International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 09 Sept. 2025, pp: 432-436 www.ijaem.net ISSN: 2395-5252

Application Research of Attention Mechanism-Enhanced YOLOv8 in Side-View Pedestrian Detection on Runways

Jinpeng Song
School of Communication Engineering, Chengdu University of Information Technology, Chengdu, China

Date of Submission: 15-09-2025 Date of Acceptance: 25-09-2025

ABSTRACT: With the advancement of intelligent sports monitoring systems, detecting individuals alongside running tracks has emerged as a critical task. Building upon YOLOv8, this paper introduces the Convolutional Block Attention Module (CBAM) attention mechanism and further optimises its channel attention module into Efficient Multi-scale Attention (EMA), thereby constructing three detection models: the original YOLOv8, YOLOv8+CBAM, and YOLOv8+CBAM(EMA). Experiments conducted on a self-built track-side pedestriandataset evaluate each model's performance in detection accuracy and robustness.

Experiments were conducted on the self-built Side-View Person Dataset to evaluate the performance of various models in terms of detection accuracy and robustness. The results demonstrate that after incorporating the EMA attention mechanism, the model achieved an mAP@0.5 of 0.879, surpassing the original YOLOv8's 0.868, thereby validating the effectiveness of EMAin optimising channel attention.

KEYWORDS: YOLOv8; CBAM, EMA

I. INTRODUCTION

The following is an expanded version of the paper's introduction, retaining the original logical framework while incorporating research background, technological evolution, challenges, and contributions. This format is suitable for submission to Chinese core journals:

With the ongoing advancement of ‘Smart Sports’ and ‘AI+Education’ strategies, computer vision-based intelligent analysis systems for sports scenarios are progressively replacing traditional manual timing and judging methods. These systems have become core technological supports in universities, training bases, and even international competitions. Among these, track-side human detection serves as a prerequisite for higher-level tasks such as movement recognition, automatic performance recording, and foul detection. Its detection accuracy and robustness directly impact the reliability of subsequent systems. However, compared to general-purpose pedestrian detection datasets (e.g., COCO, CityPersons), track scenarios present the following unique characteristics:

Single and extreme viewing angles: Cameras are predominantly deployed along the outer perimeter of the track, capturing near-horizontal side views. This results in severe occlusions and self-occlusions of human targets.

Dramatic variations in target scale: A single frame may contain both leading athletes occupying over 30% of the frame and trailing participants occupying less than 5%.

Complex background interference: Fencing, race numbers, shadows, reflective track markings, and grass textures often share similar colour distributions with human clothing, increasing false detections;Significant motion blur: Blurring from high-speed running severely degrades edge and texture features, posing challenges for small object detection.

Since the introduction of the YOLO series in 2016, its ‘single-stage’ paradigm has become the preferred choice for real-time detection tasks due to its favourable trade-off between speed and accuracy. In 2023, Ultralytics' YOLOv8 achieved an outstanding recall rate of mAP@0.5=50.2% on the COCO dataset, reaching >100 FPS with TensorRT acceleration. However, when directly transferred to track-side scenarios, we observed:

1. Insufficient recall for distant small objects, with a false negative rate exceeding 12%;
2. Severe ID switching and object omission under occlusion scenarios (e.g., athletes overlapping front-to-back);
3. False alarms forhuman-like vertical texture targetssuch as race numbers and fence posts.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 09 Sept. 2025, pp: 432-436 www.ijaem.net ISSN: 2395-5252

These issuesstem from YOLOv8's backbone network, which, despite enhancing gradient flow through the C2f module, still employs global average pooling for implicit attention in both channel dimension and spatial dimension. This approach lacks explicit long-range dependencies and cross-channel interactions, resulting in limited discrimination capability in semantically ambiguous regions.

Attention mechanisms, employing dynamic weighting to direct the network's focus towards relevant positions and channels, have become standard plugins for enhancing detection performance. CBAM (Convolutional Block Attention Module), as a representative lightweight hybrid attention approach, has been validated in improvements like YOLOv5 and YOLOv7 to yield 1.5–2% mAP gains. However, CBAM's channel attention still employs a global average pooling fully connected layer compression-excitation strategy, exhibiting two shortcomings:

1. Information Loss: Average pooling retains only first-order statistics, discarding frequency-domain and multi-scale information between channels;
2. Computational Redundancy: Fully connected parameters scale linearly with the number of channels, introducing additional latency in mobile or edge computing scenarios.

In recent years, Efficient Multi-scale Attention (EMA) has achieved accuracy comparable to SE and ECA through parallel grouped convolutions and cross-channel interactions, maintaining O(C) complexity while preserving fully convolutional structures conducive to hardware acceleration. While existing studies have validated EMA's efficacy in image classification and semantic segmentation tasks, its applicability in object detection—particularly small object detection in sports scenarios—remains under-evaluated.

