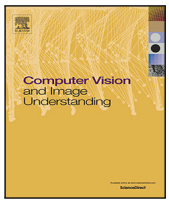
Computer Vision and Image Understanding 264 (2026) 104657

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



TOTNet: Occlusion-aware temporal tracking for robust ball detection in sports videos☆

Hao Xu [a,*], Arbind Agrahari Baniya [a], Sam Wells [b], Mohamed Reda Bouadjenek [a], Richard Dazeley [a], Sunil Aryal [a]

[a] School of Information Technology, Deakin University, Melbourne, Australia
[b] Paralympics Australia, Melbourne, Australia

ARTICLE INFO

Keywords:
Sports
Object tracking
Sports video analysis
Temporal context modeling

ABSTRACT

Ball tracking is a fundamental problem in computer vision, particularly in sports analytics, where it underpins tasks such as analyzing ball movement in soccer and basketball or detecting bounce locations in tennis and table tennis. Most existing methods are developed and evaluated on resource-rich, commercial sports footage with ideal camera angles, high-resolution imagery, and multiple viewpoints. In contrast, many other sports contexts, including semi-professional leagues, local amateur competitions, and Paralympic sports, lack these resources. Footage in these settings often comes from single, fixed, and suboptimal viewpoints, where occlusion becomes a dominant challenge for automated tracking. Existing methods frequently fall short in such conditions because their architectures and training strategies do not explicitly account for prolonged or full occlusion. To address this gap, we present the **Table Tennis Australia (TTA) dataset**, the first professionally annotated Paralympic table tennis benchmark with dense visibility labels, captured under realistic single-view conditions. With **2,396** occluded instances (including 998 fully occluded), TTA is the most occlusion-rich publicly available dataset to date. Alongside the dataset, we propose the **Temporal Occlusion Tracking Network (TOTNet)**, a novel tracking system designed to maintain localization accuracy even under extended occlusion. Through comprehensive experiments on four sports tracking datasets, TOTNet achieves state-of-the-art performance, with substantial gains in full-occlusion scenarios. We release the dataset, code, and evaluation scripts to foster reproducibility and future research in occlusion robust tracking for low resource sports; all materials are available at https://github.com/AugustRushG/TOTNet.

1. Introduction

Automated ball tracking is a core capability in sports analytics, enabling downstream tasks such as possession analysis, trajectory prediction, and event detection (Naik et al., 2022; Kamble et al., 2019a). While existing research has achieved strong performance on broadcast-quality, multi-camera footage in sports like tennis, basketball, and soccer, most methods and benchmarks assume ideal capture conditions: high-resolution video, optimal viewing angles, and minimal occlusion. In contrast, many real world settings, including Paralympic competition, semi-professional tournaments, and amateur leagues, operate under single-view, fixed-angle capture with frequent, prolonged occlusions. These conditions severely limit the effectiveness of existing approaches and reduce the reliability of analytics in contexts where they could have the greatest practical impact.

We address this gap by introducing the task of *visibility-aware occlusion-robust ball tracking* for racket sports. This task explicitly measures tracking robustness across varying visibility levels and prolonged occlusions, a setting largely ignored in current literature. To support research in this domain, we present **TTA** (Table Tennis Australia), the first professionally annotated Paralympic table tennis dataset with dense frame level visibility labels. TTA contains 12,414 samples, including 2396 occlusion cases (998 fully occluded). Over 19% of frames are captured under realistic single-view conditions This high occlusion density makes TTA uniquely suited for benchmarking robustness, where other datasets provide little to no occlusion coverage (Huang et al., 2019; Sun et al., 2020; Tarashima et al., 2023). Examples of the dataset are shown in Fig. 1. TTA reflects the true constraints of low-resource

☆ This article is part of a Special issue entitled: 'CV for Sports' published in Computer Vision and Image Understanding.

* Corresponding author.
E-mail addresses: august.xu@research.deakin.edu.au (H. Xu), a.agraharibaniya@deakin.edu.au (A.A. Baniya), sam.wells@paralympic.org.au (S. Wells), reda.bouadjenek@deakin.edu.au (M.R. Bouadjenek), richard.dazeley@deakin.edu.au (R. Dazeley), sunil.aryal@deakin.edu.au (S. Aryal).

https://doi.org/10.1016/j.cviu.2026.104657
Received 18 September 2025; Received in revised form 23 December 2025; Accepted 8 January 2026
Available online 12 January 2026

Fig. 1. TTA dataset examples including different tournaments such as Paralympics Pairs, World Para Elite, World Para Future.

sports analytics and is intended as a **benchmark for visibility-aware evaluation** in practical settings.

Previous ball tracking approaches struggle under these conditions. Frame-based object detectors such as YOLO (Redmon, 2016), Faster R-CNN (Ren et al., 2016), and SSD (Liu et al., 2016) rely heavily on spatial cues from a single frame, making them brittle when the ball is invisible due to occlusion (Saleh et al., 2021). Temporal CNN approaches (Liu and Wang, 2022; Huang et al., 2019; Sun et al., 2020; Tarashima et al., 2023; Voeikov et al., 2020; Raj et al., 2025; Chen and Wang, 2023) improve robustness by stacking frames, but this treats temporal information as static features, losing dynamic motion patterns critical for predicting the ball's location when occluded. Kalman filter-based methods (Naik and Hashmi, 2023; Hu et al., 2024; Li et al., 2023) offer motion prediction but are limited to linear dynamics and cannot capture the complex trajectories caused by spin, sudden deflections, or rapid velocity changes common in racket sports.

Alongside the dataset, we define a **visibility-specific evaluation protocol** that reports performance across visibility tiers, enabling systematic benchmarking of occlusion robustness. We further provide **TOTNet** (Temporal Occlusion Tracking Network) as a strong reference baseline for this benchmark. TOTNet integrates motion modeling with occlusion targeted data augmentation and is explicitly designed for *offline* sports analytics applications such as post-match analysis, referee support, and tactical review, where inference speed is secondary to accuracy and robustness.

Contributions. The main contributions of this work are:

1. **A new task: visibility-aware occlusion-robust ball tracking for racket sports** – Formalizing a benchmark problem setting not addressed in existing literature.
2. **TTA: a new occlusion-rich benchmark dataset** – First professionally annotated Paralympic table tennis dataset with dense visibility labels, including 2396 occlusion cases (998 fully occluded) — the highest occlusion density among racket-sport benchmarks.
3. **Visibility-specific evaluation protocol** – Enabling fine-grained, tiered evaluation of robustness under different occlusion severities.
4. **TOTNet as a strong reference system** – Motion-aware temporal tracking baseline achieving SOTA performance across four racket-sport datasets, particularly under occlusion.
5. **Real-world deployment evidence** – Integrated into an elite-level Paralympic table tennis analytics workflow, reducing annotation time per match from 3–4 h to under 6–7 min.

