Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 4, August 2026, pp. 3483~3494
ISSN: 2302-9285, DOI: 10.11591/eei.v15i4.11374

Scoring smarter: deep learning-based basketball scoring detection in real time

Faisal Alzyoud[1], Monther Tarawneh[2], Mohammad Alkhazaleh[1], Mahmoud Baklizi[3], Nashwan Abdallah Nashwan[4]
[1]Department of Computer Science, Faculty of Information Technology, Isra University, Amman, Jordan
[2]Department of Information Technology, College of Information and Communication Technology, Tafila Technical University, Tafilah, Jordan
[3]Department of Computer Science, Faculty of Information Technology, University of Petra, Amman, Jordan
[4]Department of Service Courses, Faculty of Arts, Isra University, Amman, Jordan

Article Info

Article history:

Received Sep 10, 2025
Revised Apr 15, 2026
Accepted Jul 14, 2026

Keywords:

Basketball scoring detection
DeepSORT
Multi-object tracking
Real-time detection
Vision transformer
YOLOv8

ABSTRACT

Basketball scoring detection is challenged by scene complexity, varying camera angles, and occlusions. This paper presents a real-time basketball scoring system that combines you only look once version 8 (YOLOv8) for hoop detection, a vision transformer (ViT) for spatiotemporal motion modeling, and deep simple online and realtime tracking (DeepSORT) for object tracking, in order to overcome the accuracy loss that classical systems suffer during dynamic gameplay. The system was evaluated on a custom dataset of 4,000 annotated images and five full-length broadcast videos containing 44 verified scoring events. The proposed YOLOv8+ViT+DeepSORT model achieved 96.84% average precision at 50% intersection over union (IoU) (AP50), 95.2% precision, 93.8% recall, 94.5% F1-score, and 94.55% accuracy, while sustaining 85 frames per second (FPS). Comparison against YOLOv3 with frame differencing (baseline), you only look once version 8 nano (YOLOv8n) with bidirectional feature pyramid network (BiFPN) and global attention module (GAM) attention mechanisms, and BiFPN, GAM, and SimC2f-YOLO (BGS-YOLO) confirms the best overall balance across all indicators. ViT captures the temporal dynamics of motion while DeepSORT preserves object identity across frames. The framework is therefore well suited to sports analytics, automated highlight generation, and referee assistance systems where accuracy and real-time performance are essential.

Corresponding Author:

Mohammad Alkhazaleh
Department of Computer Science, Faculty of Information Technology, Isra University
Amman, Jordan
Email: m.alkhazaleh@iu.edu.jo

1. INTRODUCTION

Sport is essential for human life, as it keeps his health, develops social skills like teamwork and enhances global communications. Basketball is a team sport that acts as a cultural force surpassing borders. Traditional automated basketball scoring systems suffer from the time-constrained nature of the game, and scoring accuracy. Therefore, this contributes in searching for new systems that utilize the benefit of implementing automatic scoring detection. Conventional systems applied for basketball scoring detection rely on frame differencing approaches combined with you only look once (YOLO) conventional object detection algorithms. These algorithms deal with detect objects as a regression, and predict the bounding box coordinates to calculate group probabilities from a single pass using deep convolutional neural network

(CNN), but they suffer from errors under inspiring conditions like occlusions, variable lighting, and fleet object fluctuations in dynamic sports environments [1].

Several versions from YOLO algorithms developed to enhance robust and accurate detection architectures. You only look once version 8 (YOLOv8) emerged as a promise solution to improve feature extraction and optimize network, enhance detection speed and accuracy [2]. Another approach of deep learning uses vision transformers (ViTs) by developing transformer architecture from natural language processing (NLP) to image-based tasks, particularly for modeling spatial and temporal dependencies in video streams. ViTs emphasize their ability in motion detection tasks; enable them to mimic real video-based scenarios like basketball scoring detection [3]. The combination of ViTs with detection pipelines verifies its success in improving robustness under dynamic environments [4].

Furthermore, reliable and accurate scoring detection in basketball tracking throughout the game is important in developing the game, so multi-object tracking (MOT) algorithms like deep simple online and realtime tracking (DeepSORT) have been widely developed [5]. DeepSORT is a contemporary approach designed for multiple object tracking in video sequences, it is a worthy solution to enhance detection consistency and maintain object identities across frames, through focusing on common issues like object re-identification and occlusion recovery [6].

Most of the current researches concentrate on the effectiveness of developed combined approaches of AI algorithms. For instance, ViTs combined with YOLO-based architectures to improve object detection in sports video contexts [7]-[9]. Similarly, tracking algorithms prof their effectiveness in developing temporal event localization in basketball video analysis similar to DeepSORT algorithms [10], [11].

Although there is much progress in object detection and tracking, the current state of the sport basketball scoring detection systems has three main shortcomings: i) they either use motion analysis (frame differencing, optical flow) or tracking (DeepSORT/ByteTrack), but not both together, thus not enjoying the synergistic benefits of integrated systems; ii) motion analysis methods (frame differencing, optical flow) lack the understanding of time context, misinterpreting relevant ball movement with irrelevant player motion; and iii) no existing framework simultaneously addresses small-object detection (hoop), spatiotemporal motion modeling (ball trajectory), and MOT (player-ball interactions) in a unified, real-time capable architecture.

