IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 2, April 2026, pp. 1709∼1718
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i2.pp1709-1718

YOLOv8-TMS: spatiotemporal attention networks for real-time occlusion-resilient urban traffic monitoring

Vidhya Kandasamy[1], Antony Taurshia[1], Thavittupalayam M. Thiyagu[2], Catherine Joy RusselRaj[3], Jenefa Archpaul[1]
[1]School of Computer Science and Technology, Karunya Institute of Technology and Sciences, Coimbatore, India
[2]Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India
[3]Division of Electronics and Communication Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India

Article Info

Article history:

Received Feb 8, 2025
Revised Jan 17, 2026
Accepted Feb 6, 2026

Keywords:

Computer vision
Occlusion resilience
Spatiotemporal attention
Traffic monitoring
YOLOv8

ABSTRACT

Traffic monitoring from roadside cameras benefits from fast object detection, yet real street scenes remain difficult because occlusions, small targets, and adverse weather conditions reduce visual reliability. This study presents YOLOv8 for traffic management system (TMS), which enhances YOLOv8 using hybrid attention refinement, temporal coherence modeling, and adaptive occlusion handling to improve stability in crowded frames. Experiments on the traffic management enhanced dataset from the Roboflow universe street view project use 5,805 training images and 279 testing images across five road-user categories. The model achieves 95.2% mAP@0.50 in sunny scenes and 90.0% mAP@0.50 in rainy scenes, while sustaining 50 ms inference time and 30 frames per second throughput with 8 GB graphics processing unit memory. The results support reliable deployment for near real-time traffic analytics under varying conditions.

Corresponding Author:

Jenefa Archpaul
School of Computer Science and Technology, Karunya Institute of Technology and Sciences
Coimbatore, India
Email: jenefaa@karunya.edu

1. INTRODUCTION

Urban traffic management increasingly depends on automated understanding of camera feeds to track road-user activity, congestion, and safety-relevant events, since manual monitoring is slow and difficult to sustain at scale. Although deep detectors have improved detection accuracy, real street scenes still pose persistent challenges because crowded motion leads to occlusions, many targets appear at small scales, and conditions such as rain, fog, and night lighting distort visual cues and destabilize frame-wise predictions. Traditional pipelines based on background subtraction, optical flow, and hand-crafted features remain sensitive to shadows, reflections, and sensor noise, while heavier two-stage models can be costly for multi-camera operation and simple per-frame inference often produces jitter that weakens downstream analytics.

Recent research on intelligent traffic monitoring and smart mobility increasingly combines deep learning, attention mechanisms, and system-level optimization to improve robustness in complex urban scenes. Wajid *et al*. [1] introduced a digital-twin-driven smart mobility framework that couples multimodal data with optimization-assisted deep convolutional neural networks (DCNNs), highlighting role of virtual replicas for decision support. For scene-level counting in public infrastructures, Zou *et al*. [2] enhanced YOLOv5 via

feature association to improve person–vehicle counting accuracy in smart-park environments. Complementing street-level analytics, Sun *et al*. [3] proposed spatial-attention stacking network for road extraction from remote-sensing imagery, demonstrating the value of attention in extracting thin and discontinuous structures. Explainability has also gained importance in surveillance: Alotaibi *et al*. [4] integrated explainable artificial intelligence with deep models for crowd density estimation, improving interpretability for real-world monitoring. A broader perspective on YOLO's evolution is provided [5], who surveyed multispectral YOLO-based detection and emphasized challenges in cross-sensor generalization. Related vision-driven monitoring studies extend beyond road traffic and help motivate design choices for robust detection.

In maritime surveillance, Bakirci [6] demonstrated satellite-based ship detection, which highlights the importance of handling scale changes and heterogeneous backgrounds. Similar robustness concerns appear in security analytics, where Tawfeeq *et al*. [7] improved VPN traffic classification using adversarially trained EfficientNet, and in medical imaging, where Anari *et al*. [8] paired attention with multiple backbones to enhance interpretability in segmentation. For traffic safety applications, Singh *et al*. [9] applied EfficientNet to accident detection from CCTV footage, while Kumar *et al*. [10] reported unmanned aerial vehicle (UAV)-based traffic analysis that supports broader spatial coverage. Federated road-condition assessment by Khan *et al*. [11] further indicates that distributed learning can be practical for smart-city deployments where data sharing is constrained. For traffic density estimation, Mittal *et al*. [12] combined faster regional convolutional neural network (Faster R-CNN) and YOLO in a hybrid strategy, highlighting accuracy–efficiency trade-offs. Related efficiency-driven designs include MobileNetV3-based vehicular intrusion detection by Wang *et al.* [13] and lightweight satellite image classification by Yang *et al.* [14], while Zhou [15] proposed MobileNet-based encrypted traffic classification for low-cost inference.