2. Related work

2.1. Sports tracking datasets

Existing sports ball tracking datasets can be broadly grouped into those for large-field team sports (e.g., soccer, basketball) and racket sports (e.g., tennis, badminton, table tennis). In team sports, prolonged occlusions and large playing areas make ball position prediction far more ambiguous, whereas in racket sports, short-term motion cues and player context often allow occluded trajectories to be estimated reliably. Most existing racket sports datasets share common traits: broadcast-quality footage with optimal viewing angles, limited or incidental occlusions rarely annotated explicitly, and no visibility-specific evaluation protocol. Table 1 summarizes representative datasets, showing that none combine dense visibility annotations with a large number of full occlusions in realistic single-view capture. Our proposed TTA dataset fills this gap, offering the first benchmark for visibility-aware ball tracking in low-resource racket sports.

2.2. Single object tracking in sports videos

The development of deep learning-based image detectors such as YOLO (Redmon, 2016), SSD (Liu et al., 2016), and R-CNN (Girshick et al., 2014) has significantly advanced ball tracking in sports videos. These methods follow the tracking-by-detection (TBD) paradigm, where detections are obtained from individual frames and subsequently linked to form trajectories (Naik and Hashmi, 2023; Buric et al., 2018; Teimouri et al., 2019; Reno et al., 2018; Komorowski et al., 2019). However, TBD methods process frames independently, which limits their ability to leverage temporal information and results in temporally inconsistent tracking, especially during partial or full occlusions.

To overcome these limitations, recent works have explored the integration of temporal information. Methods like TrackNet (Huang et al., 2019), TrackNetV2 (Sun et al., 2020), and MonoTrack (Liu and Wang, 2022) incorporate multiple consecutive frames as inputs to CNNs, capturing short-term motion patterns. Other approaches use advanced temporal modeling techniques, such as optical flow (Dosovitskiy et al., 2015), Recurrent Neural Networks (RNNs), convolutional LSTMs (Patraucean et al., 2015), and temporal convolutions (Lea et al., 2017), to better model object motion over time (Kukleva et al., 2019; Li et al., 2023). Additionally, transformers (Vaswani, 2017) have introduced spatiotemporal attention mechanisms, enabling models to learn correlations within and across frames and predict object movements more effectively (Yu et al., 2024; Chao et al., 2024). Distinct from end-to-end deep learning paradigms, a separate category of methods relies on global optimization to ensure trajectory coherence. Maksai et al. (2016) formulated ball tracking as a Mixed Integer Program (MIP), effectively capturing long-term dependencies by jointly optimizing ball and player interactions over the entire video sequence. Similarly, Zou et al. (2024) recently employed a graph-based message-passing framework to refine candidate detections extracted via classical computer vision heuristics. While these approaches demonstrate high tracking precision by leveraging global temporal context (offline processing), they differ fundamentally from online, causal trackers which must operate in real-time without access to future frames. Despite these advances, current approaches still struggle with occlusion handling, particularly

Table 1
Comparison of representative racket-sports ball tracking datasets. TTA is the only dataset with dense occlusion labeling, including a large number of fully occluded frames, captured under realistic single-view conditions.

Dataset	Sport	Capture setup	FPS	Resolution	#Samples	#Occ. Cases	#Full Occ.	Occ. Rate (%)
TrackNet (Huang et al., 2019)	Tennis	Broadcast	30	1280 × 720	19,835	1474	82	7.43
TrackNetV2 (Sun et al., 2020)	Badminton	Broadcast	30	1280 × 720	68,763	0	0	0.00
TT (Voeikov et al., 2020)	Table tennis	Broadcast	120	1920 × 1080	52,061	0	0	0.00
TTA (Ours)	Table tennis	Handheld	25	1920 × 1080	12,414	**2396**	**998**	**19.30**

in sports where the ball frequently disappears due to rapid motions and interactions with players. This highlights the need for methods that can effectively leverage both temporal and contextual information to improve tracking consistency and its robustness to occlusion challenges.

2.3. Occluded object tracking

Occluded object tracking remains a significant challenge in video-based object detection, despite advancements in the field (Saleh et al., 2021). The difficulty lies in collecting and labeling datasets with sufficient occlusion diversity, as creating comprehensive real-world datasets for all occlusion scenarios is nearly impossible. As a result, many studies are based on synthetic datasets or automatically generated occluded samples (Saleh et al., 2021). To address this, Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have been employed to generate occluded data. For instance, Wang et al. (2017) augmented the COCO dataset with occluded objects using GANs, improving model robustness through enhanced training data. Similarly, Li and Malik (2016) created synthetic occlusions by overlaying object masks from one image onto another, producing amodal data to improve occlusion handling.

Compositional models also show promise. These models detect partially occluded objects by leveraging a generative, modular approach. For example, Kortylewski et al. (2020) used a differentiable generative compositional layer instead of the fully connected layer in a CNN, enabling robust classification of occluded objects and accurate localization of occluders. In another approach, Cui et al. (2021) framed object tracking as a Markov decision process within a deep reinforcement learning framework. Their AD-OHNet tracker utilized temporal and spatial contexts from action-state histories prior to occlusion, enabling accurate tracking even during complete occlusion. For multiperson tracking, Zhou et al. (2018) proposed a deep alignment network that combines an appearance model with a Kalman filter-based motion model. This hybrid approach effectively handles occlusions and improves motion reasoning during tracking.

Occlusion is a common challenge in sports scenarios, significantly impacting ball and player detection. For instance, Halbinger and Metzler (2015) introduced a two-stage approach for detecting soccer balls under occlusion. The first stage detects the ball when fully visible, while the second stage identifies occluded parts as "bumps" on player silhouettes using a Hough transform for circular shape detection, followed by Freeman Chain Code analysis to confirm the ball's dimensions. In Kamble et al. (2019b), the authors tackled ball occlusion by tracking the player occluding the ball, maintaining continuity in the tracking process. However, this post-processing approach is less effective in sports with frequent and complex occlusions that demand precise, real-time localization. To address occlusions, Naik and Hashmi (2023) integrated YOLOv3 with a Kalman filter to predict the ball's position during occlusion. This hybrid approach demonstrated improved robustness in challenging scenarios by combining detection and prediction. In racket sports, leveraging temporal information is critical due to frequent occlusions of fast-moving objects. For example, Huang et al. (2019) demonstrated that by stacking multiple consecutive frames as input enables CNN models to capture trajectory patterns, aiding in the detection and tracking of occluded objects. However, the mechanisms by which trajectory patterns are learned remain unexplored and the occlusion samples in the dataset is too small to show its effectiveness. Building on this, Sun et al. (2020) improved prediction accuracy by adopting a U-Net architecture combined with a multiple-input, multiple-output (MIMO) framework. Although this design introduced a slight reduction in processing speed, it significantly enhanced overall performance. Extending the TrackNet family, Chen and Wang (2023) and Raj et al. (2025) further boosted performance by incorporating motion features through frame differencing. Additional progress was achieved by Liu and Wang (2022), who incorporated residual connections within U-Net blocks to improve tracking accuracy in badminton videos.