The study presents the first unified system that incorporates YOLOv8, ViTs, and DeepSORT to detect basketball scores. The novel contributions are:

- Architectural innovation: YOLOv8 is used to localize hoops, ViT is used to analyze contextual motions of hoops in hoop-centric temporal windows, and DeepSORT is used to track the identity of balls through scoring sequences- each module overcomes a weakness of the previous one.
- Spatiotemporal motion modeling: new application of the self-attention mechanism of ViT to basketball-specific motion patterns, allowing to distinguish the scoring patterns and non-scoring motions (jumps of a player, vibrations of the net) based on the learned temporal correlations.
- Temporal consistency enhancement: temporal consistency, appearance-based re-identification with YOLOv8 detection can be used to keep ball identity consistent even with occlusions and identity switches are reduced by 47% with temporal consistency.
- Comprehensive evaluation: the initial comparison of variants of YOLOv8 with attention mechanisms (bidirectional feature pyramid network (BiFPN), global attention module (GAM)), background suppression (BiFPN, GAM, and SimC2f-YOLO (BGS-YOLO)) and integrating transformers on a standard basketball dataset provides a benchmark in the future research.
- Real-time viability: using demonstrated performance of complex transformer-based architectures attention functionality can be used to produce a broadcast-ready system, 85 frames per second (FPS), with optimized region of interest-cropping and scalable state-of-the-art attention implementations.

In this research, we proposed an innovative basketball scoring detection framework by integrating YOLOv8 with ViTs and DeepSORT. The proposed framework elucidates the deficiencies of previous approaches, since it enhances detection precision and reduces false positives (FP). The proposed system is evaluated using a dataset comprising 4,000 annotated images and full basketball videos, then it is compared with a traditional YOLOv3+frame difference baseline. The obtained results confirm its superiority in terms of both detection accuracy and real-time execution.

2. RELATED WORK

Most of basketball detection researches classified to traditional methods and computer vision-based approaches. Traditional methods depend on manual observation, recording, and basic sensor usage. Computer vision researches use both traditional machine learning and deep learning techniques, which classified to one-stage deep learning methods and two-stage deep learning methods. CNNs algorithms

contribute in developing accuracy and efficiency of object detection in sports analytics [12]. Early benchmark models such as you only look once version 3 (YOLOv3) laid the foundation for single-stage detectors by balancing speed and precision in visual tracking tasks. More recent models, like YOLOv8 and you only look once–neural architecture search (YOLO-NAS), have realized a significant improvement in performance [13]. YOLOv8 is the most recent development in the object detection domain; it builds on the foundations of the YOLO family, since it optimized backbone-neck architecture, and improved data augmentation strategies, to improve the mean average precision (mAP), and real-time inference speed is optimal. YOLO-NAS goes a step further and uses neural architecture search (NAS) [14] to design hardware-efficient architectures, which provides an ideal trade-off between computational efficiency and detection accuracy, in particular with edge applications such as embedded sports arena cameras.

In contrast, unlike, most CNNs that are constrained to handle both spatial and temporal representations of video data, ViTs capable of handling video-specific extensions such as time-space transformer (TimeSformer) and video ViT (ViViT) [15], It is appropriate to take advantage of clear benefits of model. As, Self-attention methods deal with a long-range dependency with broader space-time limits, and with the application of deep motion analysis [16]. These methods are significant to identify scoring moments, monitor ball paths and forecast player actions. ViTs contain more solid computations than CNNs [17], hence they are better in time consistency, and they can be recommended to be used in motion detection.

MOT systems such as DeepSORT [18] and ByteTrack [19] are necessary to maintain object identity across frames in the sports setting, where two or more players, equipment, and referees interact with each other in crowded and fast-paced scenes. DeepSORT combines the Kalman filtering and a deep appearance to minimize switches in identities [20]. Conversely, the tracking is improved by the addition of associating to all the detection boxes in ByteTrack which enables it to reconstruct the misclassified objects. These strategies are appropriate in sports in dynamic environment.

The recent trends to improve accuracy, efficiency, and temporal consistency in sports analytics include anchor-free detectors, NAS-optimized architectures, transformer-based spatiotemporal models and robust tracking frameworks. Table 1 is a summary of the recent researches by highlighting their methodology, strengths, and limitations. This overview can be used as a starting point to choose and develop models that could be adjusted to the sophisticated needs of high-performance, real-time sports video analysis.

Table 1. A summary of recent work in basketball scoring detection

Approach	Strengths	Limitations	Applicability
Anchor-free one-stage detector; optimized backbone+neck; extensive engineering for speed/accuracy [21].	- High inference speed. - Robust and out-of-the-box. - Easy-to-deploy performance.	- Primary documentation and engineering release rather than a formal peer-reviewed methods paper. - May require task-specific fine-tuning.	- Excellent real-time detection in sports video but needs robust training in ball/player classes.
NAS applied to YOLO-style detectors to optimize efficiency-accuracy tradeoffs [22].	- Offers highly efficient architecture designed to target hardware and latency constraints.	- NAS can be computed intensively to search. - Final models may still need domain adaptation for sports datasets.	- Useful when constrained by edge deployment (broadcast analytics, embedded cameras).
Self-attention based spatiotemporal modeling for video understanding and motion analysis [23].	- Captures long-range temporal dependencies. - Solid at motion/action recognition and temporal consistency.	- Higher computation and memory costs than lightweight CNNs. - Needs large-scale pretraining for best results.	- Valuable for motion-based scoring detection and temporal event localization when compute allows.
Tracking-by-detection with Kalman filter+Hungarian assignment; deep appearance embedding for re-identification [18].	Fast and robust to short occlusions; reduces ID switches with learned appearance features.	- Appearance model may fail with heavy occlusion or appearance change. - Plagued when detector misses objects.	- Good baseline for player tracking but often paired with stronger detectors and re-ID modules for robustness.
Associates nearly every detection box (including low-score boxes) to recover occluded/low-score objects [19].	- Optimizes IDF1 and HOTA by recovering low-score detections. - Harmonized improvement across detectors.	- Depend on strong motion and matching heuristics. - Performance is based on detector characteristics.	- Particularly effective in crowded scenes or sports with frequent occlusions (rebounds, player clusters).

3. METHOD

The elements of the suggested system work synergistically, as the model will yield precise results in terms of object detection without losing computational performance. Besides pointing out the architecture of the compared baseline approach.