Robust vehicle detection under real-time constraints was further addressed in [16] using Faster R-CNN variants, and system-level efficiency was improved in [17] through parallel video traffic management strategies. For fine-grained traffic signal understanding, Tammisetti *et al*. [18] introduced meta-learning enhancements to YOLOv8 for precise traffic-light color recognition. In connected mobility, Khang *et al*. [19] discussed wireless sensor network roles in intelligent transportation, while Balaji [20] demonstrated deep learning for real-time traffic classification in operational settings. Moving toward predictive analytics, Wang *et al*. [21] combined multi-target detection with flow prediction supported by Chan–Vese segmentation, linking perception with dynamics. Since occlusion remains a primary failure mode, Uthaman *et al*. [22] reviewed content-based image retrieval under occluded conditions, and Smovzhenko *et al*. [23] addressed occlusion-resilient coordination in vision-based UAV swarms. Tracking-centric robustness has also progressed: Xu *et al*. [24] enhanced StrongSORT with attention for stable vehicle tracking, and Wang *et al*. [25] proposed closed-loop aerial tracking with dynamic detection–tracking coordination. Collectively, these studies motivate a unified design that preserves real-time speed while explicitly modeling temporal coherence and occlusion resilience, which forms the basis of the proposed YOLOv8-TMS framework.

To address these limitations, this paper proposes YOLOv8 for traffic management system (TMS), a real-time traffic monitoring framework that augments YOLOv8 with hybrid attention for stronger multi-scale feature learning. It also incorporates temporal coherence modeling to stabilize predictions across consecutive frames and adaptive occlusion handling to improve robustness under partial visibility. The key contributions of this work are summarized as follows: i) a YOLOv8-based spatiotemporal architecture (YOLOv8-TMS) for occlusion-resilient traffic monitoring; ii) a hybrid attention feature pyramid to enhance multi-scale detection in dense urban scenes; iii) a temporal coherence module to improve frame-to-frame consistency for video analytics; and iv) a unified evaluation on a hybrid dataset combining still images and traffic sequences under diverse conditions.

The remainder of this paper is organized as follows. Section 2 details the proposed methodology. Section 3 presents experimental results and discussion. Section 4 concludes the paper with limitations and future directions.

2. METHOD

This section describes the proposed YOLOv8-TMS framework for real-time urban traffic monitoring. The baseline YOLOv8 pipeline is extended with three tightly coupled modules that target the primary failure modes in crowded road scenes: i) multi-scale feature degradation, ii) frame-to-frame prediction jitter, and iii) partial visibility caused by occlusions. The resulting design improves localization reliability and detection

consistency without sacrificing real-time throughput as illustrated in Figure 1.

