Journal Pre-proof


# Hybrid-YOLO: Lightweight Mamba-Transformer Hybrid with Multi-Scale Fusion for Real-World Traffic Detection

Hongqing Wang , JunKit Chaw , Marizuana Mat Daud , Liantao Shi , Nannan Huang , Tin Tin Ting , Liuzhen Pu

PII: S2405-9595(25)00131-6
DOI: https://doi.org/10.1016/j.icte.2025.09.002
Reference: ICTE 903

To appear in: *ICT Express*

Received date: 10 July 2025
Revised date: 20 August 2025
Accepted date: 5 September 2025

Please cite this article as: Hongqing Wang , JunKit Chaw , Marizuana Mat Daud , Liantao Shi , Nannan Huang , Tin Tin Ting , Liuzhen Pu , Hybrid-YOLO: Lightweight Mamba-Transformer Hybrid with Multi-Scale Fusion for Real-World Traffic Detection, *ICT Express* (2025), doi: https://doi.org/10.1016/j.icte.2025.09.002

The Korean Institute of Communications and Information Sciences

ScienceDirect

ICT Express * (2025)

www.elsevier.com/locate/icte

# Hybrid-YOLO: Lightweight Mamba-Transformer Hybrid with Multi-Scale Fusion for Real-World Traffic Detection

Hongqing Wang 1, JunKit Chaw 1*, Marizuana Mat Daud 1, Liantao Shi 2*, Nannan Huang 3, Tin Tin Ting 4 and Liuzhen Pu 5

*Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Bangi, Selangor 43600, Malaysia*
*Institute for Carbon-Neutral Technology, Shenzhen Polytechnic University, ShenZhen, 51800, China*
*Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor 43600, Malaysia*
*Faculty of Data Science and Information Technology, INTI International University, 71800 Nilai, Negeri Sembilan, Malaysia*
*Faculty of Information Technology and Engineering, Ningde Vocational and Technical College, Ningde 355000, China*

**Abstract**

Vehicle detection in complex traffic scenes remains challenging due to frequent occlusions, lighting variations, and extreme weather. We present Hybrid-YOLO, a real-time detection framework that unifies Mamba-based state space modeling, Transformer-driven global attention, and multi-scale feature fusion to achieve high accuracy at low computational cost. At its core, Hybrid-YOLO introduces a Dynamic Residual Stem (DR Stem) for adaptive feature calibration, a Hexa-Scan Selective Block (HSSBlock) for six-directional structural perception, and a Selective State Space Model (SSM) for efficient long-range dependency modeling. A Cross-Stage Scales Feature Extraction (CSSFE) module enriches spatial semantics for small-object detection, while a Sparse-Queries Cascade Self-Attention (SCS) module focuses computation on informative regions, enhancing robustness to clutter and background noise. Extensive experiments on KITTI, BDD100K, and IITM-HeTra show that Hybrid-YOLO achieves 90.11 mAP@0.5 at 66.3 FPS, surpassing state-of-the-art methods in both accuracy and efficiency, and offering a promising solution for real-world intelligent transportation systems.

*Keywords: Intelligent vehicle detection technology, YOLO, complex traffic environment, mamba, transformer, small object detection, Industry, Innovation and Infrastructure*

## 1. Introduction

The rapid growth in vehicle ownership has posed significant challenges for intelligent transportation systems (ITS), especially in complex urban scenarios where occlusion, poor lighting, and adverse weather impair detection accuracy. Traditional traffic monitoring methods relying on manual feature engineering were limited in capturing subtle environmental variations [1–3]. The appearance of convolutional neural networks (CNNs) has transformed computer vision tasks, including vehicle detection. Two-stage detectors (e.g., R-CNN, Faster R-CNN) achieve high accuracy through region suggestion and ROI refinement, but are computationally intensive [4] [5]. One-stage models (e.g., YOLOv1-YOLOv11) achieve real-time performance through direct classification and regression on feature maps, but often struggle with small or occluded objects [4] [16].

*Corresponding author
*E-mail addresses:* p120388@siswa.ukm.edu.my, chawjk@ukm.edu.my, marizuana.daud@ukm.edu.my, xiaoshi1108@outlook.com, P120598@siswa.ukm.edu.my, tintin.ting@newinti.edu.my and mainland0507@outlook.com.

https://doi.org/xxxx/xxxx

Recent approaches attempt to balance speed and accuracy by integrating lightweight structures and attention mechanisms. YOLOv7-Ghost introduces the Ghost module and BiFPN to improve efficiency [11], while YOLOv8 employs CSPNet and PAN-FPN for enhanced multiscale feature fusion [12]. However, handling fine-grained spatial dependencies and high-resolution images remains a challenge. YOLOv9 and YOLOv11x further improve detection performance through the use of GELAN and CSPv11, respectively, albeit at the cost of increased model complexity [4].

Transformers, particularly the Swin Transformer [13], have advanced vision applications through hierarchical self-attention mechanisms. RepViT [14] and EfficientViT [15] introduced mobile-friendly attention cascades, yet still suffer from high latency due to the quadratic complexity of self-attention. To address this issue, Mamba-based State Space Models (SSMs) have been developed, enabling linear-time sequence modeling while effectively capturing long-range dependencies [7]. However, achieving accurate detection under real-world constraints remains difficult.

To address the above issues, we propose Hybrid-YOLO, a lightweight detection framework that integrates the advantages of CNN, Transformer, and State Space Modeling. Our

approaches include.