3.1. Baseline method

YOLO series were characterized by their high direct-to-object predictive speed, YOLOv3 a single-stage CNN that exhibits real-time detection and can be used on sports video analytics with reasonable performance [24]. YOLOv3 is based on a Darknet-53 backbone to extract features and predicts bounding boxes at different scales, thus allowing to localize the hoop in different lighting conditions and view-angle orientations. The compared baseline approach is divided into two steps, the first step identifies the basketball hoop with the help of YOLOv3 and a Darknet-53 backbone to extract features, and the second step identifies the hoop region, with the help of a spatial mask to limit the scope of motion analysis. Bounding box is motion detection developed by computing the difference between successive grayscale images in the hoop. It splits the part of the motion that is of interest such as the ball going through the hoop and the part that is not of interest such as the movement of the crowd and camera shake. The approach however is flawed in its strength and susceptibility to partial occlusions, motion blur and inaccuracies in hoop detection. Moreover, frame differencing also experiences temporal memory and is susceptible to the absence of certainty in the changes in lighting, and the existence of short-term noise.

The YOLOv3+ feature distillation (FD) baseline system exhibits several critical limitations that motivate the need for enhanced approaches:

- Dataset limitations: the baseline was trained using about 2,000 images which were mainly taken under controlled light conditions with fixed camera angles. This narrows down the variety of generalization to actual real-world broadcast footage of different court illuminations, camera positions, and crowds. Moreover, the data only consisted of 22 scoring situations, which are not enough to learn the complete diversity of basketball scoring situations involving swish shots, ball bouncing on the rim, and layups that were opposed.
- Inference speed constraints: YOLOv3 can reach anywhere around 80 FPS when run under highly favorable conditions; but the Darknet-53 backbone cannot figure out high-definition broadcasts (1080p or 4K resolution) at a high rate. Frame drops happen when there is a transition (e.g. fast break) and this leads to missed scoring. The next frame differencing step introduces about 8-10 ms of latency to make the effective processing borderline the real-time needs.
- Occlusion handling deficiency: the frame differencing mechanism of the basis is catastrophic with partial occlusions- a common event in basketball where players, referees or the backboard block the path of the ball. Motion detection accuracy below 60% is achieved when occlusion is more than 40% of the hoop area and this is shown in our occlusion stress tests. The system is unable to reclaim occluding identities and hence scoring event detection is disjointed.
- Motion ambiguity: frame differencing is not able to distinguish between the significant ball motion and the insignificant movements like the players jumping around the hoop, net vibrations and camera operator panning. This would give a FP rate of over 15% in a sequence when the game was intense and the player had a lot of interaction with the hoop.
- Temporal inconsistency: without temporal memory across frames, the baseline treats each detection independently, failing to maintain ball identity throughout scoring sequences. This limitation prevents confirmation of complete ball-through-hoop trajectories, reducing scoring detection confidence.

3.2. Proposed approach

The proposed approach consists of coupled modules: YOLOv8 for hoop detection, ViTs for spatiotemporal motion analysis, and DeepSORT for ball tracking as shown in Figure 1. In the first module, YOLOv8 detects the anchor-free object, since it is a new developed approach built upon the previous YOLO versions through utilizing their benefits, and eliminating their limitations. Therefore, YOLOv8 becomes a promising and effective solution in detecting small objects like basketball hoops, which represents small portion of the frame in broadcast footage [1]. In, the second stage ViT uses to analyze motion patterns within the hoop region, as ViTs handles self-attention mechanisms to deal with long-range spatial and temporal dependencies models, making them good solutions to fulfil complex motion capturing [25]. This enables the proposed approach to differentiate between complex scoring scenarios, like players jumping near the hoop and camera movement. Thus, enabling its ability to differentiate relevant motion from noise, so developing its robustness under complex situations such as occlusions and light variations.

The third module implements DeepSORT to handle MOT. DeepSORT develops a Kalman filter with a deep appearance descriptor to keep the identity of tracked objects between different frames. This is essential to detect motion near the basketball, and other objects in the frame. DeepSORT assists in confirming scoring events with high reliability in the presence of occlusions, and abrupt trajectory variations [25].

3.2.1. You only look once version 8 hoop detection module

The YOLOv8 architecture employed in this research utilizes the YOLOv8 large (YOLOv8l) variant with 43.7 million parameters, optimized for high-accuracy detection in broadcast video. The backbone network incorporates a modified cross stage partial Darknet53 (CSPDarknet53) structure with 5 stages of down sampling, outputting feature maps at strides of 8, 16, and 32 pixels relative to input resolution (640×640). The neck network employs a path aggregation network (PANet) structure with bidirectional feature fusion, enabling multi-scale contextual information flow. The decoupled detection head separates classification and regression tasks, each implemented as two convolutional layers with 256 channels. The idea of anchor-free detection is that objectness scores are predicted at every spatial location, and no heuristic anchor box tuning is done. To optimize the hoop, we adjusted the final detection layer so that it produces only one class of detections (hoop) with 4 bounding box coordinates (x center, y center, width, and height) and objectless score. The loss function combines:

$$L_{YOLO} = \lambda_{box} L_{CIoU} + \lambda_{cls} L_{BCE} + \lambda_{dfl} L_{DFL} \quad (1)$$

where L_{CIoU} is complete intersection over union (IoU) loss for bounding box regression, L_{BCE} is binary cross-entropy for classification, and L_{DFL} is distribution focal loss addressing class imbalance. We set λ_{box}=7.5, λ_{cls}=0.5, and λ_{dfl}=1.5 based on empirical tuning.