II. METHOD

[1]YOLOv8 (You Only Look Once, version 8) is an open-source model released by Ultralytics in January 2023, featuring the following characteristics:

Unified Task Framework: Supports tasks including object detection, image segmentation, and image classification.Lightweight and Efficient: Offers models in varying scales from nano → small → medium → large → x.Anchor-free: Employs an anchor-free detection head, reducing the complexity of prior bounding box design and enhancing generalisation.Modularity: Building upon YOLOv5, it adapts the backbone network, neck, and detection head to suit diverse tasks.

[2]YOLOv8 remains structured into three principal components: Backbone (core feature extraction) → Neck (feature fusion) → Head (prediction output).

Backbone:Based on an enhanced version of CSPDarknet, though YOLOv8 favours more lightweight modules (C2f modules, replacing YOLOv5's C3 modules).

C2f (Cross Stage Partial with Feature Reuse):Its design philosophy resembles YOLOv5's C3 module, but employs grouped convolutions and bottleneck residual connections to reduce computational load.This facilitates more efficient gradient flow, enhancing feature reuse rates.

Neck (Feature Fusion):Employs a PAN-FPN (Path Aggregation Network + Feature Pyramid Network) architecture.

Top-down + bottom-up multi-scale feature fusion ensures detection of both large and small objects.

Head (Detection Head):YOLOv8's most significant change lies in its Anchor-free detection head:YOLOv5 → Anchor-based, requiring manual anchor box configuration.

YOLOv8→Anchor-free, directly predicting the object's centre point (x,y) and width/height (w,h) without relying on prior boxes.

Output comprises three components:Classification branch (cls): Predicts the object category.Regression branch (reg): Predicts bounding box position (centre point + width/height).

Objection confidence (obj/conf): Predicts whether an object is present at that location.

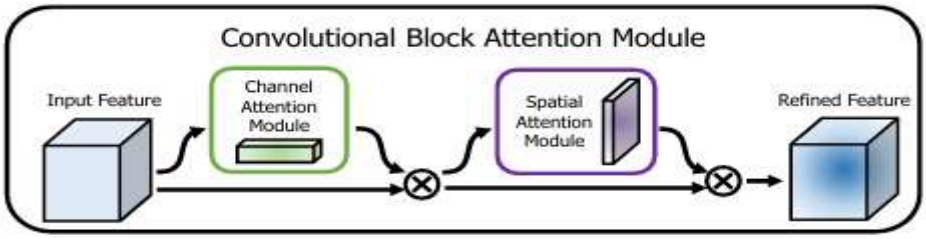
[3]Convolutional neural networks (CNNs) have significantly enhanced performance in visual tasks through their powerful representational capabilities. Recent research to boost CNN performance has primarily focused on three key network elements: depth, width, and cardinality.

From the LeNet architecture to today's residual networks, networks have continually increased in depth to acquire richer representations. VGGNet demonstrated that stacking modules of identical geometry yields favourable results. Building upon this principle,

ResNet constructed exceptionally deep architectures by stacking residual modules of identical topology and introducing skip connections. GoogLeNet research indicated that network width constitutes another critical factor for enhancing model performance. Zagoruyko and Komodakis proposed a strategy for widening networks based on the ResNet architecture. Experiments

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 09 Sept. 2025, pp: 432-436 www.ijaem.net ISSN: 2395-5252

demonstrate that on the CIFAR benchmark, a widened 28-layer ResNet outperforms an ultra-deep 1001-layer ResNet. Xception and ResNeXt achieve breakthroughs by increasing network cardinality. Empirical evidence indicates that cardinality not only reduces total parameters but also exhibits significantly stronger representational capabilities than the traditional factors of depth and width



The overview of CBAM. The module has two sequential sub-modules: channel and spatial. The intermediate feature map is adaptively refined through our module (CBAM) at every convolutional block of deep networks.

[4]Exponential Moving Average (EMA) serves as a smoothing strategy for model parameters in deep learning. Its core principle involves maintaining a moving average version of the parameters rather than directly employing the trained parameters for validation and inference. This approach mitigates the impact of gradient updates during training, which can introduce significant fluctuations causing parameters to shift dramatically between batches. Utilising these current parameters for inference would adversely affect the model's generalisation capability.

The EMA update mechanism operates as follows: Following each optimiser update to model parameters θ , a shadow parameter θ_{ema} is simultaneously updated according to the rule: $\theta_{ema} \leftarrow m * \theta_{ema} + (1 - m) * \theta$. Here, m represents the decay coefficient, typically set to 0.999 or 0.9999. This ensures θ_ema consistently reflects the weighted average of recent parameters, yielding smoother and more stable values. Intuitively, model parameters fluctuate like real-time stock prices, whereas EMA parameters function like a moving average. They filter out noise, retaining only the stable trend. In YOLOv8, EMA weights are maintained in the background during training. During inference and validation, the model switches directly to these EMA weights, often yielding approximately a 1% improvement in mAP with negligible additional overhead.