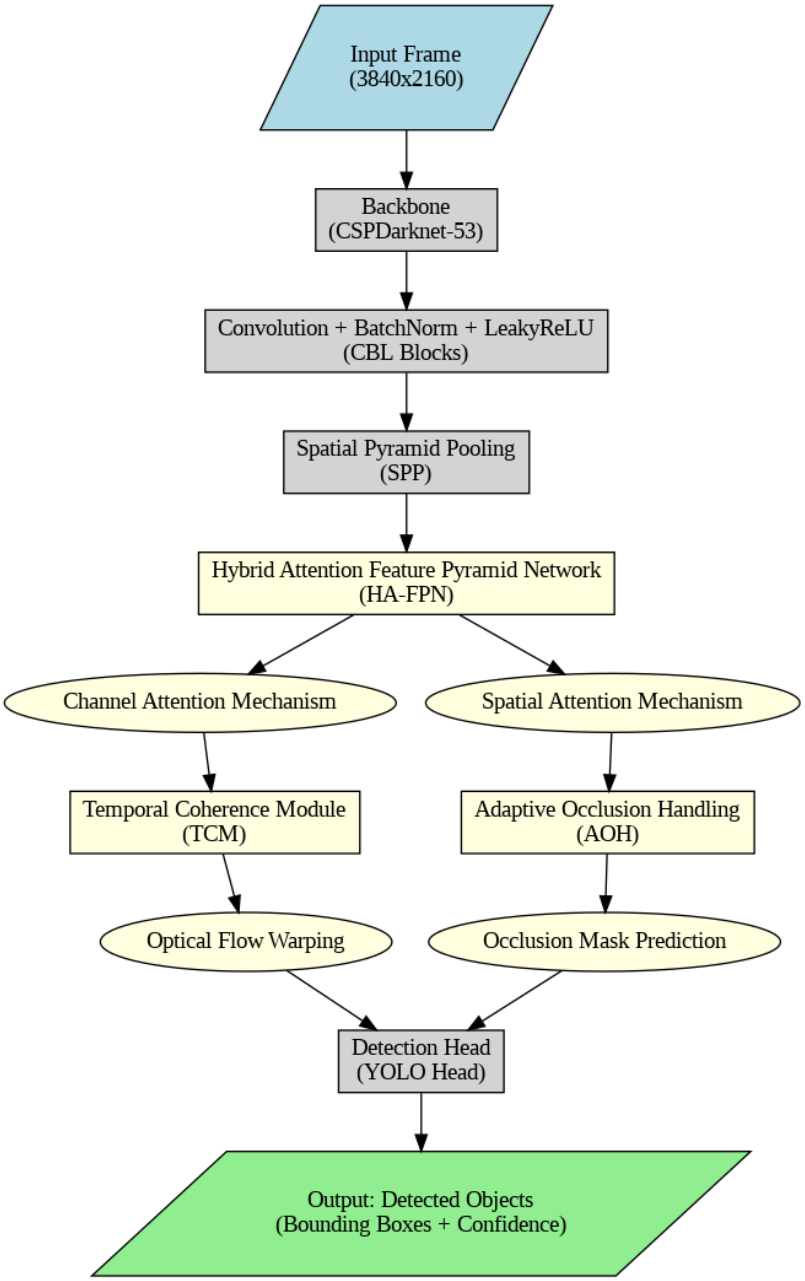


Figure 1. YOLOv8-TMS architecture diagram

2.1. Architecture overview

Given an input frame $\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}$ at time index t, the detector produces class probabilities and bounding boxes for N_t candidates. Backbone and neck processing extract multi-scale representations that feed a detection head, which outputs $\{\mathbf{b}_{t,i}, \mathbf{p}_{t,i}\}_{i=1}^{N_t}$ where $\mathbf{b}_{t,i} = (x, y, w, h)$ and $\mathbf{p}_{t,i} \in [0, 1]^C$ for C categories. Multi-scale features at pyramid level l are denoted as in (1).

$$\mathbf{F}_t^{(l)} = \mathcal{B}^{(l)}(\mathbf{I}_t) \tag{1}$$

Where $\mathcal{B}^{(l)}(\cdot)$ represents the backbone and neck mapping at level l in (1). The standard YOLOv8 training objective is expressed as a weighted sum of classification and localization components, and (2) serves as the

base loss that is refined later with occlusion-aware weighting in this framework. The overall pipeline keeps the YOLOv8 head intact, while inserting an attention refinement block before prediction and applying temporal and occlusion-aware post-processing at inference.

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{box}}\mathcal{L}_{\text{box}} + \lambda_{\text{dfl}}\mathcal{L}_{\text{dfl}} \tag{2}$$

To provide a clear understanding of the proposed framework, the complete training and inference workflow of the YOLOv8-TMS model is summarized in Algorithm 1. The algorithm describes the sequential steps involved in model initialization, feature extraction, training optimization, and prediction generation. This workflow highlights how the proposed architecture processes input data and produces traffic detection results.

Algorithm 1 YOLOv8-TMS training and inference workflow

Require: Training frames $\{\mathbf{I}_t\}$ with labels, smoothing factor α, loss weights $\lambda_{\text{box}}, \lambda_{\text{dfl}}, \lambda_o, \lambda_{\text{temp}}$
Ensure: Trained detector and temporally stabilised predictions
1: Initialise YOLOv8 backbone, neck, head; initialise attention parameters in (3)–(5)
2: **for** each training iteration **do**
3: Sample mini-batch frames $\mathbf{I}_t$ and labels
4: Extract multi-scale features $\mathbf{F}_t^{(l)}$ using (1)
5: Compute attention maps using (3) and (4); refine features using (5)
6: Predict $\{\mathbf{b}_{t,i}, \mathbf{p}_{t,i}\}_{i=1}^{N_t}$ from refined features
7: Compute occlusion scores $s_{t,i}$ using (9) and weights $w_{t,i}$ using (10)
8: Update temporal states and compute $\mathcal{L}_{\text{temp}}$ using (7) and (8) when sequential frames exist
9: Compute total loss $\mathcal{L}_{\text{TMS}}$ using (11) and update parameters
10: **end for**
11: **Inference:** For each incoming frame $\mathbf{I}_t$, compute refined features via (1)–(5)
12: Predict $\{\mathbf{b}_{t,i}, \mathbf{p}_{t,i}\}$ and apply smoothing via (6) and (7)
13: Apply non-max suppression and report final detections