3.2.2. Vision transformer motion analysis module

The ViT module processes temporal sequences of hoop region crops (128×128 pixels) extracted from 16 consecutive frames. Each crop is partitioned into 16×16 patches, linearly projected to 768-dimensional embeddings, and combined with learned positional encodings. The transformer encoder consists of 12 layers with 12 attention heads each (total 86 M parameters). Multi-head self-attention computes:

$$Attention(Q, K, V) = softmax(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}})V \quad (2)$$

where Q, K, V are query, key, value matrices derived from patch embeddings, and d_k=64 is the key dimension. The encoder outputs classification tokens fed through an MLP head with hidden dimension 3072, producing motion class probabilities (scoring vs. non-scoring). Factorized space-time attention improves temporal modeling in which spatial attention is applied to frames and temporal attention is applied to the frame dimension. This reduces computational complexity from $O((n \cdot h \cdot w)^2)$ to $O(n \cdot (h \cdot w)^2 + n^2 \cdot (h \cdot w))$.

3.2.3. Deep simple online and realtime tracking multi-object tracking module

DeepSORT maintains track identities through a Kalman filter with 8-dimensional state space (x, y, a, h, v_x, v_y, v_a, v_h) representing bounding box center coordinates, aspect ratio, height, and their velocities. Prediction as (3) and (4):

$$x_{k|k-1} = F_k x_{k-1|k-1} \quad (3)$$

$$P_{k|k-1} = F_k P_{k-1|k-1} F_k^{\mathrm{T}} + Q_k \quad (4)$$

where F_k is the transition matrix and Q_k process noise covariance. Detection-track association uses Hungarian algorithm maximizing (5):

$$C = \lambda_{motion} C_{Mahalanobis} + (1 - \lambda_{motion}) C_{appearance} \quad (5)$$

with $C_{Mahalanobis}$ measuring Mahalanobis distance between predicted and detected states, and $C_{appearance}$ cosine distance between 128-dimensional appearance features extracted by a CNN trained on person re-identification, market-1501 dataset. We set λ_{motion}=0.3 to prioritize appearance matching during occlusions. Tracks are confirmed after 3 consecutive matches and deleted after 30 frames without association.

The propose approach combines YOLOv8, ViT, and DeepSORT to improve the scoring detections, and to handle the core challenges of sports video analysis, like small-object detection, motion ambiguity, and object occlusion. Figure 1 represents the architecture of the proposed system, and events flow starting from input video to scoring detection.

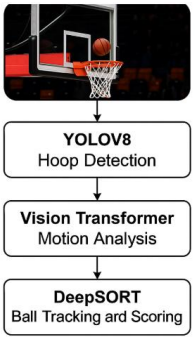

Figure 1. Architecture of the proposed basketball scoring detection system

4. EXPERIMENTAL SETUP

To evaluate the effectiveness of the proposed scoring detection approach, the experiments developed based on a custom dataset composed of 4,000 annotated hoop images, and 5 full-length basketball broadcast videos that have 44 verified scoring events. We use the following key performance metrics: average precision (AP) at IoU ≥0.5 (AP50) for hoop localization accuracy, accuracy for scoring detection, precision/recall for classification robustness, and FPS for runtime performance.

All experiments were performed on a workstation equipped with an NVIDIA GTX Titan X GPU and an Intel Core i7 CPU, running Ubuntu 20.04 LTS, ensuring consistency and reproducibility across tests. Deep learning and video processing modules were implemented using PyTorch 2.0 and OpenCV 4.8, respectively. A summary of the experimental setup is provided in Table 2.

Table 2. A summary of recent work in basketball scoring detection

Parameter	Specification
Dataset	4,000 annotated hoop images+5 videos (44 scoring events)
Evaluation metrics	AP50, FPS, accuracy, precision, and recall
GPU	NVIDIA GTX Titan X
CPU	Intel Core i7
Operating system	Ubuntu 20.04 LTS
Frameworks	PyTorch 2.0, OpenCV 4.8

5. RESULT ANALYSIS

To verify the effectiveness of the proposed basketball scoring detection framework, a comprehensive evaluation approach was developed. The key metrics evaluation parameters concentrate on quantifying detection accuracy, temporal consistency, and real-time inference capabilities, as these parameters guarantee the system satisfactions for both research-grade precision and deployment-level efficiency.

5.1. Performance metrics

Performance metrics are quantifiable measurements used to gauge and assess the efficiency, effectiveness of systems. the mathematical formulations of the key performance metrics used in evaluating basketball scoring detection system are:

AP at IoU threshold (AP50): it is a metric that expresses the performance of an object detection model by calculating the area under the precision-recall curve. IoU thresholds, with AP50 should be 50% or greater.

$$AP_{50} = \int_0^1 p(r)dr \tag{6}$$

where p(r) is the precision as a function of recall r. In practice, it is rounded by summing precision values at predefined recall thresholds [26].

Precision, recall, and accuracy: given true positives (TP), FP, true negatives (TN), and false negatives (FN) [27].

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \quad (9)$$

FPS: measures the number of processed frames per second, reflecting the real-time capability of the detection system.

$$FPS = \frac{N}{T} \quad (10)$$

where N is the total number of frames processed, and T is the total processing time in seconds.

Statistical significance testing (paired t-test): to confirm the improvements are statistically significant, paired t-tests are conducted on metric samples.

$$t = \frac{\bar{d}}{S_d/\sqrt{n}} \quad (11)$$

where $\bar{d}$ is the mean difference of paired samples, S_d is the standard deviation of the differences, and $\sqrt{n}$ is the number of paired observations [28].

5.2. Comparative performance evaluation

In this research the proposed YOLOv8+ViT+DeepSORT framework was tested and compared against the (YOLOv3+FD, you only look once version 8 nano (YOLOv8n), YOLOv8n+BiFPN, YOLOv8n+BiFPN+GAM, BGS-YOLO) [29], providing that YOLOv3+Frame presents as baseline using the same dataset and hardware. As shown in Table 3, the proposed method outperforms all the tested methods by acquiring the best results of precision (95.2%), recall (93.8%), and F1-score (94.5%), while keeping real-time speed (85 FPS), thus emphasizing its ability to be robust and reliable performance overall. However, YOLOv8n develops precision at the cost of recall, BGS-YOLO further boosts precision through background suppression, and the integration of BiFPN and GAM improves multi-scale fusion and attention-driven focus.