Contrasting with encoder–decoder architectures, Tarashima et al. (2023) argued that such designs often lack sufficient feature diversity for effective tracking. They employed high-resolution feature extraction methods from HRNet (Wang et al., 2020; Yu et al., 2021), combined with position-aware model training and temporal consistency, achieving SOTA results across multiple sports datasets.

Most existing approaches leverage temporal information by stacking multiple frames along the channel dimension for 2D convolutions. However, this method limits the ability to fully capture the rich temporal dynamics inherent in video data. Stacking frames treats temporal information as static features rather than dynamic sequences, failing to explicitly model the evolution of object motion over time. As a result, these approaches struggle to track objects accurately during occlusion, where the model must rely on contextual information from preceding and succeeding frames to predict the occluded object's position.

3. Methodology

3.1. Datasets

To evaluate performance across varied racket sports, we use four datasets: three existing ones for table tennis, tennis, and badminton, and a newly introduced TTA dataset designed to emphasize occlusion scenarios. The TT dataset from Voeikov et al. (2020) includes five training and seven testing videos, yielding 36,224 training, 3232 validation, and 3720 test samples. Captured at 120 fps (1920 × 1080) from a side view, it features minimal occlusion and lacks visibility labels. The tennis dataset, accessed via (Tarashima et al., 2023) from Huang et al. (2019), contains 10 clips (30 fps, 1080 × 720) with ball coordinates and visibility labels: 0 (out-of-frame), 1 (visible), 2 (partially visible), 3 (fully occluded). We created balanced splits detailed in Table 2. The badminton dataset (Sun et al., 2020) has 26 training and 3 testing matches (30 fps, 1280 × 720) with binary visibility labels (0 = not visible, 1 = visible). It contains a higher proportion of invisible frames than other datasets.

The **TTA dataset**, manually annotated for this study, contains 12,414 samples (25 fps, 1920 × 1080) from 17 professional-level Para table tennis matches across major tournaments including the Paralympics Pairs, World Para Elite (WPE), and World Para Future (WPF). Each frame is annotated with both ball coordinates and a visibility label. The dataset includes **2396 occlusion samples** — substantially more than any existing benchmark — arising from challenging camera angles, the small ball size, and highly dynamic gameplay. All annotations were reviewed by national team analysts to ensure reliability and correctness.

Annotation Protocol: We annotate each frame using three visibility levels: *(i) fully visible*, where the ball is completely unobstructed and easily identifiable; *(ii) partially occluded*, where the ball is partially

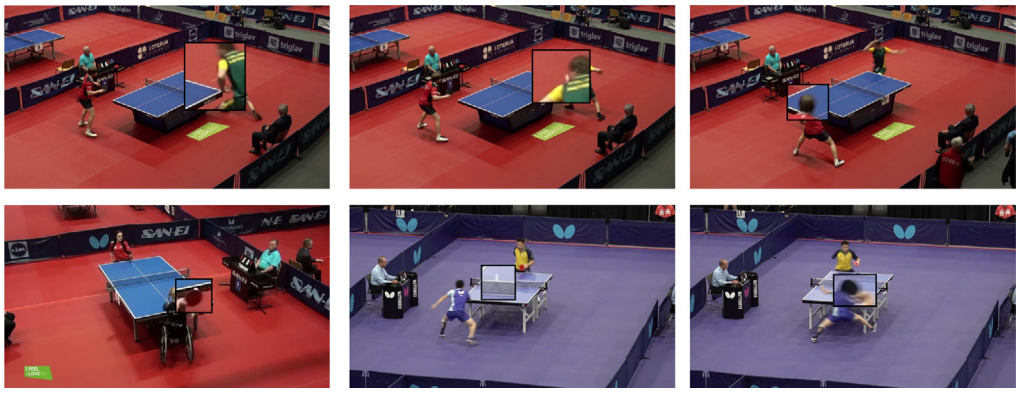


Fig. 2. Examples of partially occluded cases (Columns 1–2) and fully occluded cases (Column 3). The approximate ball location in each frame is zoomed in.

Table 2
Visibility-level distribution across Tennis, TTA, and Badminton datasets.

Dataset	Visibility level	Train	Val	Test
Tennis	Out of frame	587	107	135
	Fully visible	11,641	2940	3054
	Partially occluded	861	193	338
	Fully occluded	55	22	5
TTA (ours)	Fully visible	6787	1692	1539
	Partially occluded	976	252	168
	Fully occluded	689	177	132
Badminton	Out of frame	8052	2022	2188
	Fully visible	54,794	13,690	10,468

hidden by a player, racket, or other object but still provides local visual cues; and *(iii) fully occluded*, where the ball is completely hidden from view yet remains in play. Examples of partially and fully occluded is shown in Fig. 2.

For fully visible and partially occluded frames, annotators mark the ball center directly. For fully occluded frames, where no pixel-level evidence exists, the ball location is inferred through *trajectory-consistent interpolation*: annotators reference the ball positions in preceding and subsequent visible frames to estimate a physically plausible continuation of its motion, respecting table geometry, player contact timing, and feasible speed changes. When an occlusion extends beyond three consecutive frames, annotators additionally estimate instantaneous velocity and direction to maintain temporal coherence. All long-occlusion annotations are reviewed by a second annotator to ensure reliability.

This procedure ensures that fully occluded labels are not arbitrary but reflect the most probable continuation of the ball's true trajectory.

Design Rationale: While TTA is comparable in overall scale to existing broadcast-based datasets, it is uniquely curated for challenge density. More than 19% of its frames involve occlusions, including 998 fully occluded cases—orders of magnitude higher than previous datasets. This concentration of difficult samples minimizes redundancy, enables more efficient occlusion-specific learning, and establishes TTA as the most focused benchmark for evaluating single-view tracking under realistic, low-resource sports conditions.

3.2. Occlusion augmentation

We propose a novel augmentation technique to enhance performance in occlusion scenarios. For a sequence of frames, we simulate occlusion by masking the ball area in the target frame with a randomly sized shape filled with the surrounding mean pixel values. Examples are shown in Fig. 3. This augmentation forces the model to rely on temporal information from adjacent frames and the spatial context surrounding the occluded region rather than solely depending on the current frame's spatial features.

To prevent the model from adapting too strongly to this augmentation, we introduce additional noise by randomly selecting areas in the other frames and filling them with mean pixel values. This ensures that the model learns to generalize better and robustly utilize both temporal and spatial features across all frames.