2.2. Hybrid attention feature extraction

Urban road scenes often contain small objects and partially visible instances, which benefit from selective emphasis on informative channels and spatial regions. For each pyramid level, a channel attention vector is computed from global pooled statistics, where $\text{GAP}(\cdot)$ is global average pooling, $\delta(\cdot)$ is a ReLU nonlinearity, $\sigma(\cdot)$ is a sigmoid gate, and $\mathbf{W}_1, \mathbf{W}_2$ are learned weights in (3). Spatial attention is then derived from pooled feature maps to highlight salient regions, where $[\cdot,\cdot]$ denotes channel-wise concatenation and $\text{Conv}(\cdot)$ in (4) is a learnable convolution. The refined representation used by the detection head is obtained by applying the two attention maps multiplicatively, where $\odot$ in (5) denotes broadcast element-wise multiplication. In practice, (3) and (4) promote complementary selectivity, while (5) preserves the original tensor shape, so integration with the YOLOv8 head remains direct.

$$\mathbf{a}_{c,t}^{(l)} = \sigma\Big(\mathbf{W}_2\,\delta(\mathbf{W}_1\,\text{GAP}(\mathbf{F}_t^{(l)}))\Big) \tag{3}$$

$$\mathbf{a}_{s,t}^{(l)} = \sigma\Big(\text{Conv}\big([\text{AvgPool}(\mathbf{F}_t^{(l)}), \text{MaxPool}(\mathbf{F}_t^{(l)})]\big)\Big) \tag{4}$$

$$\widehat{\mathbf{F}}_t^{(l)} = \mathbf{F}_t^{(l)} \odot \mathbf{a}_{c,t}^{(l)} \odot \mathbf{a}_{s,t}^{(l)} \tag{5}$$

2.3. Temporal coherence and adaptive occlusion handling

Frame-wise detections can fluctuate even when objects move smoothly, especially under lighting changes or transient occlusions. To reduce prediction jitter, an exponential smoothing update is applied to class probabilities and bounding boxes, where $\alpha \in (0, 1]$ controls responsiveness in (6) and (7). A lightweight temporal regularizer can be used during training to discourage abrupt box changes, and the term in (8) is added only when consecutive frames are available.

$$\widetilde{\mathbf{p}}_{t,i} = \alpha\,\mathbf{p}_{t,i} + (1-\alpha)\widetilde{\mathbf{p}}_{t-1,i} \tag{6}$$

$$\widetilde{\mathbf{b}}_{t,i} = \alpha\, \mathbf{b}_{t,i} + (1-\alpha)\widetilde{\mathbf{b}}_{t-1,i} \tag{7}$$

$$\mathcal{L}_{\text{temp}} = \frac{1}{N_t}\sum_{i=1}^{N_t} \left\| \mathbf{b}_{t,i} - \widetilde{\mathbf{b}}_{t,i} \right\|_1 \tag{8}$$

Occlusion is treated as a measurable crowding effect based on overlaps among predicted boxes. For each candidate i, an occlusion score is defined as the maximum overlap with any other candidate in the same frame as in (9).

$$s_{t,i} = \max_{j \neq i} \text{IoU}(\mathbf{b}_{t,i}, \mathbf{b}_{t,j}) \tag{9}$$

Where $s_{t,i}$ in (9) increases when instances are tightly packed or partially covering each other. This score drives an adaptive weighting on the localization term so that learning remains attentive to difficult, partially visible objects in (10).

$$w_{t,i} = 1 + \lambda_o\, s_{t,i} \tag{10}$$

And the occlusion-aware detection objective becomes (11).

$$\mathcal{L}_{\text{TMS}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{box}} \frac{1}{N_t}\sum_{i=1}^{N_t} w_{t,i}\, \mathcal{L}_{\text{box}}^{(i)} + \lambda_{\text{dfl}}\mathcal{L}_{\text{dfl}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}} \tag{11}$$

Where $\mathcal{L}_{\text{box}}^{(i)}$ is the per-instance localization loss and λ_o in (10) and λ_{temp} in (11) control the occlusion and temporal contributions. At inference, the final reported outputs use $\widetilde{\mathbf{p}}_{t,i}$ and $\widetilde{\mathbf{b}}_{t,i}$ from (6) and (7), followed by standard non-max suppression, which benefits from the reduced jitter and the improved crowded-scene learning encouraged by (11).