1) Hybrid-YOLO is composed of three parts, backbone, neck and detection head. In this study, CSSFE module and HSSBlock are integrated into the backbone to enhance feature extraction. The incorporation of the state space mechanism enables dynamic focus on essential information while mitigating redundancy. The SCS module receives multi-scale inputs from the backbone and constructs the global fusion network as illustrated in the Fig. 1. The resulting feature maps are then forwarded to the PAFPN neck, where the HSSBlock is employed once again for advanced feature fusion [6]. Finally, the refined features are passed through a decoupled detection head to generate the output.
2) In the proposed Hybrid-YOLO, each module is carefully designed and interconnected to address three key challenges in complex traffic scenes: occlusion, extreme weather, and lighting changes. The Dynamic Residual Stem (DR Stem) serves as the basic feature adjustment component, stabilizing gradient flow and suppressing low-level noise by adaptively balancing the residual path and identity path. This component is embedded in both the HSSBlock and the SCS modules, ensuring consistent feature quality from spatial feature extraction to global feature fusion. The HSSBlock combines the DR Stem with a six-direction scanning strategy (horizontal, vertical, and diagonal) to enhance spatial structural perception capabilities, enabling the model to maintain sensitivity to target contours even in scenarios with partial occlusion or complex backgrounds. These directional perception features are further processed by the Selective SSM, which dynamically updates the hidden matrix based on both historical and current inputs to efficiently capture long-range dependencies, thereby maintaining contextual consistency in scenarios such as strong lighting variations. The CSSFE module enriches spatial semantic information through hierarchical pooling and multi-scale feature fusion, improving the detection capability of small and distant targets under adverse conditions.
3) The SCS module again utilizes the DR Stem for stable feature preprocessing and focuses on the most informative regions through a sparse query attention mechanism, thereby reducing computational overhead while enhancing robustness against lighting and environmental interference. Within the overall architecture, the DR Stem ensures stable feature quality, the HSSBlock and CSSFE enhance spatial representation, the SSM preserves global temporal consistency, and the SCS achieves efficient global attention. These modules work collaboratively to mitigate the adverse effects of occlusion, extreme weather, and lighting changes on detection accuracy.

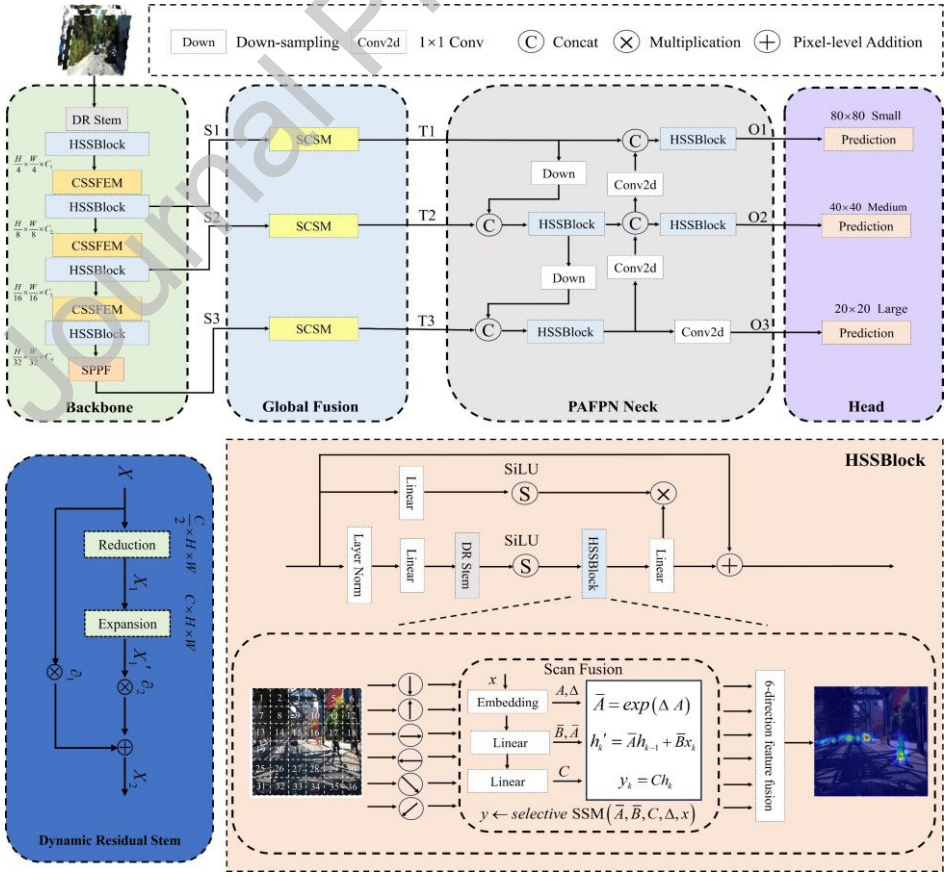


Fig. 1. Architecture of the Hybrid-YOLO.

## 2. Methods

### *2.1 Overall Architecture*

The overall architecture of Hybrid YOLO comprises four main components: a multi-scale fusion Mamba backbone (CSSFEM-HSSBlock), a global fusion module for modeling long-range dependencies, a PAFPN-based neck, and a decoupled detection head, as illustrated in Fig. 1.

The HSSBlock processing flow is divided into an upper branch and a bottom branch, with the core module of HSSBlock included in the bottom branch. The upper branch takes input feature $S \in \mathbb{R}^{H\times W\times C}$, performs a point-by-point linear mapping (1×1 convolution $W_1(\cdot)$) directly, and obtains $S' = \phi\left(W_1 * S\right) \in \mathbb{R}^{C\times H\times W}$ through SiLU activation $\phi(\cdot)$.

In the bottom layer, the input first is processed by layer normalization, and then a linear projection $W_2(\cdot)$, to obtain $\overline{S} = \left(W_2\left(LayerNorm\left(S\right)\right)\right)$. Then, it is input into the dynamic Residual Stem (DR Stem), where channel reduction and expansion are performed, and the residual and identity channels are adaptively fused through two gated weights to obtain $\overline{S}' = \partial_1 \cdot DR\left(\overline{S}, \omega\right) + \partial_2 \cdot \overline{S}$. The $DR(\cdot)$ includes a reduction channel and an expansion channel, both of which are composed of 1×1 convolutions. The $\omega$ is the correlation coefficient related to input $\overline{S}$. The $\partial_1$ and $\partial_2$ are adaptive weights balancing residual and identity pathways. DR Stem works as an adaptive feature calibration unit that can stabilize gradient flow, suppress low-level noise, and enhance task-related structures before directional scanning.