3.3. BCE loss with visibility-based weighting

We employ a visibility-aware weighted binary cross-entropy (BCE) loss to account for the imbalance and inherent uncertainty present in occlusion-heavy ball tracking. Prior works (Huang et al., 2019; Sun et al., 2020; Tarashima et al., 2023; Voeikov et al., 2020) typically supervise ball locations using 2D Gaussian target maps for all frames. Our formulation follows this consistent strategy: *all annotated ball positions — visible, partially occluded, and fully occluded — are represented as normalized Gaussian heatmaps*. This provides smooth spatial gradients, avoids unstable one-hot targets, and leads to more stable optimization.

Our loss design therefore ensures:

- Consistent Gaussian supervision across all annotated visibility levels.
- Balanced learning through visibility-dependent weighting.

Target definition. Let $P \in \mathbb{R}^{H\times W}$ denote the predicted heatmap obtained after applying a spatial softmax over all $H \times W$ logits. For any annotated ball position (T_x, T_y), the target heatmap $T_{\text{map}} \in [0,1]^{H\times W}$ is defined as a normalized 2D Gaussian:

$$T_{\text{map}}[i,j] = \frac{1}{Z}\exp\left(-\frac{(i-T_y)^2+(j-T_x)^2}{2\sigma^2}\right), \tag{1}$$

where the normalization term is:

$$Z = \sum_{u=0}^{H-1}\sum_{v=0}^{W-1}\exp\left(-\frac{(u-T_y)^2+(v-T_x)^2}{2\sigma^2}\right). \tag{2}$$

Visibility-based weighting. Each visibility level $v \in \{0,1,2,3\}$ is assigned a scalar weight. We define the visibility-weight vector as:

$$\mathbf{w} = [w_{\text{oof}},\ w_{\text{vis}},\ w_{\text{partia}},\ w_{\text{occ}}]. \tag{3}$$

For a frame with visibility label v, the corresponding weight is the scalar:

$$w_v = \mathbf{w}[v]. \tag{4}$$

Final loss. The final loss for a frame is:

$$L = w_v \cdot \text{BCE}\big(P,\ T_{\text{map}}\big), \tag{5}$$

where $\text{BCE}(\cdot)$ returns a scalar loss between the predicted heatmap P and the Gaussian target T_{map} (if present).

Fig. 3. Augmentation examples for three different datasets where the ball is circled in red.

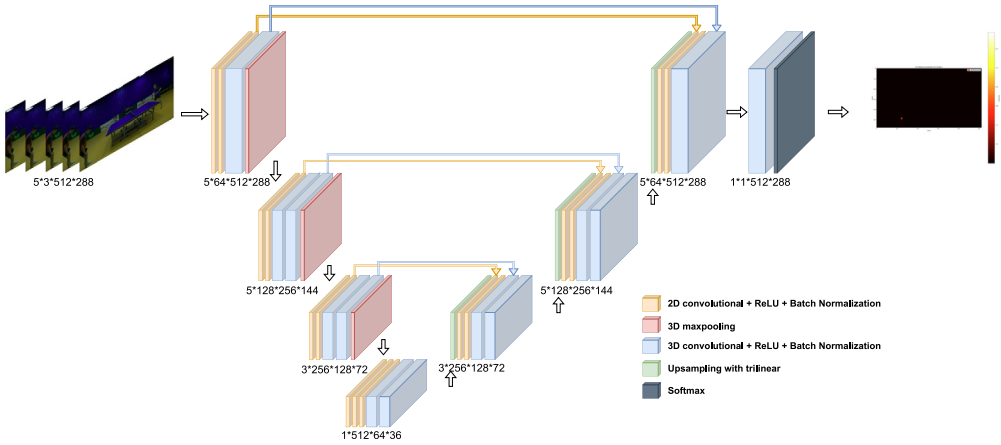


Fig. 4. Overview of the proposed **TOTNet** architecture. TOTNet preserves the temporal dimension throughout the encoder–decoder pipeline by combining 2D spatial convolutions with 3D temporal convolutions. Residual connections ensure spatial detail retention, while skip connections between encoder and decoder layers support both spatial and temporal feature reconstruction. The final output is a visibility-aware heatmap representing the predicted ball position.

3.4. Architecture

Our model extends prior work (Huang et al., 2019; Sun et al., 2020) by explicitly preserving the temporal dimension throughout the network, enabling richer motion-aware feature learning. Unlike earlier approaches that collapse temporal information by stacking frames along the channel axis, TOTNet maintains a structured temporal pathway that captures frame-to-frame dynamics.

Motion is modeled through lightweight shortcut 3D convolutional blocks integrated into both the encoder and decoder. These 3D convolutions operate over short temporal windows to extract local motion cues — such as instantaneous velocity and direction changes — which are then fused with spatial features via residual connections. This spatio-temporal fusion helps maintain trajectory continuity during rapid movement or occlusions, where visual evidence alone is insufficient.

The overall architecture builds upon a U-Net backbone (Ronneberger et al., 2015), using encoder–decoder blocks with skip connections to preserve spatial detail while propagating temporal context. An overview of the full architecture is provided in Fig. 4.

3.4.1. Encoder

Each encoder block integrates spatial and temporal convolutions, where the spatial convolution is implemented as a 2D convolution, and the temporal convolution is implemented as a 3D convolution. Initially, spatial convolutions are applied to each frame independently to extract spatial features. These features are then passed through temporal convolutions, keeping the temporal dimension intact, to capture dependencies across frames and effectively combine spatial and temporal information for better object localization. A 3D max pooling operation is then applied to reduce both the spatial size and temporal frame size while preserving the most valuable information. Additionally, a residual connection is included within each encoder block, where the output from the spatial convolution is added to the output of the temporal convolution to ensure that spatial information is not lost. A detailed flow of the encoder block is specified in Fig. 5.

As the model deepens on the encoder side, the number of channels increases while the spatial resolution and temporal sequence decrease. The kernel size for spatial convolutions becomes smaller to focus on finer details in smaller spatial regions, whereas the kernel size for temporal convolutions decreases as the temporal dimension gets reduced.

3.4.2. Bottleneck

The bottleneck block processes the highly abstracted information from the encoder. By the time the input reaches the bottleneck block, the temporal dimension has been reduced to 1, and the spatial dimensions have been significantly downsized. This block contains more spatial layers than the other blocks, focusing on extracting the most critical features. Since the temporal dimension is reduced to 1, the kernel size for temporal convolution is set to (1,1,1), effectively performing pointwise convolution. This operation facilitates mixing information across all channels, enabling the block to generate rich feature representations that combine temporal and spatial information effectively.