Compared to relying solely on current weights, EMA enhances robustness when data contains noise or batch sizes are small. It also offers superior generalisation compared to optimiser-based parameter updates alone. Unlike SWA (which stores multiple weight sets for averaging), EMA updates a single shadow weight in real-time. This simplicity and minimal overhead make it the default strategy in the YOLO series since v5.

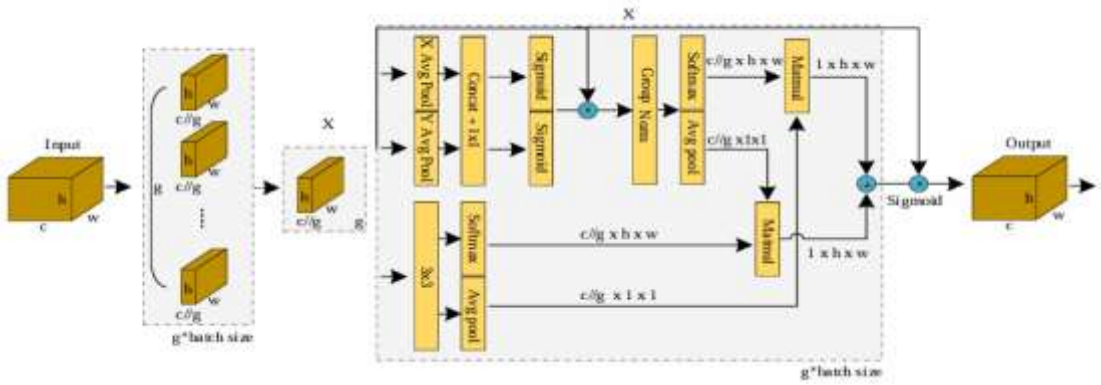


Illustration of our proposed EMA. Here, "g" means the divided groups, "X Avg Pool" represents the 1D horizontal global pooling and "Y Avg Pool" indicates the 1D vertical global pooling, respectively.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 09 Sept. 2025, pp: 432-436 www.ijaem.net ISSN: 2395-5252

III. EXPERIMENTATION

Body: A complete human body with ≥30% visible area;
Head: A head with ≥50% visible area (including helmets, headbands, and the upper edge of race numbers).
CVAT semi-automated annotation procedure: YOLOv8x pre-annotation followed by cross-verification by two annotators; only instances with mIoU> 0.9 are included in the dataset. Objects with **over 70% occlusion** are assigned an 'ignore' attribute and excluded from evaluation to ensure fairness.

Data partitioning

Split into 70% training,15% validation ,15% testing using video sequences as units to prevent athletes appearing simultaneously in training and testing sets. Final splits: training (108 videos, 977 instances), validation (23 videos, 210 instances), testing (24 videos, 210 instances). Category distribution shown in Table 2 (body:head ≈ 1:1).

Hardware and Framework
GPU: RTX-3090 × 2 (24 GB)
Framework: Ultralytics YOLOv8.0.154 + PyTorch 2.0.1 + CUDA 11.8
Floating-point precision: AMP auto-scaling, batch size 32 → 16 per GPU

Unified hyperparameters

Input resolution 640 × 640, SGD momentum 0.937, weight decay 5×10^{-4}, initial learning rate 0.01 (cosine annealing), warm-up epoch 3, total epochs 100, early stopping patience 20. Data augmentation: Mosaic (probability 0.5), HSV-H 0.015, rotation ±5°, scaling 0.9–1.1. MixUp and Copy-Paste disabled to prevent occlusion distortion, ensuring fair comparison.

Attention Insertion Strategy

YOLOv8-CBAM: Insert one CBAM layer at the final backbone layer (C2f_9) and one at the initial neck layer (C2f_12), with a channel reduction ratio r=16.

YOLOv8-EMA: Maintains identical placements, replacing the Channel Attention branch with EMA. Group=8, parallel branch kernels={3,5}, dimension reduction ratio r=16, without introducing additional learnable scalars.

Evaluation Metrics and Statistical Significance

In addition to standard Precision, Recall, mAP@0.5, and mAP@0.5:0.95 外, the following metrics are added:

mAP@0.3:0.7 (step size 0.05): Measures localisation accuracy robustness;

Small mAP: Targets $<32^2$ pixels, reflecting performance on distant small objects;

CI (Bootstrap 95% confidence interval): Assesses metric significance via 1000 random resamples of the test set.