The core processing flow of the HSSBlock is as follows. First, perform a six-direction scan (including horizontal, vertical, and diagonal directions, with each direction containing forward and reverse sequences) $d \in \{1, 2, ..., 6\}$. utilizing selective scanning operations, unfold the feature map into a token sequence, as shown in the following equation.

$$x^d = Scan_d\left(\overline{S}'\right) \in \mathbb{R}^{L_d\times C} \qquad (1)$$

Where $L_d$ is the sequence length in direction d. In each direction, a discrete selective state space model (SSM) is applied for feature updating, as shown below.

$$h_k^d = \overline{A}^d h_{k-1}^d + \overline{B}^d x_k^d \qquad (2)$$

$$y_k^d = \overline{C}^d h_k^d \qquad (3)$$

Where $x_k^d$ denotes the k-th input token in direction d, $h_k^d$ is the hidden state, and $\overline{A}^d, \overline{B}^d, \overline{C}^d$ are the state, input and output matrices, respectively. Among them, $\overline{B}^d$ and $\overline{C}^d$ are dynamically generated via the selective scanning mechanism from the input features, enabling direction-specific feature enhancement. The obtained sequence $\left(y_k^d\right)_{k=1}^{L_d}$ is reshaped into a feature map form $F^d = reshape\left(\left(y_k^d\right)\right)_{k=1}^{L_d} \in \mathbb{R}^{H\times W\times C}$, and normalized via Layer Normalization $\overline{F}^d = LayerNorm\left(F^d\right)$. Normalized feature maps from all six directions are concatenated along the channel dimension [7].

$$F_{HSS} = Concat\left(\overline{F}^1, \overline{F}^2, ..., \overline{F}^6\right) \in \mathbb{R}^{H\times W\times 6C} \qquad (4)$$

Followed by a 1×1 convolutional reduction of the number of channels, the final output is $\overline{F}_{HSS} = Conv_{1\times1}\left(F_{HSS}\right) \in \mathbb{R}^{H\times W\times C}$. This design achieves parallel modeling and dynamic enhancement of global features in six directions, thereby significantly improving the model's global receptive field and directional sensitivity while maintaining local details.

Feature fusion with the upper branch is achieved through pixel multiplication operations to obtain $S_{HSS} = \overline{F}_{HSS} \odot S'$. Finally, the residual connection is fused with the original input to obtain the final output $O_{HSS} = S_{HSS} \oplus S$.

In HSSBlock, the DR Stem stabilizes gradient flow, suppresses noise, and preserves fine-grained structural features beneficial for detecting small or occluded objects through a dual-channel gated adaptive balancing of the residual and identity channels, while maintaining efficiency with a lightweight compression–expansion design. The six-direction selective SSM models target contours and geometric structures across multiple directions, efficiently model long-range dependencies in linear time, enhance contextual consistency, suppress background interference, and improve boundary perception capabilities.

### *2.2 Cross-stage Scales Spatial Feature Extraction Module (CSSFEM)*

In traffic detection, small targets like vehicles and pedestrians occupy few pixels, limiting available detail and context. To address this, we propose a lightweight Cross-stage Scales Spatial Feature Extraction (CSSFE) module (Fig. 2), which replaces conventional convolutions in the backbone with multi-stage feature extraction. CSSFE enables efficient multi-scale representation learning and spatial modulation, allowing the model to remain lightweight while adapting to complex traffic scenarios. The specific equations for CSSFEM are presented below.

$$X' = Conv_1\left(\partial_1\left(Conv_3\left(X, \kappa\right)\right)\right) \oplus X \qquad (5)$$

Eq. 1, shows the processing flow of the extraction. It is

assumed that the input is $X \in \mathbb{R}^{H\times W\times C}$, where $Conv_1(\cdot)$ and $Conv_3(\cdot)$ denote the convolution of 1×1 and 3×3. The $\kappa$ is the correlation coefficients of it corresponding input. The $\partial_1$ represents GELU activate function.

The initial position of the CSSFE module is Layer normalization. Given input $X' \in \mathbb{R}^{H\times W\times C}$ is processed by layerNorm, and then the number of channels is equally divided into four parts, $X_1, X_2, X_3$ and $X_4$ .As shown in Fig. 2, the feature outputs corresponding to four different colors and they all have a spatial dimension of $X_i \in \mathbb{R}^{H\times W\times C/4}$. In the CSSFE module, the processing sequence is structured from bottom to top. The first feature map remains unaltered, serving as an original representation $X_1 \in \mathbb{R}^{H\times W\times C/4}$. The second feature map processes by a 2× max pooling operation to produce $\hat{X}_2 \in \mathbb{R}^{H\times W\times C/4}$, effectively reducing its spatial dimensions while retaining salient features. Similarly, the third feature map is subjected to a 4× max pooling, resulting in $\hat{X}_3 \in \mathbb{R}^{H\times W\times C/4}$ and the fourth feature map is processed with a 6× max pooling to yield $\hat{X}_4 \in \mathbb{R}^{H\times W\times C/4}$. This hierarchical pooling strategy enables the extraction of multi-scale features. The specific formula is shown in Eq. 2, where $i \in (2,3,4)$.

$$maxpool(X_i) \in \mathbb{R}^{H/2^{(i-1)}\times W/2^{(i-1)}\times C/4} \quad (6)$$

$$X_i' = Up\left(Conv_3\left(\hat{X}_i, \rho\right)\right) \quad (7)$$

$$X_m = Conv_1\left(Concat\left(X_1, X_2, X_3, X_4\right)\right) \quad (8)$$

$$X_{out} = \partial_1\left(X_m\right) \otimes X' \quad (9)$$

The Eq. 3 used a 3×3 convolution for feature fusion and sequentially recovered the resolution using upsampling. where $\rho$ denotes the correlation coefficient and $Up(\bullet)$ denotes the upsampling. Multi-scale feature maps are concatenated and processed with a 1×1 convolution to integrate contextual

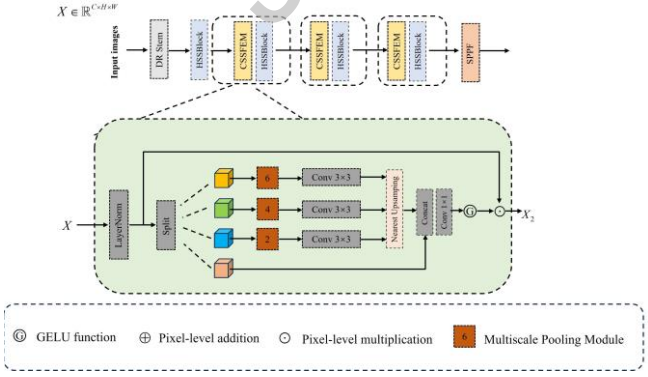


Fig. 2. Architecture of the CSSFEM.

information. This produces a global attention map that adaptively weights the original features, enhancing focus on key regions while preserving spatial structure. The CSSFEM and mamba-based HSSBlock form our backbone.