3.4.3. Decoder

After passing through the bottleneck block, the decoder blocks begin the 3D upsampling process to restore both the temporal and spatial

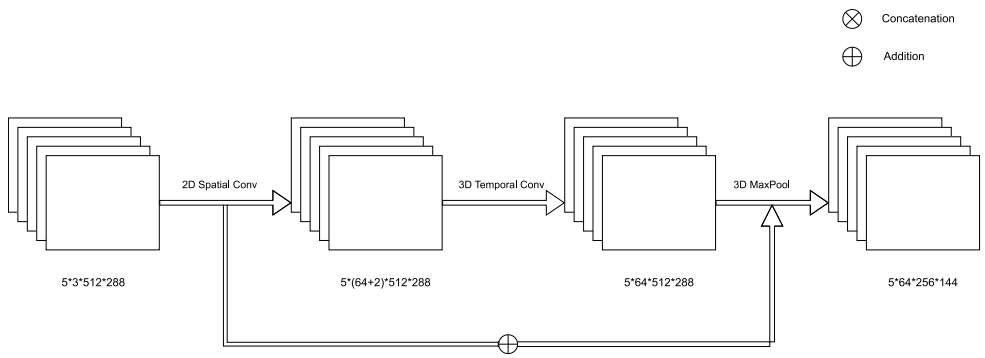


Fig. 5. Detailed structure of the first encoder block in **TOTNet**. Each frame is first processed by a 2D convolution to extract spatial features, followed by a 3D temporal convolution to capture inter-frame dependencies. A 3D max-pooling layer reduces both spatial and temporal dimensions, while a residual connection from the spatial path to the temporal path ensures preservation of fine-grained spatial details.

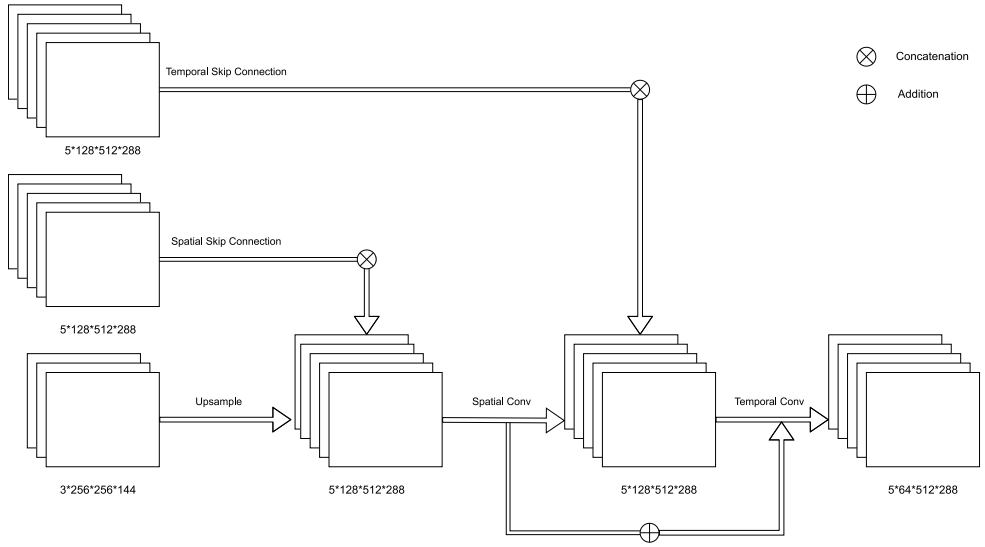


Fig. 6. Structure of the last decoder block in **TOTNet**. Features are upsampled using trilinear interpolation, followed by 2D spatial and 3D temporal convolutions. Skip connections from earlier encoder layers provide both spatial and temporal context, enabling accurate reconstruction. A final temporal convolution reduces the channel dimension, and a softmax is applied to produce the visibility-aware probability heatmap of the ball location.

dimensions. Features are upsampled using the trilinear method, as the model focuses on generating a heatmap for the object's likely position rather than precise segmentation. Each decoder block mirrors the layers of the encoder block, and skip connections are utilized for both spatial and temporal convolutions. These skip connections concatenate the encoder features along the channel dimension, allowing the decoder to leverage information from earlier layers for enhanced feature reconstruction both spatially and temporally. A detailed decoder structure is in Fig. 6. In the final stage, a temporal convolutional block is employed to reduce the channel dimension to one, retaining the most critical information. A softmax activation is applied to the resulting heatmap to ensure it represents a normalized probability distribution.

4. Experiments

4.1. Implementation details

For all datasets, video frames and corresponding ball coordinates are resized to 288 × 512 pixels. The number of input frames is treated as a hyperparameter, with five frames selected as the optimal balance between temporal context and computational efficiency. Alongside occlusion augmentation, we apply standard data augmentations, including color jittering, random cropping/resizing, and horizontal or vertical flipping. Models are trained using the AdamW optimizer (Loshchilov, 2017) with a learning rate of 5×10^{-4}, weight decay of 5×10^{-5}, a batch size of 8, and a cosine learning-rate scheduler. Training is performed for 30 epochs, with the best checkpoint selected based on validation performance. The default weighting for BCE loss is set to $[1, 2, 2, 3]$. All experiments are conducted on a single NVIDIA A100 GPU.

4.2. Evaluation

We evaluate performance using RMSE and Percentage of Correct Keypoints (PCK). RMSE is the square root of the mean of the squared distances between predicted and ground truth coordinates, while PCK measures the proportion of predictions that fall within a defined distance threshold from the ground truth:

$$\text{dist} = \sqrt{(x_{\text{pred}} - x_{\text{label}})^2 + (y_{\text{pred}} - y_{\text{label}})^2}. \tag{6}$$

Predictions within this threshold are treated as correct. For fully visible and partially occluded balls, we adopt a 4-pixel threshold, consistent with the ball's approximate 4–5 pixel diameter at a resolution of 288 × 512. For fully occluded frames, we relax the threshold to 10 pixels — roughly twice the ball's physical size — to account for the inherent uncertainty in annotating invisible objects and the variation introduced by trajectory-based interpolation. This avoids overpenalizing predictions when even human annotators cannot reliably localize the ball.

Table 3
Performance comparison across Tennis, TTA, Badminton and TT datasets. RMSE and PCK are reported separately for each visibility level. For fully visible and partially occluded cases, PCK is computed at a 4-pixel threshold, while for fully occluded cases a 10-pixel threshold is used. Because the TT datasets includes only visible frames and does not distinguish visibility levels, the TT (overall) column reports performance across all TT frames. The rightmost column reports model efficiency in terms of parameter count (M) and average inference speed (FPS).

