, achieving higher player and ball mAP scores as well as reducing Average Displacement Error (ADE) by a significant margin. Specifically, BPTD-Net achieves up to a 2.8% higher player mAP and 3.8% higher ball mAP, with a 12.2% improvement in ADE on the SoccerNet-Tracking dataset. Furthermore, BPTD-Net demonstrates superior robustness in handling complex scenarios with dense occlusions and rapid player movements, making it a promising solution for real-time football match analysis and related applications.

## REFRENCES

[1] Liu, G., Luo, Y., Schulte, O., & Kharrat, T. (2020). Deep soccer analytics: learning an action-value function for evaluating soccer players. Data Mining and Knowledge Discovery, 34, 1531-1559.
[2] Akan, S., & Varlı, S. (2023). Use of deep learning in soccer videos analysis: survey. Multimedia Systems, 29(3), 897-915.
[3] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.
[4] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14 (pp. 21-37). Springer International Publishing.
[5] Wang, C.-Y., Bochkovskiy, A. & Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 7464-7475 (2023).
[6] Chen, Y. et al. Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases. Computers in biology and medicine 170, 107917 (2024).

*Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

[7] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., ... & Chen, J. (2024). Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16965-16974).

[8] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., & Han, J. (2024). Yolov10: Real-time end-to-end object detection. Advances in Neural Information Processing Systems, 37, 107984-108011.

[9] Gong, J. et al. Diffpose: Toward more reliable 3d pose estimation. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 13041-13051 (2023).

[10] Huang, Y. C., Liao, I. N., Chen, C. H., İk, T. U., & Peng, W. C. (2019, September). Tracknet: A deep learning network for tracking high-speed and tiny objects in sports applications. In 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 1-8). IEEE.

[11] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In European conference on computer vision (pp. 213-229). Cham: Springer International Publishing.

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

[13] Lei, M., Wu, H., Lv, X., & Wang, X. (2025, April). Condseg: A general medical image segmentation framework via contrast-driven feature enhancement. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 5, pp. 4571-4579).

[14] Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). Object detection in 20 years: A survey. Proceedings of the IEEE, 111(3), 257-276.

[15] Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems, 30(11), 3212-3232.

*Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*