### *2.3 Sparse-queries Cascade Self-Attention Module (SCSM)*

To efficiently model global context while minimizing computational burden, the SCSM module introduces sparse self-attention based on BiFormer. It filters irrelevant Keys and Values, focusing on informative regions, thereby enhancing semantic understanding in complex traffic scenarios with minimal overhead [8].

First, the input feature map $I$ is processed through two dynamic residual (DR) connections.

$$I_1 = \partial_1 \cdot DR\left(I, \omega\right) + \partial_2 \cdot I \quad (10)$$

This improves gradient flow and model robustness. After performing the same DR Stem operation again, we obtain $I_2$.

The feature input $I_2$ is output via layer normalization $LN(\bullet)$ to obtain $I_2'$, which is fed into Bi-wise routing index attention. As can be seen from the left-hand side in Fig. 3, it divides the input feature map $I_2' \in \mathbb{R}^{H\times W\times C}$ into $K^2$ patches, each of which is linearly embedded into a one-dimensional feature vector with a feature dimension $HW/K^2$. Integrating each linearized patch, we obtain the token matrix $I_3 \in \mathbb{R}^{K^2\times HW/K^2\times C}$. Passing through the three linear layers $W_Q$, $W_K$ and $W_V$, respectively, we obtain the three matrices Q, K and V . The specific calculation equations are as follows.

$$Q = I_3 \cdot W_Q, K = I_3 \cdot W_K, V = I_3 \cdot W_V \quad (11)$$

A region-level attention map is constructed to evaluate inter-region relevance. Coarse-grained pruning retains the top-t most relevant regions based on attention scores, reducing computation. Fine-grained routing then indexes the corresponding K and V vectors. The specific calculation equations are as follows.

$$K' = Index(K, I_{rim}), V' = Index(V, I_{rim}) \quad (12)$$

Where $I_{rim}$ represents the path index matrix, the final sparse self-attention is calculated as follows:

$$Attention(Q, K', V') = Softmax\left(\frac{QK'^T}{\sqrt{d}}\right)V' + V_l \quad (13)$$

The $V_l$ is a local context enhancement term used to compensate for important information that may be lost in "non-attention areas" by sparse attention. The SCS module acts as a global fusion component, enhancing sensitivity to small targets

while reducing resource usage via sparse query attention.

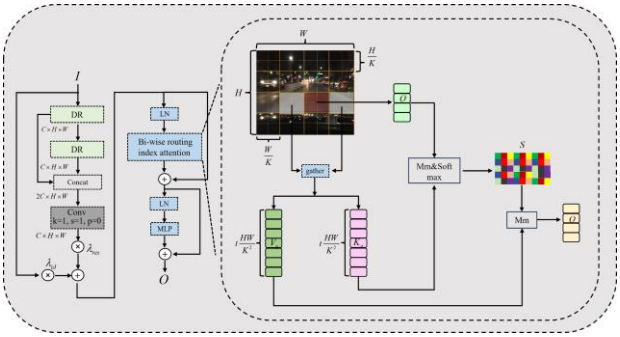


Fig. 3. Architecture of the SCSM.

## 3. Experimental

### *3.1 Experimental platforms and datasets*

All experiments were conducted on an NVIDIA RTX 4090 GPU (24 GB) using PyTorch 2.0.1, CUDA 11.8, and Python 3.9. The model was trained for 100 epochs with a batch size of 4, an input resolution of 640×640, momentum of 0.937, an SGD optimizer, and a learning rate that decayed from 0.01 to 0.0002. The reported 66.3 FPS was obtained from inference on the same hardware, with all images resized to 640×640, an inference batch size of 16, and only forward inference (excluding data loading, preprocessing, and postprocessing). The detection head adopts an NMS-free design, eliminating the traditional non-maximum suppression step, which keeps FPS stable across different scene complexities and significantly improves speed in dense scenes. To ensure robustness in different traffic scenarios, a mixed dataset containing 4,495 images was constructed, including 2,190 images from BDD-100K (covering various weather, lighting, and nighttime conditions), 1,663 images from KITTI (urban vehicle and pedestrian scenes), and 642 images from IITM-hetra [17] [18] [19].

### *3.2 Loss fuction and evaluation metrics*

To improve positioning accuracy in complex traffic scenarios, Hybrid-YOLO employs the MPDIoU loss function for bounding box regression. This function is based on the CIoU loss and incorporates the center distance, shape similarity, and aspect ratio differences between the predicted and ground-truth bounding boxes [9]. The specific equation is as follows.

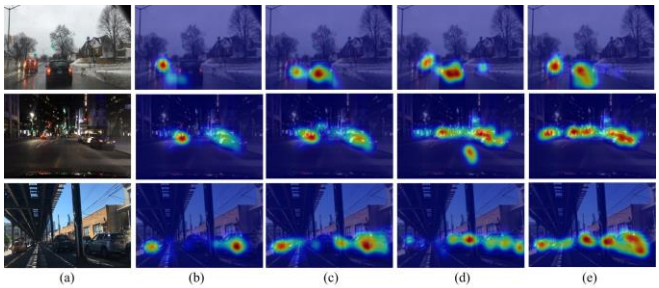


Fig. 6 Heatmap showing the effect on the BDD-100K dataset before and after model.

$$L_{MPDIoU} = 1 - IoU + \frac{\rho^2}{c^2} + \alpha v \qquad (14)$$

where $\rho$ is the center point distance between predicted and ground truth boxes, $c$ is the diagonal of the smallest enclosing box, and $v$ is the aspect ratio penalty term weighted by $\alpha$. This loss function enhances robustness in detecting small and occluded targets.