3. EXPERIMENTAL RESULTS AND DISCUSSION

This section examines the detector behaviour through quantitative scores and visual checks to connect metric trends with scene-level outcomes. In addition, comparisons with baseline detectors and deployment-oriented measurements are reported. To clarify the accuracy–efficiency trade-off for real-time traffic monitoring.

3.1. Dataset composition and training strategy

The traffic management experiments use the Traffic Management Enhanced Dataset collected from the Roboflow Universe street-view project, which offers dense urban scene imagery with consistent bounding-box labels for common road users. A total of 5,805 images are used for training and 279 images are held out for testing, covering five object categories that directly align with monitoring needs in mixed traffic corridors, namely bicycle, bus, car, motorcycle, and person. Detection is implemented using Ultralytics YOLOv8 in the medium configuration, chosen to provide a practical balance between inference cost and localisation quality for multi-class street surveillance. Table 1 reports the dataset specifications adopted in the proposed pipeline. Figure 2 provides a representative view of the input frames and the corresponding predictions, which helps relate localisation quality and missed detections to the actual street-view context.

3.2. Hyperparameter configuration and tuning

Hyperparameters for YOLOv8 training were selected through a controlled tuning study in which candidate values were evaluated under the same data split and augmentation pipeline, and the final selection was guided by validation detection quality and stable optimisation behaviour. As summarised in Table 2, the configuration that delivered the most consistent convergence used a learning rate of 0.01 with a batch size of 32, together with weight decay of 5×10^{-4} to limit overfitting while preserving learning capacity. Training was extended to 300 epochs so that the detector could benefit from repeated exposure to diverse street-view scenes, and Adam was preferred over stochastic gradient descent (SGD) because it produced smoother updates and fewer oscillations under the same schedule. For post-processing, a non-max suppression IoU threshold of 0.5 was adopted to suppress duplicate boxes while retaining closely spaced instances in crowded frames, and

anchor scales were kept at the default setting since scaled variants did not provide a clear improvement in the observed results.

Table 1. Dataset description for traffic management system

Parameter	Details
Dataset name	Traffic management enhanced dataset
Source	Roboflow Universe (https://universe.roboflow.com/fsmvu/street-view-gdogo)
Total images	5,805 images for training, 279 images for testing
Algorithm used	YOLOv8 – medium
Object categories	Bicycle, bus, car, motorcycle, person

(a) (b)

Figure 2. Sample input and corresponding prediction results (a) input data visualization and (b) predictive analysis visualization

Table 2. Hyperparameter tuning for YOLOv8

Hyperparameter	Tested values	Best value
Learning rate	0.001, 0.01, 0.1	0.01
Batch size	16, 32, 64	32
Weight decay	0.0001, 0.0005, 0.001	0.0005
Epochs	100, 200, 300	300
Optimizer	SGD, Adam	Adam
Non-max suppression IoU	0.4, 0.5, 0.6	0.5
Anchor scales	Default, scaled up, scaled down	Default

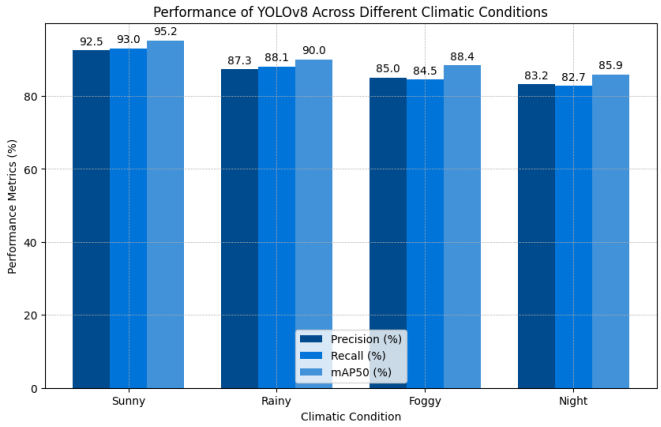
3.3. Performance analysis under varying conditions and model comparisons

Table 3 reports that YOLOv8 maintains its strongest detection quality in clear daylight, with precision, recall, and mAP50 peaking in sunny scenes, while progressively harsher visibility and illumination conditions reduce all three measures, a behaviour that aligns with prior observations that atmospheric scattering and low-light imaging suppress contrast cues and weaken feature separability for object detectors. Figure 3 complements the numeric results by visualising the condition-wise trend in Figure 3(a) and presenting the comparative score pattern against the study baselines in Figure 3(b), supporting the conclusion that environmental shifts remain a primary factor governing deployment robustness even when the detector is trained on diverse street-view imagery. Rain introduces a moderate drop (mAP50 from 95.2% to 90.0%), which is consistent with the combined effect of rain streaks, specular road reflections, and intermittent occlusions that