Model	Metric	Tennis			TTA			Badminton		TT (Overall)	Parameters M/FPS
		Visible	Partial Occ.	Full Occ.	Visible	Partial Occ.	Full Occ.	Visible	Not visible		
TTNet (Voeikov et al., 2020)	RMSE	25.40	72.70	48.77	34.28	36.33	39.61	40.76	43.31	4.02	7.62/40.75
	PCK	0.23	0.19	0.17	0.39	0.26	0.21	0.23	0.74	0.85	
TrackNetV2 (Sun et al., 2020)	RMSE	24.48	109.93	232.70	4.47	19.15	35.22	34.32	32.95	2.43	11.34/52.34
	PCK	0.84	0.49	0.00	0.93	0.71	0.57	0.86	0.88	0.91	
TrackNetV4 (Raj et al., 2025)	RMSE	9.51	71.43	175.68	3.71	14.21	26.63	**17.99**	284.34	**0.93**	11.34/49.88
	PCK	0.95	**0.64**	0.17	0.91	0.74	0.56	**0.92**	0.39	**0.99**	
MonoTrack (Liu and Wang, 2022)	RMSE	43.28	138.39	227.71	3.83	**10.07**	32.45	40.13	**27.67**	2.23	2.84 /139.75
	PCK	0.78	0.42	0.00	0.92	**0.74**	0.45	0.84	**0.90**	0.95	
WASB (Tarashima et al., 2023)	RMSE	16.58	105.73	264.45	4.18	20.11	37.67	27.17	56.50	3.11	1.48/68.27
	PCK	0.92	0.52	0.17	0.93	0.70	0.42	0.88	0.79	0.84	
TOTNet	RMSE	**6.07**	**63.41**	**27.98**	**2.55**	10.95	**17.17**	23.43	44.55	1.67	8.65/28.50
	PCK	**0.95**	0.61	**0.67**	**0.94**	0.68	**0.60**	0.89	0.84	0.97	

We do not report precision or recall because, for the Tennis, TTA, and TT datasets, every frame in the benchmark is assumed to contain exactly one valid ball annotation. These datasets are rally-centric and comprise only ball-in-play segments in which out-of-frame occurrences are effectively negligible. As a result, each model prediction is matched directly to a single ground-truth coordinate, and binary classification metrics (e.g., true negatives, multiple detections) are not meaningful. Under this evaluation protocol, precision and recall reduce to the information already conveyed by localization metrics such as PCK and RMSE, and thus offer no additional insight.

In contrast, the badminton dataset contains a substantial number of out-of-frame instances. For such frames, we assign the ball a canonical coordinate of (0,0) and compute RMSE and PCK accordingly.

4.3. Other models

In this work, we used official implementations whenever available; otherwise, we re-implemented the models following the original descriptions. The set of baselines includes TrackNetV2 (Sun et al., 2020), TrackNetV4 (Raj et al., 2025), MonoTrack (Liu and Wang, 2022), WASB (Tarashima et al., 2023), and TTNet (Voeikov et al., 2020). Among these, **only TrackNetV2** was re-implemented, as its original codebase is written in TensorFlow while our framework is built in PyTorch, requiring a reproduction for consistent evaluation. All other methods were used through their official or publicly released PyTorch implementations. Implementation details, adaptations, and resolution settings are provided in our GitHub repository. WASB (Tarashima et al., 2023) and TrackNetV4 (Raj et al., 2025) represent the most recent and advanced approaches, serving as strong baselines for comparison.

4.4. Results

The results across all four datasets are summarized in Table 3. Our proposed model, TOTNet, consistently outperforms existing state-of-the-art methods, particularly under occlusion. On the tennis dataset, TOTNet reduced the RMSE for partially occluded objects from 105.73 to 63.41. More notably, for fully occluded cases, it achieved an RMSE of 27.98 and improved PCK to 0.67, demonstrating its ability to recover ball trajectories even under complete occlusion. On the TTA dataset, which features the highest density of occlusion, TOTNet again outperformed all other models. In the fully occluded setting, it achieved an PCK of 0.60 and an RMSE of 17.17, underscoring its robustness in challenging real-world conditions. Similar improvements were observed on the TT and badminton datasets, where TOTNet matched or surpassed state-of-the-art performance across all visibility levels. These results highlight the effectiveness of leveraging temporal and spatial context for accurate tracking in both clear and occluded scenarios.

Although TOTNet achieves state-of-the-art performance across all benchmarks, we observe several recurring failure modes. First, during extended full occlusions (>5–6 frames), the predicted trajectory may drift when neither short-term motion cues nor contextual features provide sufficient constraints. Second, in a small number of fast rallies, motion blur combined with visually ambiguous backgrounds (e.g., white advertising boards or spectator clothing) can temporarily cause loss of the ball as shown in Fig. 7. Third, in the badminton dataset, occasional out-of-frame sequences lead to temporary tracking discontinuities despite the model's ability to re-acquire the object once it re-enters the frame. These cases are rare but highlight the fundamental limits of single-view tracking under extreme occlusion and appearance ambiguity.

4.5. Ablation studies

4.5.1. Online vs. Offline

We conducted an ablation study on the TTA dataset to analyze the effect of temporal context length (N frames) and inference directionality. To clarify the mechanism, we define our two inference modes as follows:

- **Online (uni-directional):** This mode simulates real-time prediction. To predict the ball's position in a target frame t, the model is given a sequence of N consecutive frames ending with the target frame: $[t - N + 1, \ldots, t]$. This approach is causal, using only past and present information.
- **Offline (bi-directional):** This mode is designed for post-processing applications where the entire video is available. To predict for a target frame t, the model uses a centered sliding window of N frames, accessing both past and future context: $[t - \frac{N-1}{2}, \ldots, t, \ldots, t + \frac{N-1}{2}]$. This requires access to future frames relative to the target, making it non-causal but allowing for more robust predictions, especially during occlusion.

The results of this study are shown in Table 4. As expected, offline inference, which leverages both past and future context, outperforms online inference when the number of input frames is moderate. However, we observe a decline in offline performance as the input length increases to 9 frames, suggesting that excessive context may introduce noise or reduce temporal precision.

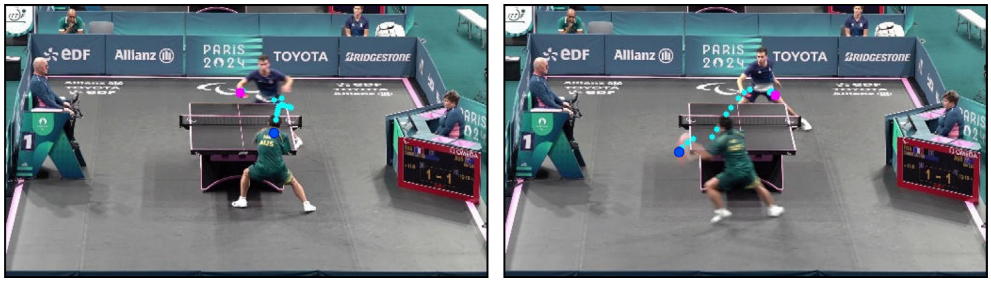


Fig. 7. Failure cases observed in fast rally sequences, where motion blur combined with occlusion leads to temporary tracking errors. Pink circles indicate the predicted ball positions, and black circles denote the ground-truth locations.