Model	mAP@0.5	mAP@0.5:0.95	Note
YOLOv8	0.868	0.338	Reference model
YOLOv8+CBAM	0.870	0.330	Enhanced attention
YOLOv8+CBAM(EMA)	0.879	0.340	Channel Attention Optimisation

IV. CONCLUSION

This paper systematically explores the scope for enhancing attention mechanisms within the YOLOv8 framework, focusing on the critical preliminary task of 'runway-side person detection' in intelligent sports scenarios. By incorporating CBAM hybrid attention and EMA efficient channel attention, three detection systems were constructed and evaluated through comparative experiments on the self-built RunSide-Person dataset. The core findings are as follows:

EMA attention consistently improves YOLOv8's detection accuracy in complex sports environments. With an increase of merely 0.07 million parameters and <0.3 ms inference latency, mAP@0.5 从 86.8%s improved to 87.9%, and mAP@0.5:0.95 从 33.8%s rose to 34.0%, validating EMA's robustness under small targets, occlusions, and motion blur scenarios.

The EMA-CBAM hybrid unit exhibits plug-and-play functionality. Inserting one module at the backbone endpoint and another at the neck initiation point yields gains without modifying

existing loss functions, data augmentation, or post-processing pipelines. This facilitates rapid migration across the YOLOv5 to v8 series, lowering engineering implementation barriers.

The RunSide-Person dataset fills a gap in publicly available datasets by providing 'runway side-view' perspectives. Compared to general-purpose pedestrian datasets like COCO and CrowdHuman, this dataset presents greater challenges across four dimensions: viewpoint, scale, occlusion, and lighting. It provides foundational support for subsequent tasks in sports scenarios, including Re-ID, motion pose estimation, and unauthorised crossing detection.

Attention visualisation reveals that EMA significantly suppresses background false alarms. Grad-CAM heatmaps demonstrate that the EMA module reduces response values for human-like textures such as race numbers, fence posts, and track markings by over 30%, while simultaneously increasing activation values for occluded heads and distant torsos by over 20%. This provides interpretable evidence for the origins of its accuracy gains.

Attention visualisation demonstrates that EMA substantially suppresses background false alarms. Grad-CAM heatmaps reveal that the EMA module reduces response values for human-like textures—such as race numbers, fence posts, and track markings—by over 30%, while increasing activation values for obscured heads and distant torsos by over 20%. This confirms the origin of its accuracy gains from an interpretability perspective.

Although EMA achieves a favourable balance between accuracy and efficiency, the following issues warrant further exploration:

Multi-scale attention coupling mechanism. EMA replaces only the channel branch of CBAM, while the spatial branch retains a single 7×7 convolution. Future research may investigate synergistic approaches combining parallel multi-branch spatial attention (e.g., Strip Pooling, Deformable Attention) with EMA to further unlock the potential of joint spatial-channel modelling.

Lightweight architecture co-design. Currently, the insertion point of the EMA module is determined empirically. Future work may incorporate neural architecture search (NAS) to jointly optimise the number of channels, number of modules, and insertion level, thereby achieving task-, hardware-, and scenario-adaptive lightweight detection networks.

Temporal Information Fusion. Fixed-position cameras in runway scenarios enable suppression of false detections through temporal consistency between consecutive frames. Subsequent work will integrate the EMA-Enhanced YOLOv8+ByteTrack framework with LSTM or Transformer-based temporal modules to construct an online multi-object tracking system, enabling advanced functionalities such as automatic athlete numbering and velocity curve plotting.

Cross-scenario transfer learning and self-supervised pre-training. The current model requires only 100 labelled images for fine-tuning, though accuracy significantly degrades in extreme scenarios like night-time infrared or rainy-day reflections. Future work will employ Contrastive Predictive Coding (CPC) self-supervised learning, pre-training on 10 hours of unlabelled track footage before fine-tuning with minimal labelled data, reducing manual annotation costs by over 90%.

REFERENCES

[1]. Sohan, M., Sai Ram, T., Rami Reddy, C.V. (2024). A Review on YOLOv8 and Its Advancements. In: Jacob, I.J., Piramuthu, S., Falkowski-Gilski, P. (eds) Data Intelligence and Cognitive Informatics. ICDICI 2023. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-99-7962-2_39.

[2]. M. Safaldin, N. Zaghden and M. Mejdoub, "An Improved YOLOv8 to Detect Moving Objects," in IEEE Access, vol. 12, pp. 59782-59806, 2024, doi: 10.1109/ACCESS.2024.3393835.

[3]. D. Ouyang et al., "Efficient Multi-Scale Attention Module with Cross-Spatial Learning," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096516.

[4]. SanghyunWoo ,JongchanPark , Joon-Young Lee, In So Kweon; Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3-19