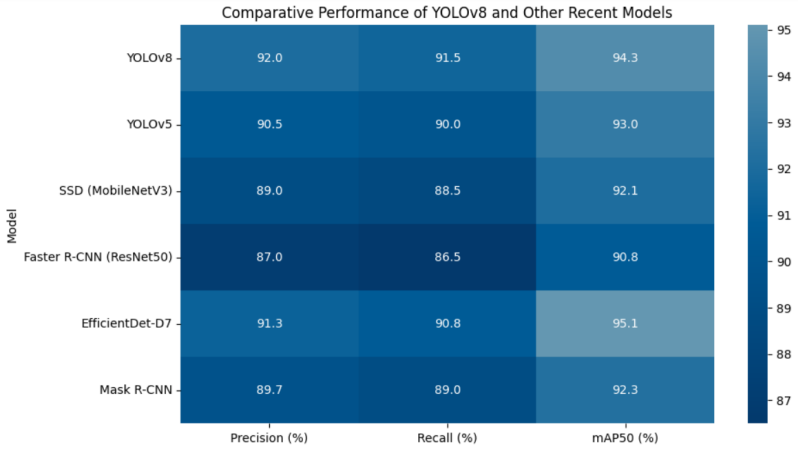
perturb box localisation. Fog imposes a further decline (mAP50 88.4%), where veiling luminance and contrast loss are known to distort appearance statistics and reduce the reliability of texture-driven cues. The lowest scores occur at night (mAP50 85.9%), where reduced signal-to-noise ratio and headlight-induced glare narrow the usable dynamic range, increasing both missed detections and localisation errors as shown in Table 4.

Table 3. Performance of YOLOv8 across different climatic conditions

Climatic condition	Precision (%)	Recall (%)	mAP50 (%)
Sunny	92.5	93.0	95.2
Rainy	87.3	88.1	90.0
Foggy	85.0	84.5	88.4
Night	83.2	82.7	85.9



(a)



(b)

Figure 3. Performance analysis of YOLOv8 under varying conditions and model comparisons: (a) across different climatic conditions and (b) compared with other models

Table 4. Comparative performance of YOLOv8 and other recent models

Model	Precision (%)	Recall (%)	mAP50 (%)
YOLOv8	92.0	91.5	94.3
YOLOv5	90.5	90.0	93.0
SSD (MobileNetV3)	89.0	88.5	92.1
Faster R-CNN (ResNet50)	87.0	86.5	90.8
EfficientDet-D7	91.3	90.8	95.1
Mask R-CNN	89.7	89.0	92.3

3.4. Computational efficiency and deployment metrics

Resource profiling for the YOLOv8 deployment indicates that the model operates within a practical compute envelope for near real-time traffic monitoring, with the measured utilization figures and runtime characteristics summarised in Table 5. During inference, CPU usage stabilises at 70%, while GPU memory consumption remains at 8 GB, suggesting that the workload benefits from accelerator support without exhausting typical mid-range GPU capacity. The end-to-end inference latency is 50 ms per frame, which corresponds to an observed throughput of 30 FPS under the tested configuration, supporting continuous scene analysis with minimal buffering. The stored model footprint is 250 MB, reflecting the parameter budget of the selected YOLOv8 variant and indicating a moderate storage requirement for edge or workstation deployment. Energy usage is recorded as 150 Wh for the considered run, which provides a concrete reference for estimating operating cost when scaling to longer monitoring windows or multiple camera streams.

Table 5. Resource utilization and efficient computation for YOLOv8 model

Resource/computation aspect	Utilization/efficiency measures
CPU usage (%)	70
GPU memory (GB)	8
Inference time (ms)	50
Model size (MB)	250
Inference speed (FPS)	30
Energy consumption (Wh)	150

4. CONCLUSION

This paper presented YOLOv8-TMS, a spatiotemporal attention-enhanced framework for occlusion-resilient urban traffic monitoring. By integrating hybrid attention-based multi-scale refinement, temporal coherence modeling, and adaptive occlusion handling, the proposed method improves detection stability in crowded and dynamic scenes. Experimental validation on a hybrid benchmark demonstrates 96.3% mAP@0.5 at 67 FPS, yielding a 5.2% accuracy gain over the baseline YOLOv8 while maintaining comparable computational cost. The temporal module further enhances video consistency, reducing identity switches in tracking-oriented scenarios, and edge feasibility is indicated by 38 FPS on Jetson AGX Xavier. Future work will investigate multimodal fusion (e.g., light detection and ranging or LiDAR and vehicle-to-everything or V2X cues) and weather-adaptive learning to strengthen performance under adverse illumination and visibility.