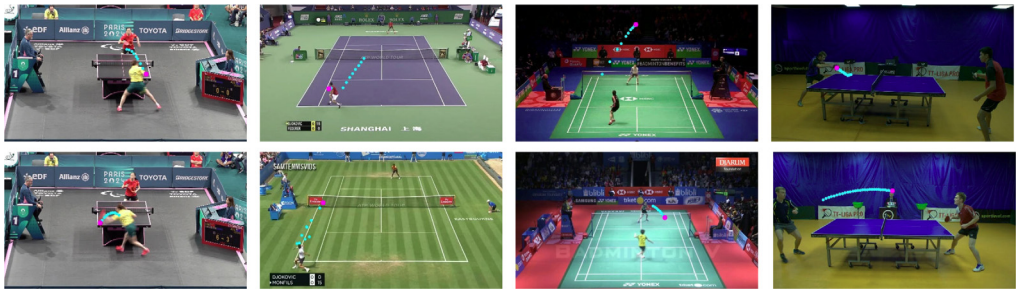


Fig. 8. Tracking examples from trained TOTNet including TTA, Tennis, Badminton.

Table 4
Ablation study on temporal context length (N frames) and directionality. Results are reported as RMSE and PCK. For fully visible and partially occluded cases, PCK is computed at a 4-pixel threshold, while for fully occluded cases a 10-pixel threshold is used. FPS is measured on an A100. Online = uni-directional (real-time), Offline = bi-directional (offline).

Frames N	Direction	Metric	Visible	Part. Occ.	Full Occ.	FPS
3	Online	RMSE	2.80	11.47	29.56	35.37
		PCK	0.93	0.71	0.52	
	Offline	RMSE	2.28	**8.72**	16.02	
		PCK	0.93	**0.87**	0.52	
5	Online	RMSE	2.55	10.95	**17.17**	28.50
		PCK	0.94	0.68	0.60	
	Offline	RMSE	2.21	9.78	17.49	
		PCK	0.95	0.75	0.57	
7	Online	RMSE	2.87	13.54	18.90	18.81
		PCK	0.93	0.69	0.53	
	Offline	RMSE	**2.03**	24.41	21.90	
		PCK	**0.94**	0.76	**0.65**	
9	Online	RMSE	3.84	12.53	18.36	13.16
		PCK	0.93	0.68	0.59	
	Offline	RMSE	3.57	23.72	22.65	
		PCK	0.93	0.79	0.64	

4.5.2. Component isolation

To validate each component's effectiveness, we conducted an ablation study on the TTA datasets using TOTNet as the base model. Table 5 shows how each progressive modification improved performance. First, the visibility-weighted BCE loss prioritized occluded and low-visibility frames, enhancing robustness under occlusion. Next, occlusion augmentation simulated diverse scenarios, forcing the model to rely on temporal and spatial context, significantly boosting accuracy in challenging conditions. Finally, integrating optical flow did not improve performance. Instead, it introduced additional noise that led to unstable predictions, particularly under rapid motion. Moreover, the inclusion of optical flow significantly reduced inference speed — dropping the model to 22 FPS — making it impractical for real-world deployment where real-time operation is required. These components complementarily enhanced TOTNet's performance, with the base model already outperforming the best existing methods (Tarashima et al., 2023), highlighting its superior architecture.

5. Deployment workflow

TOTNet, trained on the TTA dataset, has been deployed within an elite-level Paralympic table tennis analytics workflow as an automated ball-position detection system as shown in Fig. 8. Previously, professional sports analysts manually annotated ball positions — particularly bounce locations — for every rally, a process that could take 3–4 h for a single match. With our approach, a complete match can now be processed in approximately 6–7 min. When combined with an event-detection module, bounce locations relative to the table can be extracted in under 30 s for a single game. This integration not only reduces annotation time by over 90% but also enables rapid turnaround for post-match analysis, tactical review, and performance feedback.

6. Conclusion

We introduce the task of *visibility-aware occlusion-robust ball tracking* for racket sports, addressing a gap in current sports analytics where most benchmarks lack dense occlusion annotation and visibility-specific evaluation. To support research in this domain, we present the TTA dataset — the first professionally annotated benchmark from Paralympic table tennis with 2396 occlusion cases (998 fully occluded), over 19% of total frames, captured under realistic single-view conditions. Alongside the dataset, we define an evaluation protocol that measures performance across visibility levels, enabling systematic benchmarking of occlusion robustness.

Table 5
Ablation study on the TTA dataset across different visibility levels. RMSE and Accuracy are reported separately for each method. WBCE = Weighted Binary Cross Entropy, Aug. = Occlusion Augmentation, OF. = Optical Flow.

Method	WBCE	Aug.	OF.	Visible	Part. Occ.	Full Occ.
RMSE						
TOTNet (Baseline)	–	–	–	3.20	14.36	24.82
TOTNet (WBCE)	✓	–	–	2.75	13.51	20.74
TOTNet (Aug.)	–	✓	–	2.83	11.26	17.89
TOTNet (WBCE + Aug.)	✓	✓	–	2.55	10.95	17.17
TOTNet (WBCE + Aug. + OF.)	✓	✓	✓	2.78	13.25	27.46
PCK						
TOTNet (Baseline)	–	–	–	0.926	0.667	0.508
TOTNet (WBCE)	✓	–	–	0.945	0.671	0.546
TOTNet (Aug.)	–	✓	–	0.927	0.682	0.591
TOTNet (WBCE + Aug.)	✓	✓	–	0.936	0.685	0.599
TOTNet (WBCE + Aug. + OF.)	✓	✓	✓	0.925	0.732	0.561

To establish a strong baseline, we propose TOTNet — a system integrating temporal preservation, motion modeling, and occlusion-targeted augmentation — and evaluate it across four racket-sport datasets. Results show that TTA exposes substantial performance gaps in existing methods and that TOTNet significantly improves tracking under challenging occlusion scenarios.

Beyond Paralympic table tennis, this problem formulation, dataset, and protocol can generalize to other low-resource and adaptive sports where broadcast-quality footage is unavailable and occlusions are frequent. This work enables reproducible benchmarking and provides the sports analytics community with tools to advance fair officiating, performance analysis, and technology access in underrepresented domains.

Future work will expand the current dataset to a larger scale with additional matches from different tournaments and varied camera angles, enhancing its value as a comprehensive benchmark. We also plan to explore lighter architectures and more recent models, including transformer-based approaches, to establish stronger and more diverse baselines for benchmarking.

CRediT authorship contribution statement

Hao Xu: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Arbind Agrahari Baniya:** Writing – review & editing, Methodology. **Sam Wells:** Resources. **Mohamed Reda Bouadjenek:** Writing – review & editing. **Richard Dazeley:** Writing – review & editing. **Sunil Aryal:** Writing – review & editing, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Hao Xu reports statistical analysis was provided by Paralympics Australia. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We gratefully acknowledge the support and collaboration of the Paralympics Australia table tennis team, whose expertise and feedback were invaluable in the collection, annotation, and validation of the TTA dataset. Their contributions greatly enhanced the quality and real-world relevance of this work.