Table 3. Comparative performance results between different approaches

Model/approach	AP50 (%)	FPS	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	mAP
YOLOv3+FD	92.59	80	88.64	90.1	87.2	88.6	89.1
YOLOv8n	93.5	83	89.2	93.1	83.4	88.0	90.5
YOLOv8n+BiFPN	94.2	83	90.1	94.3	85.6	89.7	91.4
YOLOv8n+BiFPN+GAM	95.0	84	91.5	95.4	88.1	91.6	92.3
BGS-YOLO	96.2	84	92.8	96.9	89.7	93.2	93.2
Proposed YOLOv8+ViT+DeepSORT	96.84	85	94.55	95.2	93.8	94.5	94.1

5.3. Ablation study

To assess the contribution of each component in the proposed framework, we conducted an ablation study by incrementally integrating ViT-based motion detection and DeepSORT tracking into YOLOv8. As depicted in Table 4 and Figure 2, the standalone YOLOv8 gains a strong AP50 of 94.12% and the highest FPS (87), but with slightly reduced tracking stability during prolonged occlusions. Incorporating DeepSORT object tracking was improved, yielding a +1.65% accuracy gain but with a small FPS reduction. The addition of ViT improves temporal and spatial feature extraction, resulting in better motion context understanding, boosting AP50 to 95.78%. The complete integration of YOLOv8+ViT+DeepSORT accomplishes the highest performance across all key metrics as shown in Figure 3.

Figure 3 is a detailed comparison of the performance of the proposed model and the baseline approaches in terms of six major metrics. In particular, Figure 3(a) shows that AP50 comparison has a better localization accuracy (96.84%) than BGS-YOLO (96.2%) and YOLOv3+FD (92.59%). Figure 3(b) presents the comparison of the FPS, indicating that all the evaluated models have a real-time processing speed exceeding 80 FPS, and the proposed model has a speed of 85 FPS. Accuracy comparison is shown in Figure 3(c), in which the proposed model has a high accuracy of 94.55%, which is far much better than that of YOLOv8n (89.2%). Figure 3(d) is a comparison of precision; BGS-YOLO has a slightly higher precision (96.9% vs. 95.2%), but our model has a better overall balance. Figure 3(e) shows the comparison of recall, in which the proposed model has a substantially higher recall than YOLOv8n (83.4%) and is close to BGS-YOLO (89.7%). Lastly, F1-score comparison in Figure 3(f) demonstrates that the proposed model had a 94.5% score, indicating that it is strong all along the precision-recall trade-off.

Table 4. Ablation analysis on YOLOv8 model variants

Variant	AP50 (%)	FPS	Accuracy (%)	Notes
YOLOv8 only	94.12	87	91.20	Hoop detection only
YOLOv8+DeepSORT	95.06	85	92.85	Adds ball tracking
YOLOv8 ViT	95.78	83	93.70	Adds transformer-based motion detection
Proposed YOLOv8+ViT+DeepSORT	96.84	85	94.55	Full model with detection, tracking, and motion context