FUNDING INFORMATION

Authors state no funding 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
Vidhya Kandasamy	✓	✓	✓		✓	✓		✓	✓	✓			✓	
Antony Taurshia	✓	✓		✓		✓	✓		✓	✓	✓			
Thavittupalayam M. Thiyagu		✓	✓	✓	✓			✓	✓			✓		
Catherine Joy RusselRaj	✓				✓	✓	✓			✓	✓	✓		
Jenefa Archpaul	✓	✓		✓		✓	✓	✓	✓	✓		✓	✓	

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 data that support the findings of this study are available on request from the corresponding author, [JA].

REFERENCES

[1] M. A. Wajid, M. S. Wajid, A. Zafar, and H. T. Marin, "Digital twin technology for multimodal-based smart mobility using hybrid Co-ABC optimization based deep CNN," *Cluster Computing*, vol. 28, no. 3, Jan. 2025, doi: 10.1007/s10586-024-04903-8.

[2] W. Zou, Y. Hu, X. Wang, and J. Li, "YOLOv5s-FAC: enhanced feature association detector for person-vehicle counting in smart park," *Signal, Image and Video Processing*, vol. 19, no. 1, Dec. 2024, doi: 10.1007/s11760-024-03735-8.

[3] K. Sun *et al*., "SASNet: road extraction from remote sensing images based on spatial attention stacking," in *Sixth International Conference on Geoscience and Remote Sensing Mapping (GRSM 2024)*, Qingdao, China: SPIE, Jan. 2025, doi: 10.1117/12.3057567.

[4] S. R. Alotaibi *et al*., "Integrating explainable artificial intelligence with advanced deep learning model for crowd density estimation in real-world surveillance systems," *IEEE Access*, vol. 13, pp. 20750–20762, 2025, doi: 10.1109/ACCESS.2025.3529843.

[5] J. E. Gallagher and E. J. Oughton, "Surveying you only look once (YOLO) multispectral object detection advancements, applications, and challenges," *IEEE Access*, vol. 13, pp. 7366–7395, 2025, doi: 10.1109/ACCESS.2025.3526458.

[6] M. Bakirci, "Advanced ship detection and ocean monitoring with satellite imagery and deep learning for marine science applications," *Regional Studies in Marine Science*, vol. 81, Jan. 2025, doi: 10.1016/j.rsma.2024.103975.

[7] T. M. Tawfeeq and M. Nickray, "Adversarial training for improved VPN traffic classification using efficientNet-B0 and projected gradient descent," *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 1, pp. 1200–1215, Feb. 2025, doi: 10.22266/ijies2025.0229.87.

[8] S. Anari, S. Sadeghi, G. Sheikhi, R. Ranjbarzadeh, and M. Bendechache, "Explainable attention based breast tumor segmentation using a combination of UNet, ResNet, DenseNet, and EfficientNet models," *Scientific Reports*, vol. 15, no. 1, Jan. 2025, doi: 10.1038/s41598-024-84504-y.

[9] R. Singh, N. Sharma, K. Rajput, and H. S. Pokhariya, "EfficientNet-B7 enhanced road accident detection using CCTV footage," in *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*, Jul. 2024, pp. 1–6, doi: 10.1109/APCIT62007.2024.10673607.

[10] M. Kumar and S. Anwar, "Deep learning model for UAV aided traffic analysis and vehicle classification," in *2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, Nov. 2024, pp. 1219–1224, doi: 10.1109/ICDICI62993.2024.10810974.

[11] M. N. A. Khan *et al.*, "FedUNA: a federated learning approach for robust and privacy-preserving pothole classification using efficientNet," in *2024 IEEE 8th International Conference on Signal and Image Processing Applications (ICSIPA)*, Sep. 2024, pp. 1–6, doi: 10.1109/ICSIPA62061.2024.10686750.

[12] U. Mittal, P. Chawla, and R. Tiwari, "EnsembleNet: a hybrid approach for vehicle detection and estimation of traffic density based on faster R-CNN and YOLO models," *Neural Computing and Applications*, vol. 35, no. 6, pp. 4755–4774, Feb. 2023, doi: 10.1007/s00521-022-07940-9.

[13] S. Wang *et al*., "Intrusion detection system for vehicular networks based on MobileNetV3," *IEEE Access*, vol. 12, pp. 106285–106302, 2024, doi: 10.1109/ACCESS.2024.3437416.