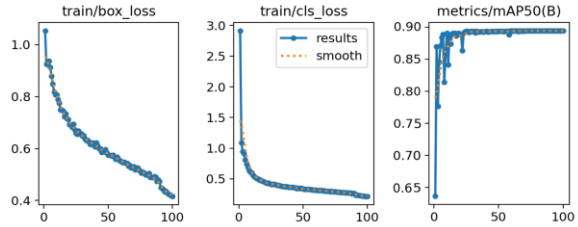


Fig. 4 Loss function and mAP@50 on mixed datasets.

Performance is evaluated using precision $\left(Pre = TP / TP + FP\right)$, recall $\left(Re = TP / TP + FN\right)$, mAP@0.5, and FPS. The mAP is the mean of category-wise average precision under IoU ≥ 0.5, while FPS measures real-time detection speed as $FPS = Images / Times$. These metrics jointly evaluate the accuracy and efficiency of the model.

Fig. 4 illustrate the convergence of loss functions and key metrics. By the 100th epoch, losses stabilize near their minimum and mAP approaches 90%, indicating effective optimization in localization and classification. The smooth curves also reflect good generalization performance.

### *3.3 Experimental results on the mixed dataset*

To verify the advantages of Hybrid-YOLO, we conducted comparative experiments on a mixed dataset combining BDD-100K and KITTI. Fig. 5 illustrates detection performance under complex conditions. In nighttime scenes (Fig. 5a–b), the model effectively handles blur, noise, and glare, successfully identifying small distant targets. Under adverse weather and alternating lighting (Fig. 5c–d), it maintains high detection accuracy despite interference from raindrops and strong illumination.

GradCAM heatmap visualization (Fig. 6) illustrates the differences in attention distribution among various models in complex traffic scenarios. (a) The original image shows the distribution of targets and backgrounds, along with common challenges in object detection, such as obstacle occlusion, light interference, and extreme weather conditions. (b) YOLOv7 displays scattered focus areas, with a substantial portion of attention directed toward background regions, leading to inadequate capture of details for small or partially occluded vehicles. (c) YOLOv8 achieves improved target localization but still allocates attention to irrelevant areas, such as road surfaces and light spots, resulting in incomplete target coverage. (d) YOLOv11 presents a relatively balanced attention distribution but tends to overlook distant small targets and fails to capture target contours effectively in crowded backgrounds. In contrast,

(e) Hybrid-YOLO produces dense attention regions with precise localization, enabling accurate detection of small, distant, and partially occluded vehicles while effectively suppressing background interference, thereby demonstrating superior robustness and recognition capability.

Quantitative results (Table 1) show Hybrid-YOLO achieves the highest mAP@0.5 of 90.11% and a real-time speed of 66.3 FPS, outperforming YOLOv11x (84.73%, 60.1 FPS) and RepViT (87.11%, 67 FPS). Although it has 55.3M parameters, the performance gain justifies the computational cost, making it suitable for real-world applications.

To further evaluate detection accuracy, we tested Hybrid-YOLO on the KITTI dataset using the standard three-level difficulty classification (easy, medium, and hard). As shown in Table 2, Hybrid-YOLO performed exceptionally well, achieving an mAP@0.5 of 88.05%, which is 3.68% and 0.73% higher than YOLOv11x and RepViT, respectively. Specifically, the model achieved an accuracy rate of 91.99% in the "person" category.

As shown in Fig. 7, Hybrid-YOLO achieves accurate detection on the KITTI dataset under various real-world interferences. (a) With partial occlusion from roadside obstacles, the model outputs bounding boxes closely matching target contours. (b) Under strong light and shadow, it maintains focus on vehicles and pedestrians without being misled by overexposure. (c) In complex backgrounds with shadows, it still detects small, distant vehicles precisely. (d) In heavily overlapping scenes with environmental distractions, Hybrid-YOLO effectively distinguishes individual objects, benefiting from the HSSBlock and CSSFE modules. These results demonstrate the model's robustness to occlusion, lighting interference, and background complexity.

Fig. 8 demonstrates the detection results of Hybrid-YOLO on the IITM-hetra dataset, showing its performance across diverse traffic targets. The results indicate that the model exhibits strong adaptability when dealing with multiple object categories and varying environmental complexities. In densely populated traffic scenes with significant inter-class scale variations, Hybrid-YOLO is able to achieve precise localization even when targets are partially occluded or located in cluttered backgrounds, highlighting its advantage in capturing fine-grained structural details. Table 3 compares the performance of Hybrid-YOLO with current state-of-the-art one-stage and two-stage detectors on the IITM-hetra dataset. Hybrid-YOLO achieves the highest mAP@0.5 of 81.34%, significantly outperforming YOLOv11x (77.35%) and Libra R-CNN (77.62%). This improvement is particularly notable in complex traffic scenarios, confirming the effectiveness of integrating the Mamba-based HSSBlock, CSSFE, and SCS modules, which collectively enable state space directional encoding, multi-scale semantic fusion, and efficient global context modeling.


Fig. 5 Hybrid-YOLO visualization results on dataset BDD-100k, where purple boxes represent small vehicles, green represents pedestrians, and dark yellow is for trucks.


Fig. 7 Hybrid-YOLO visualization results on dataset KITTI.