Data availability

All data and codes are shared through GitHub link.

References

Buric, M., Pobar, M., Ivasic-Kos, M., 2018. Ball detection using YOLO and mask R-CNN. In: 2018 International Conference on Computational Science and Computational Intelligence. CSCI, IEEE, pp. 319–323.

Chao, V., Nguyen, H.Q., Jamsrandorj, A., Oo, Y.M., Mun, K.-R., Park, H., Park, S., Kim, J., 2024. Tracking the blur: Accurate ball trajectory detection in broadcast sports videos. In: Proceedings of the 7th ACM International Workshop on Multimedia Content Analysis in Sports. pp. 41–49.

Chen, Y.-J., Wang, Y.-S., 2023. Tracknetv3: Enhancing shuttlecock tracking with augmentations and trajectory rectification. In: Proceedings of the 5th ACM International Conference on Multimedia in Asia. pp. 1–7.

Cui, Y., Hou, B., Wu, Q., Ren, B., Wang, S., Jiao, L., 2021. Remote sensing object tracking with deep reinforcement learning under occlusion. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T., 2015. Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2758–2766.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv:1311.2524 URL https://arxiv.org/abs/1311.2524.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27.

Halbinger, J., Metzler, J., 2015. Video-based soccer ball detection in difficult situations. In: Sports Science Research and Technology Support: International Congress, IcSPORTS 2013, Vilamoura, Algarve, Portugal, September 20-22, 2013. Revised Selected Papers. Springer, pp. 17–24.

Hu, Q., Scott, A., Yeung, C., Fujii, K., 2024. Basketball-SORT: an association method for complex multi-object occlusion problems in basketball multi-object tracking. Multimedia Tools Appl. 83 (38), 86281–86297.

Huang, Y.-C., Liao, I.-N., Chen, C.-H., İk, T.-U., Peng, W.-C., 2019. 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, IEEE, pp. 1–8.

Kamble, P.R., Keskar, A.G., Bhurchandi, K.M., 2019a. Ball tracking in sports: a survey. Artif. Intell. Rev. 52, 1655–1705.

Kamble, P.R., Keskar, A.G., Bhurchandi, K.M., 2019b. A deep learning ball tracking system in soccer videos. Opto-Electron. Rev. 27 (1), 58–69.

Komorowski, J., Kurzejamski, G., Sarwas, G., 2019. Deepball: Deep neural-network ball detector. arXiv preprint arXiv:1902.07304.

Kortylewski, A., He, J., Liu, Q., Yuille, A.L., 2020. Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8940–8949.

Kukleva, A., Khan, M.A., Farazi, H., Behnke, S., 2019. Utilizing temporal information in deep convolutional network for efficient soccer ball detection and tracking. In: RoboCup 2019: Robot World Cup XXIII 23. Springer, pp. 112–125.

Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D., 2017. Temporal convolutional networks for action segmentation and detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 156–165.

Li, W., Liu, X., An, K., Qin, C., Cheng, Y., 2023. Table tennis track detection based on temporal feature multiplexing network. Sensors 23 (3), 1726.

Li, K., Malik, J., 2016. Amodal instance segmentation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer, pp. 677–693.

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. Springer, pp. 21–37.

Liu, P., Wang, J.-H., 2022. MonoTrack: Shuttle trajectory reconstruction from monocular badminton video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3513–3522.

Loshchilov, I., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Maksai, A., Wang, X., Fua, P., 2016. What players do with the ball: A physically constrained interaction modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 972–981.

Naik, B.T., Hashmi, M.F., 2023. YOLOv3-SORT: detection and tracking player/ball in soccer sport. J. Electron. Imaging 32 (1), 011003.

Naik, B.T., Hashmi, M.F., Bokde, N.D., 2022. A comprehensive review of computer vision in sports: Open issues, future trends and research directions. Appl. Sci. 12 (9), 4429.

Patraucean, V., Handa, A., Cipolla, R., 2015. Spatio-temporal video autoencoder with differentiable memory. arXiv preprint arXiv:1511.06309.

Raj, A., Wang, L., Gedeon, T., 2025. Tracknetv4: Enhancing fast sports object tracking with motion attention maps. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 1–5.

Redmon, J., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39 (6), 1137–1149.

Reno, V., Mosca, N., Marani, R., Nitti, M., D'Orazio, T., Stella, E., 2018. Convolutional neural networks based ball detection in tennis games. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1758–1764.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp. 234–241.

Saleh, K., Szénási, S., Vámossy, Z., 2021. Occlusion handling in generic object detection: A review. In: 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics. SAMI, IEEE, pp. 000477–000484.

Sun, N.-E., Lin, Y.-C., Chuang, S.-P., Hsu, T.-H., Yu, D.-R., Chung, H.-Y., İk, T.-U., 2020. Tracknetv2: Efficient shuttlecock tracking network. In: 2020 International Conference on Pervasive Artificial Intelligence. ICPAI, IEEE, pp. 86–91.

Tarashima, S., Haq, M.A., Wang, Y., Tagawa, N., 2023. Widely applicable strong baseline for sports ball detection and tracking. arXiv preprint arXiv:2311.05237.

Teimouri, M., Delavaran, M.H., Rezaei, M., 2019. A real-time ball detection approach using convolutional neural networks. In: Robot World Cup. Springer, pp. 323–336.

Vaswani, A., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst..

Voeikov, R., Falaleev, N., Baikulov, R., 2020. TTNet: Real-time temporal and spatial video analysis of table tennis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 884–885.

Wang, X., Shrivastava, A., Gupta, A., 2017. A-fast-rcnn: Hard positive generation via adversary for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2606–2615.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al., 2020. Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 43 (10), 3349–3364.

Yu, J., Liu, Y., Wei, H., Xu, K., Cao, Y., Li, J., 2024. Towards highly effective moving tiny ball tracking via vision transformer. In: International Conference on Intelligent Computing. Springer, pp. 368–379.

Yu, C., Xiao, B., Gao, C., Yuan, L., Zhang, L., Sang, N., Wang, J., 2021. Lite-hrnet: A lightweight high-resolution network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10440–10450.

Zhou, Q., Zhong, B., Zhang, Y., Li, J., Fu, Y., 2018. Deep alignment network based multi-person tracking with occlusion and motion reasoning. IEEE Trans. Multimed. 21 (5), 1183–1194.

Zou, T., Jiangning, W., Yu, B., Qiu, X., Zhang, H., Du, X., Liu, J., 2024. Fast moving table tennis ball tracking algorithm based on graph neural network. Sci. Rep. 14, http://dx.doi.org/10.1038/s41598-024-80056-3.