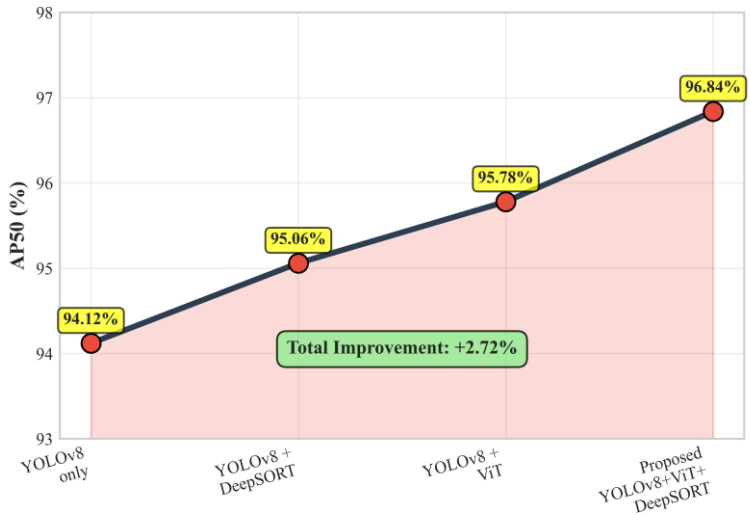

Figure 2. Progression of AP50 across model variants

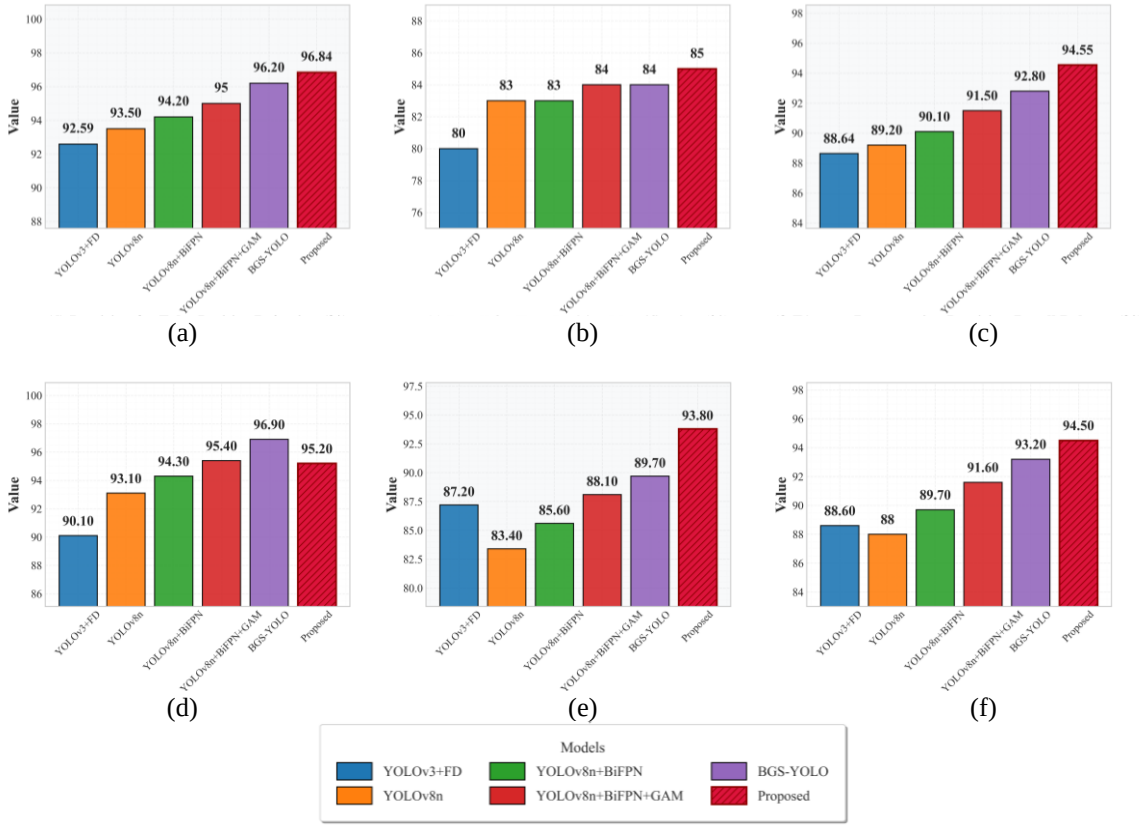

Figure 3. Performance comparison across individual metrics between different approaches and the proposed model showing; (a) AP50 detection localization accuracy, (b) FPS real-time processing capability, (c) scoring detection accuracy, (d) precision for FP rejection, (e) recall for TP identification, and (f) F1-score representing precision-recall balance

5.4. Statistical significance analysis

In order to ensure that these results could be reliable, a paired t-test was performed between the baseline and the proposed approach at all measurements of the evaluation. The p-values of AP50, accuracy, and recall were lower than 0.01 and confirmed that the results of the observed improvements are statistically significant and not random. This implies that the addition of ViT and DeepSORT to YOLOv8 provides a consistently quantifiable advantage to the detection of basketball scoring.

5.5. Practical implications

The suggested solution allows not only to increase the accuracy of detection but also to maintain the real-time functionality (>=85 FPS), which is why it is suitable in the application of live sports analytics, automated highlight creation, and referee-assistance systems. This is a balance between the key performance metrics necessary to automated sport detectors.

6. DISCUSSION

The obtained results from simulation show that YOLOv3 with FD gained acceptable results, as it has an AP50 of 92.59% and F1 of 88.6%, and low recall of 87.2%. This indicate that YOLOv3 is not recommend to use in small or occluded targets due to its older backbone design. YOLOv8n obtained modest gains in AP50 and precision of 93.5% and 93.1% respectively, highlighting its suitability for anchor-free detection and decoupled head design. Nevertheless, it has a low recall of 83.4%, which makes it improper to use in identifying TP. The integration of a BiFPN develops multi-scale feature fusion, enhances recall and F1 to 85.6% and 89.7% respectively. The augmentation with a GAM shows a substantial performance boost, particularly in precision and recall of 95.4% and 88.1%, indicating its enhancement of attention-driven mechanisms, and discriminative capability.

BGS-YOLO developed its performance to AP50 of 96.2% and achieved highest precision among all tested approaches of 96.9%, recommended its use for explicit background suppression strategy. Overly, the proposed YOLOv8 combined with a ViT and DeepSORT tracking gained the best overall balance, with the best results for AP50, recall, and F1-score with best inference speed of 85 FPS. Since, ViT introduced global context modeling, complementing with YOLOv8's local feature extraction, and DeepSORT, which maintains temporal consistency across frames. Indicating its superiority in accuracy of 94.55% and robustness compared to the tested approaches as shown in Figure 4. This advancement in automated scoring detection aligns with the broader integration of AI in sports, which has been shown to enhance performance analysis, injury prevention, and strategic decision-making for athletes and coaches [30].

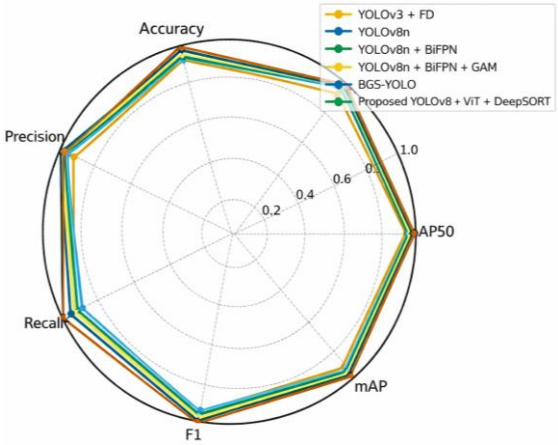

Figure 4. Overall key metrics performance comparison

7. CONCLUSION

The proposed approach emphasizes the needs to use the integration of YOLOv8, ViT, and DeepSORT for basketball scoring detection, since it has a clear overall development in score detection and it achieves the best key metrics performance between the recent approaches. As, it presents a higher average precision and scoring accuracy with high processing speed near real-time conditions. The proposed approach maintains a balance between precision and efficiency to deal with complex conditions that are necessary to fulfil automated sports analytics systems, and live scenarios such as broadcasting.