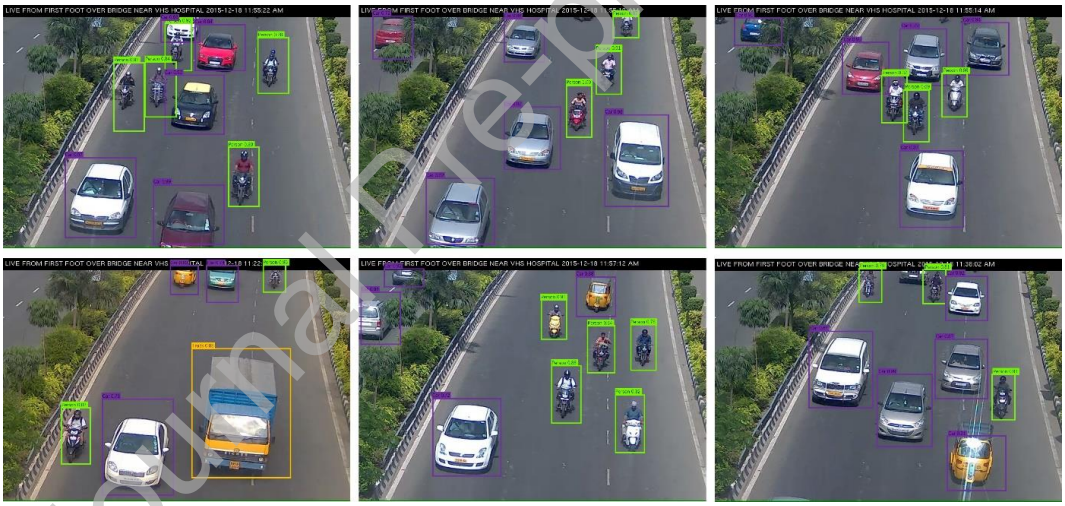

Fig. 8 Hybrid-YOLO visualization results on dataset IITM-hetra.



### *3.4 Ablation Study on Attention Modules*

To perform a more comprehensive evaluation of our proposed model, we conducted ablation experiments to assess the effectiveness of the SCS module. Specifically, we compared it against state-of-the-art lightweight global attention mechanisms, including BiFormer, CrossFormer [22], and ScalableViT [23]. As shown in Table 4, the Hybrid-YOLO model equipped with SCS achieved the highest detection accuracy (mAP@0.5: 90.11%) while maintaining real-time inference speed (66.3 FPS). Compared to CrossFormer and ScalableViT, SCS outperforms both in accuracy and speed, demonstrating its robustness and applicability in complex traffic environments. These results confirm that the sparse cascaded design of the SCS module strikes a better balance between detection accuracy and computational overhead, further underscoring its advantages in real-world traffic scenarios.

### *3.5 Visualization of Model Robustness under Challenging Conditions*

We present heatmaps generated by GradCAM for representative scenes (Fig. 9), including nighttime rain, strong illumination with specular highlights, and snow-occluded environments. Column (a) shows the original detection results, while columns (b) to (e) respectively present the attention maps of YOLOv7, YOLOv8, YOLOv11, and Hybrid-YOLO. All inputs are 640×640, consistent with the training and inference settings; heatmaps are computed on the final high-level feature before the decoupled head and overlaid after min–max normalization to ensure comparability.

YOLOv7 and YOLOv8 tend to focus on road reflections, water streaks, and snow surfaces, which leads to small targets being easily missed. YOLOv11 exhibits reduced background activation, but its attention remains scattered across large bright areas. Hybrid-YOLO is able to focus on taillights and vehicle contours while suppressing reflections; however, extremely bright specular spots still attract part of the attention, occasionally resulting in missed detections of distant vehicles.

YOLOv7 and YOLOv8 tend to over-focus on shadow boundaries, which leads to missed detections of small targets. YOLOv11 shows improvement by reducing such background bias, but its attention is still biased toward large objects, resulting in the omission of smaller ones. In contrast, Hybrid-YOLO maintains continuous attention along object contours at shadow transitions and can detect small targets, although the amount of attention allocated to these targets remains limited.

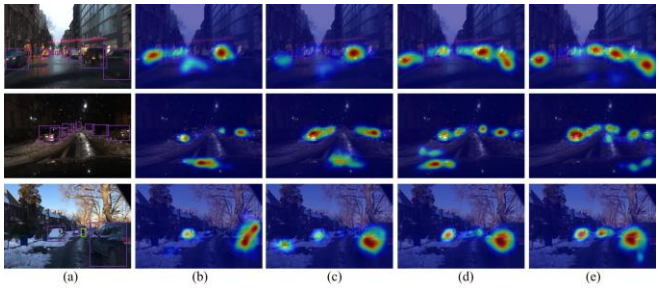


Fig. 9 Qualitative analysis based on heat maps under challenging conditions.

## 4. Discussion

Experimental results demonstrate that Hybrid-YOLO exhibits exceptional robustness in complex and challenging traffic scenarios across the BDD-100K, KITTI, and IITM-hetra datasets, including severe occlusions, lighting variations, and extreme weather conditions. Qualitative visualization results (Figs. 5–8) show that the model can accurately detect small, distant, and partially occluded objects while effectively suppressing background interference. For example, in KITTI scenes, Hybrid-YOLO maintains precise localization even in the presence of strong light-shadow contrast or densely overlapping targets, with bounding boxes closely matching target contours. Similarly, in nighttime scenes on the BDD-100K dataset, the model effectively mitigates glare, motion blur, and sensor noise, demonstrating adaptability to low-light environments.

From a quantitative perspective, Hybrid-YOLO outperforms current state-of-the-art single-stage and two-stage detectors across all difficulty levels, with particularly notable performance improvements in the “difficult” scenes of KITTI and the dense heterogeneous traffic scenes of IITM-hetra. This robustness stems from the synergistic interaction of its core components: the DR Stem module stabilizes gradient flow and suppresses low-level noise; the HSSBlock's six-direction selective SSM enhances structural perception in occluded and cluttered background scenarios; the CSSFE module enriches multi-scale semantic representations to improve small object detection performance; and the SCS module focuses attention on information-rich regions, thereby reducing interference from extreme weather and lighting conditions.

These findings align with the broader perspective of building resilient AI perception systems. Deng et al. emphasize that addressing environmental perturbations and adversarial conditions is essential for achieving trustworthy deployment [20]. The same authors also highlight the importance of integrating architectural innovations, diverse data exposure, and targeted evaluation strategies to enhance robustness [21]. The design of Hybrid-YOLO embraces these principles, combining multi-directional state space modeling, hierarchical feature fusion, and an efficient global attention mechanism. This enables the framework to deliver stable and reliable detection performance under highly variable and adverse real-world conditions, making it a promising candidate for deployment in intelligent transportation systems.