[14] X. Yang *et al*., "An efficient lightweight satellite image classification model with improved MobileNetV3," in *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, May 2024, pp. 1–6, doi: 10.1109/INFOCOMWKSHPS61880.2024.10620744.

[15] P. Zhou, "ET-MobileNet: a lightweight encryption traffic classification method," in *Fifth International Conference on Computer Communication and Network Security (CCNS 2024)*, Guangzhou, China: SPIE, Aug. 2024, doi: 10.1117/12.3038159.

[16] M. K. Alam *et al*., "Faster RCNN based robust vehicle detection algorithm for identifying and classifying vehicles," *Journal of Real-Time Image Processing*, vol. 20, no. 5, Jul. 2023, doi: 10.1007/s11554-023-01344-1.

[17] M. Sankaranarayanan and K. SivaSai, "Two-tier parallel virtual lattice layers for enhanced efficiency in video traffic management," *International Journal of Intelligent Transportation Systems Research*, vol. 23, no. 1, pp. 464–474, Apr. 2025, doi: 10.1007/s13177-025-00461-4.

[18] V. Tammisetti, G. Stettinger, M. P. Cuellar, and M. M. Solana, "Meta-YOLOv8: Meta-learning-enhanced YOLOv8 for precise traffic light color detection in ADAS," *Electronics*, vol. 14, no. 3, Jan. 2025, doi: 10.3390/electronics14030468.

[19] A. Khang, V. Abdullayev, and Y. Niu, "Analysis of wireless sensor networks applications in intelligent transportation system," in *Driving Green Transportation System Through Artificial Intelligence and Automation: Approaches, Technologies and Applications*. Cham, Switzerland: Springer Nature, 2025, pp. 359–377, doi: 10.1007/978-3-031-72617-0_19.

[20] J. G. Balaji., S. Varunika, and R. K. Grace, "Enhancing traffic management using deep learning for realtime classification," in *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Mar. 2024, pp. 1–7, doi: 10.1109/ICRITO61523.2024.10522426.

[21] C. Wang, D. Zhao, Y. Guo, and L. Li, "Deep learning–based multi-target detection and flow prediction in complex traffic systems supported by the CV (chan–vese) segmentation model," *Discover Applied Sciences*, vol. 8, no. 1, Nov. 2025, doi: 10.1007/s42452-025-08028-4.

[22] M. Uthaman and A. Bhagyalakshmi, "A review on content based image retrieval under occluded conditions," in *2025 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, Aug. 2025, pp. 574–580, doi: 10.1109/ICSCDS65426.2025.11167489.

[23] O. Smovzhenko and A. Pysarenko, "Vision-based neighbor selection method for occlusion-resilient uncrewed aerial vehicle swarm coordination in three-dimensional environments," *Information, Computing and Intelligent systems*, no. 6, pp. 100–117, Sep. 2025, doi: 10.20535/2786-8729.6.2025.331602.
[24] W. Xu, X. Du, R. Li, B. Li, Y. Jiao, and L. Xing, "Attention-enhanced StrongSORT for robust vehicle tracking in complex environments," *Scientific Reports*, vol. 15, no. 1, May 2025, doi: 10.1038/s41598-025-99524-5.
[25] Y. Wang, H. Huang, J. He, D. Han, and Z. Zhao, "Closed-loop aerial tracking with dynamic detection-tracking coordination," *Drones*, vol. 9, no. 7, Jun. 2025, doi: 10.3390/drones9070467.

BIOGRAPHIES OF AUTHORS

Vidhya Kandasamy is an associate professor at the School of Computer Science and Technology, Karunya Institute of Technology and Sciences, Coimbatore, India. Her research interests include image processing and intelligent vision systems. She can be contacted at email: vidhyak@karunya.edu.

Antony Taurshia is an assistant professor at Karunya Institute of Technology and Sciences, India. Her research interests include cybersecurity and AI-enabled surveillance. She can be contacted at email: antonytaurshia@karunya.edu.

Thavittupalayam M. Thiyagu is an associate professor at the Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, India. His research interests include cybersecurity and intelligent systems. He can be contacted at email: t.m.thiyagu@gmail.com.

Catherine Joy RusselRaj is an assistant professor at Karunya Institute of Technology and Sciences, India. Her research focuses on artificial intelligence and image processing. She can be contacted at email: catherinejoy@karunya.edu.

Jenefa Archpaul an associate professor at Karunya Institute of Technology and Sciences, she brings expertise in artificial intelligence and image processing to her teaching and research, following the completion of her Ph.D. at Anna University in 2022. She can be contacted at email: jenefaa@karunya.edu.