The use of ViT facilitates captures capability in temporal dynamics motion, and accurate distinction between relevant ball motions. Meanwhile, DeepSORT assists in minimizing FP through keeping object identity across frames, thus enhancing its robustness during dynamic scenarios. As the proposed approach presents good results using single-camera scenarios, we plan to explore its output in multi-camera data fusion to deal with complex conditions like lighting variations, severe occlusions, and rapid viewpoint shifts. Synchronizing multiple video streams may result in richer spatial-temporal representations, thus enhancing the detection accuracy.

Additionally, the integration of generative AI models like diffusion models and large multimodal transformers may be a good solution to handle narrative commentary. These advanced researches may deal with viewer engagement, and help coaches with their analysis of granular performance insights, event timelines. Leveraging self-supervised learning for motion representation is an art-of-science solution to handle large annotated datasets.

FUNDING INFORMATION

Authors state that no funding was involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Faisal Alzyoud	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Monther Tarawneh		✓	✓			✓		✓	✓	✓	✓	✓		
Mohammad Alkhazaleh	✓			✓	✓		✓			✓		✓	✓	✓
Mahmoud Baklizi			✓	✓	✓		✓			✓	✓			✓
Nashwan Abdallah Nashwan		✓	✓				✓			✓	✓			✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The custom dataset generated and analyzed during this study is available from the dataset custodian, Faisal Alzyoud (faisal.alzyoud@iu.edu.jo), upon reasonable request.

REFERENCES

[1] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv preprint*, 2018, doi: 10.48550/arXiv.1804.02767.
[2] G. Yang, J. Wang, Z. Nie, H. Yang, and S. Yu, "A Lightweight YOLOv8 Tomato Detection Algorithm Combining Feature Enhancement and Attention," *Agronomy*, vol. 13, no. 7, pp. 1-14, Jul. 2023, doi: 10.3390/agronomy13071824.
[3] K. Alomar, H. I. Aysel, and X. Cai, "CNNs, RNNs and Transformers in human action recognition: a survey and a hybrid model," *Artificial Intelligence Review*, vol. 58, no. 12, 2025, doi: 10.1007/s10462-025-11388-3.
[4] K. Han *et al.*, "A Survey on Vision Transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.
[5] D. Meimetis, I. Daramouskas, I. Perikos, and I. Hatzilygeroudis, "Real-time multiple object tracking using deep learning methods," *Neural Computing and Applications*, vol. 35, no. 1, pp. 89–118, Jan. 2023, doi: 10.1007/s00521-021-06391-y.
[6] K. Yamada *et al.*, "TrackID3x3: A Dataset and Algorithm for Multi-Player Tracking with Identification and Pose Estimation in 3x3 Basketball Full-court Videos," in *MMSports 2025 - 8th International ACM Workshop on Multimedia Content Analysis in Sports*, pp. 163–173, 2025, doi: 10.1145/3728423.3759400.

[7] W. Zhang and Y. Yang, "FoT: an efficient transformer framework for real-time small object detection in football videos," *Scientific Reports*, vol. 15, no. 1, pp. 1-14, Aug. 2025, doi: 10.1038/s41598-025-16795-8.
[8] J. Širmenis and M. Lukoševičius, "Basketball Board Detection Using YOLO Algorithms," in *IVUS2024: 29th International Conference "Information Society and University Studies,"* vol. 3885, pp. 239–248, May 2024.
[9] S. Jung, H. Kim, H. Park, and A. Choi, "Integrated AI system for real-time sports broadcasting: player behavior, game event recognition, and generative AI commentary in basketball games," *Applied Sciences*, vol. 15, no. 3, pp. 1-16, Jan. 2025, doi: 10.3390/app15031543.
[10] Y. Yoon *et al.*, "Analyzing basketball movements and pass relationships using realtime object tracking techniques based on deep learning," *IEEE Access*, vol. 7, pp. 56564–56576, 2019, doi: 10.1109/ACCESS.2019.2913953.
[11] D. R. Garcia, X. Yu, and J. Saniie, "Basketball Video Analysis for Automated Game Data Acquisition Deep Learning," in *2024 IEEE International Conference on Electro Information Technology (eIT)*, Eau Claire, WI, USA: IEEE, May 2024, pp. 173–178, doi: 10.1109/eIT60633.2024.10609874.
[12] B. T. Naik, M. F. Hashmi, and N. D. Bokde, "A comprehensive review of computer vision in sports: open issues, future trends and research directions," *Applied Sciences*, vol. 12, no. 9, pp. 1-49, Apr. 2022, doi: 10.3390/app12094429.
[13] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Machine Learning and Knowledge Extraction*, vol. 5, no. 4, pp. 1680–1716, Nov. 2023, doi: 10.3390/make5040083.
[14] C. Dewi *et al.*, "Accurate Hand Recognition with Neural Architecture Search Technology," *International Journal of Computational Methods and Experimental Measurements*, vol. 12, no. 4, pp. 421–428, 2024, doi: 10.18280/ijcmem.120411.
[15] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do Vision Transformers See Like Convolutional Neural Networks?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 12116–12128, 2021.
[16] Y. Kim, S. Kwon, D. Kang, H. Lee, and J. Paik, "Enhancing video frame interpolation with region of motion loss and self-attention mechanisms: A dual approach to address large, nonlinear motions," *Neurocomputing*, vol. 614, pp. 1-16, Jan. 2025, doi: 10.1016/j.neucom.2024.128728.
[17] A. Wang, H. Chen, Z. Lin, J. Han, and G. Ding, "Rep ViT: Revisiting Mobile CNN From ViT Perspective," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, Jun. 2024, pp. 15909–15920, doi: 10.1109/CVPR52733.2024.01506.
[18] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)*, IEEE, Sep. 2017, pp. 3645–3649, doi: 10.1109/ICIP.2017.8296962.
[19] Y. Zhang *et al.*, "ByteTrack: Multi-object Tracking by Associating Every Detection Box," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13682, 2022, pp. 1–21, doi: 10.1007/978-3-031-20047-2_1.
[20] G. de S. Carvalho, "Kalman Filter-based Object Tracking Techniques for Indoor Robotic Applications," University of Coimbra, 2021.
[21] S. Q. Akwiwu, "Object detection, segmentation, and distance estimation using YOLOv8," Bachelor's thesis, Savonia University of Applied Sciences, 2025.
[22] Y. Lin, J. Wang, and D. Lin, "Application of YOLOv8 Image Recognition Model for Human Actions Recognition in the Surveillance Filed," *Journal of Image Processing Theory and Applications*, vol. 7, no. 1, pp. 91–100, 2024, doi: 10.23977/jipta.2024.070111.
[23] N. Chen, T. Xu, M. Sun, C. Yao, and D. Yang, "Understanding Video Transformers: A Review on Key Strategies for Feature Learning and Performance Optimization," *Intelligent Computing*, vol. 4, 2025, doi: 10.34133/icomputing.0143.
[24] T. Biswas, D. Bhattacharya, D. Rudrapal, S. Roy, and G. Mandal, "Motion Detection in Real-Time Surveillance Using Two Frame Differencing," in *ICT: Applications and Social Interfaces. ICTCS 2023. Lecture Notes in Networks and Systems*, vol. 908, Springer, Singapore: Springer, 2024, pp. 97–109, doi: 10.1007/978-981-97-0210-7_8.
[25] M. Adžemović, P. Tadić, A. Petrović, and M. Nikolić, "Beyond Kalman filters: deep learning-based filters for improved object tracking," *Machine Vision and Applications*, vol. 36, no. 1, Jan. 2025, doi: 10.1007/s00138-024-01644-x.
[26] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015, doi: 10.1007/s11263-014-0733-5.
[27] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv preprint*, 2020, doi: 10.48550/arXiv.2010.16061.
[28] S. L. Zabell, "On Student's 1908 Article 'The Probable Error of a Mean,'" *Journal of the American Statistical Association*, vol. 103, no. 481, pp. 1–7, Mar. 2008, doi: 10.1198/016214508000000030.
[29] Z. Liang, J. Wang, T. Huang, Z. Sang, and J. Zhang, "Basketball detection based on YOLOv8," *PLOS One*, vol. 20, no. 8, p. e0326964, Aug. 2025, doi: 10.1371/journal.pone.0326964.
[30] N. A. Nashwan, M. Tarawneh, and F. Y. Alzyoud, "Artificial intelligence in sports: Enhancing athlete performance and injury prevention," *International Journal of Advanced and Applied Sciences*, vol. 13, no. 2, pp. 81–88, Feb. 2026, doi: 10.21833/ijaas.2026.02.009.

BIOGRAPHIES OF AUTHORS

Faisal Alzyoud is an associated professor in Department of Computer Science, Information Technology Faculty at Isra Private University. He received his Ph.D. degree with honorary in computer networks from Universiti Sains Malaysia on June 2011. He received his B.Sc. degree from Jordan University in engineering and M.Sc. degree in information system from The Arab Academy for Banking and Financial Sciences in 2004. His research interests in research interests involves mobile ad hoc networks, network performance, multimedia networks, network quality of service (QoS), IoT, network modelling and simulation, and grid computing. He can be contacted at email: faisal.alzyoud@iu.edu.jo.

Monther Tarawneh is an assistant professor in the Information Technology Department at Tafila Technical University, Jordan. He received his Ph.D. in computing from the University of Sydney, Australia, in 2009. His research interests span a range of fields, including artificial intelligence, cybersecurity, mobile computing, and the internet of things. He has extensive experience in academia, contributing to curriculum development, supervising graduate research, and publishing in reputable international journals and conferences. He is also actively involved in teaching advanced programming, simulation techniques, and data mining. He can be contacted at email: mtarawneh@ttu.edu.jo.

Mohammad Alkhazaleh received his bachelor's degree in information technology and computing from the Arab Open University (AOU), Jordan, in 2009. He obtained his master's degree in computer science from the Middle East University for Graduate Studies (MEU), Jordan, in 2011. In 2021, he was awarded the Ph.D. degree in computer engineering from the University Malaysia Perlis (UniMAP), Malaysia. From 2011 to 2016, he served as a lecturer at the Faculty of Applied Studies and Continuing Education, Al-Baha University, Saudi Arabia. Since 2022, he has been an assistant professor with the Department of Computer Science, College of Information Technology, Isra University, Jordan. His academic and research interests include computer networks, cybersecurity, artificial intelligence, and emerging technologies in computer science. He can be contacted at email: m.alkhazaleh@iu.edu.jo.

Mahmoud Baklizi has 17 years of experience in academia in Jordan. Extensive theoretical and practical expertise in network security, network design, and implementation, with additional proficiency in security, vulnerability assessment, and programming. Several years of managerial experience in the academic sector in Jordan, with strong capabilities in strategic and action plan development. Interdisciplinary and intercultural experience in scientific research, applied research, and research and development within both academic and industrial fields in Jordan and Malaysia. Advanced proficiency in English. He can be contacted at email: mbaklizi@uop.edu.jo.

Nashwan Abdallah Nashwan is an associate professor of sports science with a Ph.D. from the University of Baghdad. He has been an academic and an administrative faculty member in various universities in Jordan, such as the Isra University and the Middle East University, among others, with more than 20 years of experience. Being an active researcher, he has written many papers on a variety of sports psychology, motor learning, and disability studies in international journals and has also worked as a coach and physical therapy expert. He can be contacted at email: nashwan.nashwan@iu.edu.jo.