## Conclusion

In this work, we introduced Hybrid-YOLO, a lightweight yet high-performance framework for vehicle detection in complex traffic environments. By combining Mamba-based state space modeling, Transformer-driven global attention, and multi-scale feature fusion, the proposed design achieves an effective balance between accuracy and efficiency. The integration of the HSSBlock, CSSFE, and SCS modules enables robust detection of small, distant, and occluded targets under challenging conditions.

Future work will focus on further reducing computational overhead through model pruning and architectural optimization, as well as extending the framework to handle increasingly complex and dynamic intelligent transportation scenarios.

Table 1 Hybrid-YOLO was compared with the current state-of-the-art model on the mixed dataset

| **Method** | **Backbone** | **mAP@0.5/%** | **Params/$10^6$** | **FPS(f/s)** |
|---|---|---|---|---|
| Faster-RCNN | ResNet-50-FPN | 72.23 | 41.5 | 33 |
| Libra R-CNN | ResNet | 77.32 | 78.44 | 60 |
| VanillaNet [11] | EfficientRep | 76.00 | 17.2 | 42 |
| EfficientViT | Cascaded Attention | 83.45 | 24.2 | 43.1 |
| Swin Transformer | Swin-T | 83.32 | 42.21 | 63 |
| RepViT | MobileNetV3-L | 87.11 | 14.0 | 67 |
| YOLOv6-m | EfficientRep | 78.32 | 34.9 | **93** |
| Ghost-YOLOv7 | Ghost-BiFPN | 77.88 | 32.5 | 69 |
| YOLOv8 | CSPNet | 82.22 | 43.7 | 34.2 |
| YOLOv9 | GELAN | 81.99 | 58.9 | 36.8 |
| YOLOv10 | Modified CSP v10 | 79.22 | **7.2** | 80.7 |
| YOLOv11x | C3k2 | 84.73 | 57.0 | 60.1 |
| **Ours** | Hybrid-YOLO | **90.11** | 55.3 | 66.3 |

Table 2 FLSA-YOLO was compared with the current state-of-the-art model on the dataset KITTI.

| **Method** | **Average precision@0.5/%** | | | | | | | | | **mAP@0.5/%** | **Params/$10^6$** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Car | | | Person | | | Cyclist | | | | |
| | E | M | H | E | M | H | E | M | H | | |
| Faster R-CNN | 84.31 | 82.51 | 77.62 | 53.11 | 50.02 | 49.23 | 49.58 | 51.11 | 52.12 | 61.06 | 41.5 |
| Libra R-CNN | 92.31 | 88.42 | 74.23 | 89.33 | 85.20 | 80.58 | 89.13 | 84.56 | 79.22 | 84.80 | 78.4 |
| YOLOv6 | 83.62 | 78.43 | 71.02 | 79.55 | 78.51 | 69.11 | 82.11 | 78.56 | 71.00 | 76.88 | 34.9 |
| Ghost-YOLOv7 | 79.63 | 74.56 | 74.22 | 81.56 | 87.62 | 80.00 | 79.43 | 71.29 | 69.41 | 77.52 | 32.5 |
| YOLOv9 | 83.62 | 86.31 | 75.24 | 81.24 | 78.45 | 77.66 | 76.41 | 74.55 | 75.23 | 78.75 | 43.7 |
| YOLOv11x | 88.11 | 83.62 | 84.42 | 88.31 | 76.54 | 71.33 | 81.56 | 80.66 | 75.52 | 81.13 | 57.0 |
| Swin-Transformer | 85.42 | 78.31 | 80.23 | 80.52 | 83.56 | 80.22 | 79.31 | 76.23 | 77.84 | 80.81 | 42.2 |
| EfficientViT | 88.49 | 86.32 | 82.58 | 87.46 | 82.26 | 82.77 | 81.25 | 80.41 | 76.62 | 83.17 | 24.2 |
| RepViT | 87.62 | 88.34 | 83.25 | 91.26 | 87.29 | 81.55 | **89.72** | 81.23 | 80.77 | 86.02 | **14.0** |
| **Ours** | **93.28** | **89.99** | **86.32** | **91.99** | **88.17** | **83.67** | 89.54 | **85.68** | **84.22** | **88.05** | 55.3 |

Table 3 Hybrid-YOLO was compared with the current state-of-the-art model on the IITM-hetra dataset

| Method | Backbone | Input size | mAP @0.5/% |
|---|---|---|---|
| **One-stage** | | | |
| YOLOv4 | Darknet53 | 640×640 | 68.23 |
| YOLOv6 | EfficientRep | 640×640 | 71.55 |
| Ghost-YOLOv7 | Ghost+BiFPN | 640×640 | 67.42 |
| YOLOv8l | CSPNet | 640×640 | 74.00 |
| YOLOv11x | C3k2 | 640×640 | 77.35 |
| **Two-stage** | | | |
| Faster R-CNN | ResNet | 640×640 | 63.54 |
| Mask R-CNN | ResNet | 640×640 | 69.38. |
| Libra R-CNN | ResNet | 640×640 | 77.62 |
| **Ours** | Hybrid-YOLO | 640×640 | **81.34** |

Table 4 Ablation study on attention modules evaluated on the mixed dataset.

| Method (Backbone + Attention) | mAP @0.5/% | FPS(f/s) | Params/$10^6$ |
|---|---|---|---|
| Hybrid-YOLO w/ CrossFormer [22] | 86.77 | 52.1 | 90.1 |
| Hybrid-YOLO w/ ScalableViT [23] | 88.33 | 43.2 | 74.6 |
| Hybrid-YOLO w/ BiFormer | 87.92 | **69.8** | **53.4** |
| **Hybrid-YOLO w/ SCS (Ours)** | **90.11** | 66.3 | 55.3 |

## Conflict of interest

The authors declare that there is no conflict of interest in this paper.

## References


[1] Kamal, M. A. S., Hayakawa, T., & Imura, J. (2019). Development and evaluation of an adaptive traffic signal control scheme under a mixed-automated traffic scenario. IEEE Transactions on Intelligent Transportation Systems, 21(2), 590–602. Doi: 10.1109/TITS.2019.2896943.

[2] Jing, R., Zhang, W., Liu, Y., Li, W., Li, Y., & Liu, C. (2024). An effective method for small object detection in low-resolution images. Engineering Applications of Artificial Intelligence, 127, 107206. Doi: 10.1016/j.engappai.2023.107206.

[3] Liu R, Wu J, Lu W, et al. A Review of Deep Learning-Based Methods for Road Extraction from High-Resolution Remote Sensing Images[J]. Remote Sensing, 2024, 16(12): 2056. Doi: 10.3390/rs16122056.

[4] Wang C Y, Liao H Y M. YOLOv1 to YOLOv10: The fastest and most accurate real-time object detection systems[J], 2024. [Online] Available : http://arxiv.org/abs/2408.09332.

[5] Tian J, Jin Q, Wang Y, et al. Performance analysis of deep learning-based object detection algorithms on COCO benchmark: a comparative study[J]. Journal of Engineering and Applied Science, 2024, 71(1): 76. Doi: 10.1186/s44147-024-00411-z.

[6] Jocher, G., Chaurasia, A. and Qiu, J. (2023) YOLO by Ultralytics. https://github.com/ultralytics/ultralytics.

[7] Wang Z, Li C, Xu H, et al. Mamba YOLO: SSMs-based YOLO for object detection[J], 2024. [Online]. Available: http://arxiv.org/abs/2406.05835.

[8] Zhu L, Wang X, Ke Z, et al. Biformer: Vision transformer with bi-level routing attention[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 10323-10333. [Online]. Available: http://arxiv.org/abs2303.08810

[9] Ma, Siliang, and Yong Xu. "Mpdiou: a loss for efficient and accurate bounding box regression." arxiv preprint arxiv:2307.07662 (2023). Doi: 10.48550/arxiv.2307.07662.

[10] H. Chen, Y. Wang, J. Guo, and D. Tao, "VanillaNet: the Power of Minimalism in Deep Learning." arXiv, May 23, 2023. Accessed: Jan. 14, 2024. [Online]. Available: http://arxiv.org/abs/2305.12972.

[11] Shi Q, Zhong F, Li B, et al. Fast vehicle detection algorithm based on lightweight YOLO7-tiny[C]//Fifth International Conference on Computer Vision and Data Mining (ICCVDM 2024). SPIE, 2024, 13272: 50-58. [Online]. Available: http://arxiv.org/abs/2304.06002v3.

[12] R. Varghese and S. M., "YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness," *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, Chennai, India, 2024, pp. 1-6. Doi: 10.1109/ADICS58448.2024.10533619.

[13] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada: IEEE, Oct. 2021, pp. 9992–10002. Doi: 10.1109/ICCV48922.2021.00986.

[14] A. Wang, H. Chen, Z. Lin, J. Han, and G. Ding, "RepViT: Revisiting Mobile CNN From ViT Perspective." arXiv, Sep. 28, 2023. Accessed: Jan. 14, 2024. [Online]. Available: http://arxiv.org/abs/2307.09283.

[15] Liu X, Peng H, Zheng N, et al. Efficientvit: Memory efficient vision transformer with cascaded group attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 14420-14430. Doi: 10.48550/arxiv.2305.07027.

[16] X. Zhou, J. Kan, N. Fatin Liyana Mohd Rosely, X. Duan, J. Cai and Z. Zhou, "ILN-YOLOv8: A Lightweight Image Recognition Model for Crimped Wire Connectors," in *IEEE Access*, vol. 13, pp. 5193-5202, 2025. Doi: 10.1109/ACCESS.2025.3525564.

[17] Yu F, Xian W, Chen Y, et al. Bdd100k: A diverse driving video database with scalable annotation tooling[J]. [Online]. Available: http://doi.org/10.48550/arXiv.1805.04687.

[18] Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: The kitti dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237. Doi: 10.1177/0278364913491297.

[19] D. Mittal, A. Reddy, G. Ramadurai, K. Mitra, and B. Ravindran, "IITM-HeTra: Dataset for Vehicle Detection in Heterogeneous Traffic Scenarios," 2018. [Online]. Available:

https://www.kaggle.com/datasets/deepak242424/iitmhetra.
[20] Z. Deng, Y. Guo, C. Han, W. Ma, J. Xiong, S. Wen, and Y. Xiang, "AI Agents Under Threat: A Survey of Key Security Challenges and Future Pathways," 2024. [Online]. Available: https://arxiv.org/abs/2406.02630.
[21] Z. Deng, W. Ma, Q.-L. Han, W. Zhou, X. Zhu, S. Wen, and Y. H. Xiang, "Exploring DeepSeek: A Survey on Advances, Applications, Challenges and Future Directions," *IEEE/CAA Journal of Automatica Sinica, vol. 12, no. 5, pp. 872–893, May 2025. [Online]. Available: https://doi.org/10.1109/jas.2025.125498.
[22] Wenxiao Wang, Lu Yao, Long Chen, Binbin Lin, Deng Cai, Xiaofei He, and Wei Liu. Crossformer: A versatile vision transformer hinging on cross-scale attention. In International Conference on Learning Representations, ICLR, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2303.06908.
[23] Rui Yang, Hailong Ma, Jie Wu, Yansong Tang, Xuefeng Xiao, Min Zheng, and Xiu Li. Scalablevit: Rethinking the context-oriented generalization of vision transformer. [Online]. Available: https://doi.org/10.48550/arXiv.2203.10790.

**Authorship contribution statement**

Author contributions: Hongqing Wang: Writing – original draft, Visualization, Resources, Conceptualization; Junkit Chaw and Marizuana Mat Daud: Writing – review & editing, Supervision, Funding acquisition; Liantao Shi: Design of experiments and coding; Nannan Huang: Parameter tuning; Tin Tin Ting: Collection and organization of experimental datasets; Liuzhen Pu: Writing – review & editing.

**Acknowledgments**

This paper is supported by grant ZG-2025-018 from the Institute of Visual Informatics (IVI), Universiti Kebangsaan Malaysia (UKM).

## Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: