WH-Mamba: Wavelet-Enhanced State Space Model with Dynamic Hypergraph Fusion for Robust Traffic Object Detection

1st Huanrong Tang
Xiangtan University
Xiangtan, China
tanghuanrong@126.com

2nd Jiaxiong Lu
Xiangtan University
Xiangtan, China
jiaxionglu8@gmail.com

3rd Given Name Surname
Xiangtan University
Xiangtan, China
oyjq@xtu.edu.cn

***Abstract*—Real-time object detection in autonomous driving faces persistent challenges in balancing computational efficiency with perceptual robustness, particularly when confronting minuscule targets and adverse weather conditions. To address the limitations of existing CNN and Transformer architectures—specifically feature attenuation and quadratic complexity—we propose WH-Mamba, a novel framework synergizing State Space Models (SSMs) with advanced signal processing and graph theory. At the backbone level, we introduce the Wavelet-Cross-Mamba Block (WCMB), which integrates a Quad-Directional Scanning (QDS) strategy to align 1D selective scans with the radial topology of traffic scenes, while leveraging Discrete Wavelet Transforms to preserve high-frequency edge details often submerged in deep layers. To further mitigate environmental noise, a Frequency-Spatial Attention (FSA) module employs Fourier analysis for spectral denoising and spatial saliency sharpening. Finally, a Dynamic Hypergraph Fusion (DHF) module is designed to model complex "multi-to-multi" semantic correlations among diverse traffic agents, transcending the receptive field limits of local convolutions. Extensive evaluations on MS COCO and BDD100K demonstrate that WH-Mamba achieves 45.7% AP on COCO and 43.2% AP on BDD100K. Notably, it outperforms recent models such as YOLOv13 and the Transformer-based DEIM, showing a significant advantage in small object detection with an 30.1% AP for small objects.**

***Index Terms*—Real-time Object Detection, State Space Models, Hypergraph, Wavelet Transform, Autonomous Driving.**

I. Introduction

Driven by the accelerating pace of global urbanization and breakthroughs in artificial intelligence, Intelligent Transportation Systems (ITS) have become imperative infrastructure for enhancing road safety and efficiency. At the core of these systems lie autonomous vehicles, whose operational viability depends entirely on their ability to achieve precise, all-weather, and omnidirectional perception of complex environments. As highlighted in the recent World Models for Autonomous Driving Survey [1], the ability to predict future events and assess implications from sensor data is paramount. Among the fundamental computer vision tasks enabling this, object detection assumes paramount importance; accurately identifying diverse dynamic agents such as vehicles, pedestrians, and cyclists, along with static traffic signs, serves as the non-negotiable prerequisite for subsequent trajectory prediction and collision avoidance decision-making.

However, deploying these algorithms in real-world road scenarios presents formidable challenges that far exceed the controlled conditions of standard benchmarks. Traffic environments are characterized by extreme entropy: high-speed vehicular motion induces significant motion blur that degrades edge fidelity; distant, critical targets like traffic signs occupy minimal pixel areas, leading to the severe "small object problem" where semantic information is scarce [2]. Furthermore, adverse weather conditions—including rain, snow, and dense fog—introduce substantial high-frequency noise and contrast degradation that severely distort structural features [3], while complex urban backgrounds combined with dense traffic flows result in frequent inter-object occlusion and drastic variations in object scale.

Conventional detection architectures have historically struggled to balance these conflicting demands effectively. Since the pioneering work of YOLOv1 [4], Convolutional Neural Network (CNN)-based real-time detectors have evolved significantly. While recent iterations like YOLOv11 [5] and the attention-centric YOLOv12 [6] offer low latency, their theoretical Effective Receptive Fields (ERF) remain constrained by the locality of convolution operations. Even with the integration of Feature Pyramid Networks (FPN) [7], increasing network depth to capture global context often exacerbates the "depth versus detail" contradiction, where high-frequency cues essential for small objects are over-smoothed. Conversely, Transformer-based detectors, initiated by the seminal DEtection TRansformer (DETR) [8] and advanced by Swin Transformer [9], utilize self-attention for robust global modeling. However, despite improvements in convergence seen in models like DEIM [10], their quadratic computational complexity ($O(N^2)$) imposes prohibitive memory footprints when processing high-resolution inputs. Moreover, standard patch-based processing can disrupt the image's inherent 2D spatial structure.

To address these multifaceted limitations, this paper proposes a paradigm shift, re-examining traffic object detection through the complementary lenses of advanced signal pro-

cessing, State Space Models (SSMs), and graph theory. Our approach is driven by four synergistic motivations. First, leveraging the linear complexity of the Mamba architecture [11], we address the "spatial structure mismatch" of applying 1D sequencing to 2D images. Unlike recent orthogonal scanning approaches [12], we propose a Quad-Directional Scanning (QDS) strategy, explicitly adapting to the radial geometric topology of road scenes. Second, to tackle the "depth vs. detail" conflict, we introduce the Discrete Wavelet Transform (DWT) into the backbone, decoupling features into frequency bands to preserve high-frequency details. Third, recognizing that environmental noise and object textures are entangled, we propose a Joint Synergistic Enhancement mechanism via a Frequency-Spatial Attention (FSA) module; inspired by the Convolutional Block Attention Module (CBAM) [14] attention mechanism but extended to the spectral domain, we achieve "spectral denoising". Fourth, to capture complex non-local correlations beyond simple pairwise interactions, we introduce Hypergraph theory, drawing on the foundational Hypergraph Neural Networks (HGNN) [16], to construct adaptive hyperedges. This Dynamic Hypergraph Fusion (DHF) module clusters arbitrary nodes with similar semantics

Our contributions are summarized as follows:

- We propose the Wavelet-Cross-Mamba Block (WCMB), which utilizes the QDS strategy and dual-branch wavelet enhancement to solve the difficulties of radial feature extraction and noise interference in traffic scenarios.
- We design the FSA module, which exploits joint modeling of the frequency and spatial domains to effectively address the challenge of high-frequency detail loss for small objects in deep networks.
- We introduce the DHF module, which employs adaptive hypergraph computation to mine high-order global correlations, breaking through the locality limitations of traditional feature fusion.
- Experiments on the BDD100k and COCO datasets demonstrate that our method significantly outperforms mainstream models such as YOLOv13 and RT-DETR while maintaining real-time performance.

II. Related Work

A. Object Detection in Traffic Scenarios

Traffic object detection methodologies generally bifurcate into CNN-based real-time detectors and Transformer-based high-precision architectures. The YOLO series, originating from the seminal YOLOv1 [4], serves as the quintessential benchmark for real-time detection. The integration of Feature Pyramid Networks(FPN) [7] became a standard for multi-scale fusion. Recently, YOLOv10 [17] significantly reduced latency by introducing a Non-Maximum Suppression (NMS)-Free training strategy, while YOLOv11 [5] optimized feature extraction efficiency. YOLOv12 [6] further introduced an attention-centric design to maximize parameter utilization. Although YOLOv13 [18] attempts to mitigate locality limitations via hypergraph concepts, these models often struggle with minuscule targets in extreme scale variations.

Transformers leverage self-attention to model global context. The pioneering DETR [8] introduced an end-to-end paradigm but suffered from slow convergence. While RT-DETR [19] reconciled real-time constraints, recent advancements focus on regression refinement. D-Fine [20] enhances localization by redefining bounding box regression as a Fine-grained Distribution Refinement (FDR) task. DEIM [10] further accelerates convergence through a Dense One-to-One (Dense O2O) matching strategy. DEIMv2 [21] integrates the robust feature representations of DINOv3 [22] and employs a Spatial Tuning Adapter (STA) to optimize multi-scale feature alignment. Despite these benchmarks, the computational overhead for high-resolution processing remains a bottleneck for edge deployment.

B. Visual State Space Models and Scanning Strategies

SSMs have evolved significantly from control theory. The S4 model [23] mitigated long-sequence memory bottlenecks via structured parameterization, while the foundational Mamba [11] introduced a Selective Scan mechanism to achieve input-dependent dynamics with linear complexity. However, adapting Mamba to vision tasks faces a dimensionality mismatch. While Vision Mamba (Vim) [24] and VMamba [12] employ bidirectional or orthogonal Cross-Scan strategies, such axis-aligned approaches prove suboptimal for traffic scenarios characterized by radial perspective features. Recent works like LocalMamba [25] and Spatial-Mamba [26] highlight the need for structure-aware scanning. Inspired by these explorations, this paper proposes the Quad-Directional Scanning strategy, tailored to the geometric topology of road environments.

C. Frequency Domain Learning in Computer Vision

Frequency Domain Learning leverages transformations such as the Fast Fourier Transform (FFT) and Wavelet Transform (WT) to project spatial features into the spectral domain. While the Fourier Transform captures global distributions, the DWT is distinguished by its superior time-frequency localization. Early works like Wavelet Integrated CNNs [27] demonstrated the utility of wavelets for noise-robust classification. Recent studies like HRFAConv [13] and wavelet-based dehazing networks [15] have further proven the efficacy of wavelets in preserving high-frequency details during downsampling. In response, this paper introduces the WCMB and the FSA module, leveraging high-frequency wavelet components to accentuate the edge details of minuscule targets, realizing a mechanism of "global denoising" and "local sharpening."

D. Hypergraph Neural Networks for Feature Fusion

In the feature fusion stage, conventional architectures rely predominantly on convolutional operations, which restricts them to pairwise correlation modeling. Hypergraph Neural Networks (HGNNs) [16] transcend local limitations by introducing "hyperedges" capable of connecting an arbitrary number of nodes. Although widely applied in other domains, the integration of hypergraphs into real-time object detectors is a burgeoning frontier. Recent works like MVHFNet in remote

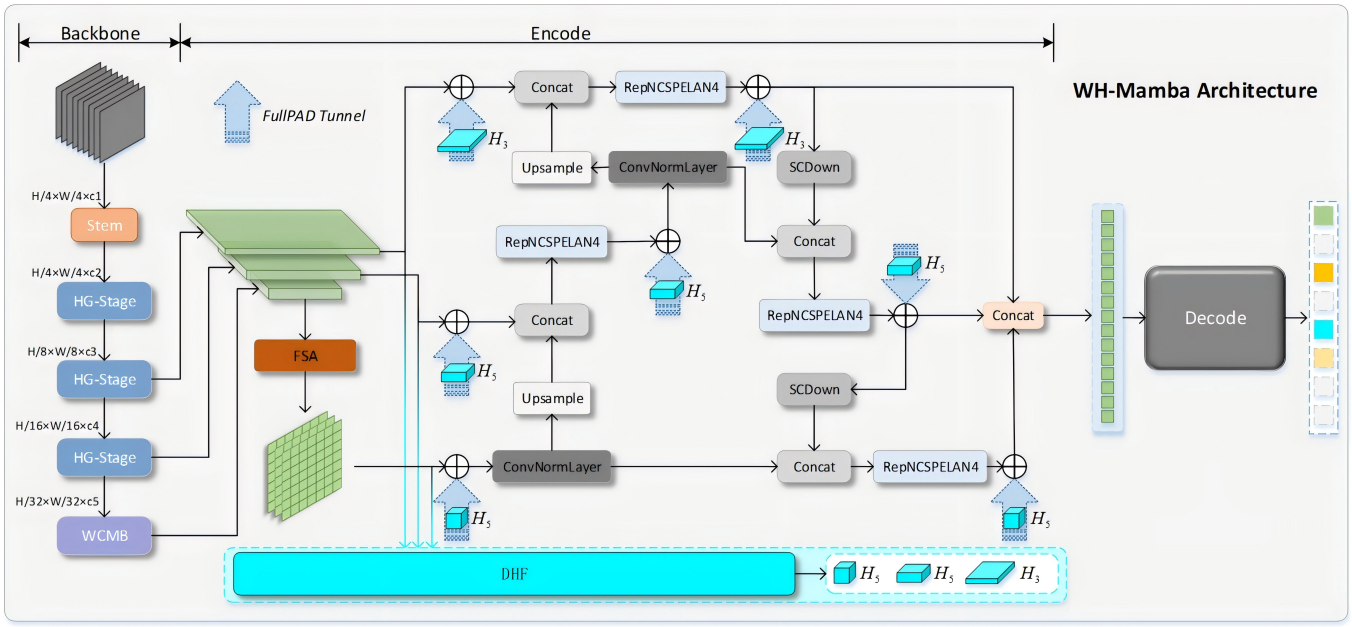

Fig. 1. The overall architecture of WH-Mamba. Utilizing a hybrid backbone, it transitions from early HG-Stages to a final WCMB for robust modeling. Key components include the FSA module for spectral-spatial refinement and the DHF for mining high-order semantics. These features are effectively distributed via the FullPAD Tunnel to facilitate object detection.

sensing and Hyper-YOLO in general detection have proven the effectiveness of hypergraph computation. We propose the DHF module. DHF synergistically combines a high-order semantic branch with a low-order detail branch, dynamically constructing hyperedges to mine cross-scale global semantics.

III. Method

A. Network Architecture

As illustrated in Fig. 1, the proposed WH-Mamba framework advances the state-of-the-art DEIM architecture. To overcome DEIM's limitations in detecting minuscule objects and handling weather-induced degradation in traffic scenarios, we retain its efficient end-to-end paradigm while fundamentally restructuring the feature extraction and fusion hierarchies. We adopt a hybrid "Early-Stage Convolution, Late-Stage Mamba" strategy for the backbone design. A pivotal architectural decision is the strategic deployment of the WCMB exclusively at the backbone's final stage, driven by two scenario-specific imperatives. First, to counteract "High-Frequency Submersion"—where deep networks prioritize low-frequency semantics at the expense of fine details—we integrate the DWT within the WCMB. By applying DWT at this semantic-rich depth, we effectively reconstruct critical edge textures of distant objects prior to fusion. Second, to achieve computationally efficient global modeling, we exploit the reduced resolution of the final stage. Implementing the QDS strategy here establishes long-range dependencies essential for capturing road topology with minimal memory overhead, thereby avoiding the prohibitive costs associated with early-stage global processing. Subsequently, features undergo dual-domain enhancement via the FSA module and cross-scale interaction via the DHF module, significantly augmenting perceptual robustness in dynamic environments while preserving linear complexity.

B. Wavelet-Cross-Mamba Block

To effectively mitigate challenges such as low pixel occupancy, motion blur, and adverse weather interference, we propose the WCMB, with its structure illustrated in Fig. 2. This module synergistically captures global dependencies and recovers high-frequency details through joint frequency-spatial modeling. Input features are bifurcated along the channel dimension: one stream undergoes Wavelet Transform Enhanced Cross-Mamba (WTEC-Mamba) for global modeling with edge enhancement, while the other utilizes Multi-Kernel Depthwise Convolution (MK-DeConv) for local perception.

WTEC-Mamba. Standard orthogonal scanning often disrupts the radial topology inherent in traffic scenes, fragmenting diagonal structures. To address this, we introduce the Cross-Mamba module equipped with a QDS strategy. By incorporating diagonal paths, QDS aligns state space modeling with physical trajectories to enhance semantic coherence. Structurally, the module employs a gated dual-branch architecture to encode global contexts. These are subsequently integrated with local spatial details via the Pixel-level Aggregation Fusion (PAF) module, which dynamically re-weights features to ensure effective multi-scale fusion. Concurrently, to counteract the high-frequency texture smoothing typical in deep networks, we integrate a DWT branch. Utilizing the Haar basis, the input feature x_I^w is decomposed into frequency sub-bands, where a selective filtering mechanism suppresses stochastic environmental noise while enhancing deterministic

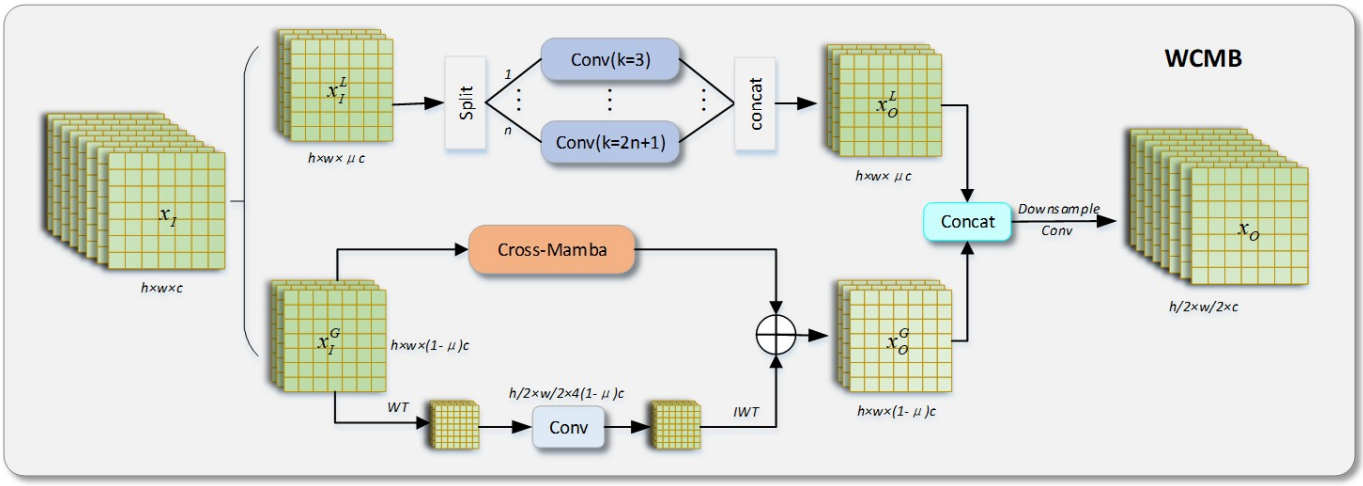

Fig. 2. Architecture of the WCMB. It splits the input features into two parts along the channel dimension. The upper branch processes features via Multi-Kernel Depthwise Convolution (MK-DeConv), while the lower branch employs the Wavelet Transform Enhanced Cross-Mamba (WTEC-Mamba).

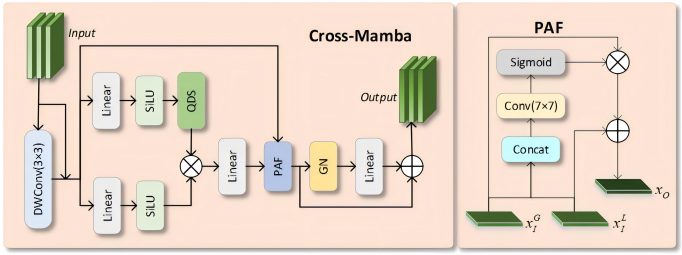

Fig. 3. The left panel illustrates the structure of the Cross-Mamba module. Input features encoded by a 3×3 DWConv are split into a QDS global branch and a SiLU gating branch. The right panel illustrates the structure of the PAF module. It performs dynamic fusion using 7×7 convolutional attention maps.

edge contours. This transformation and subsequent reconstruction are formulated as (1).

$$\begin{aligned} x_I^{wt} &= \mathrm{WT}\left(x_I^w, [f_{LL}, f_{LH}, f_{HL}, f_{HH}]\right), \\ x_O^w &= \mathrm{IWT}\left(\mathrm{Conv}(x_I^{wt}), [f_{LL}, f_{LH}, f_{HL}, f_{HH}]\right), \end{aligned} \tag{1}$$

where the refined features x_O^w are fused with the Mamba branch output x_O^G, ensuring a robust integration of global semantic context and local structural detail.

MK-DeConv. To complement the global modeling of the WTEC-Mamba branch, the feature partition $x_I^L \in \mathbb{R}^{h\times w\times \mu c}$ is dedicated to preserving fine-grained physical details in the spatial domain. We implement a MK-DeConv strategy, utilizing diverse kernel sizes to synthesize multi-scale receptive fields, as formulated in (2). This design enables the network to adaptively accommodate drastic scale variations—ranging from distant signs to nearby vehicles—inherent in dynamic traffic environments.

$$\begin{aligned} x_{Lj}^O &= \mathrm{Conv}(x_{Lj}^I, k=(2j+1)), \quad j\in\{1,\ldots,n\}, \\ x_L^O &= \mathrm{Concat}([x_{L1}^O,\ldots,x_{Ln}^O], \mathrm{dim}=-1), \end{aligned} \tag{2}$$

C. Frequency-Spatial Attention Module

Polished Text: Although Mamba architectures demonstrate superior global modeling capabilities, they frequently suffer from the attenuation of high-frequency details within deep layers, particularly when detecting small objects. To reconcile this trade-off, we introduce the FSA module. As illustrated in Fig. 4, the FSA integrates two parallel streams: a frequency domain branch specialized for enhancing specific spectral features and a spatial branch designed to localize salient regions within the pixel domain.

Frequency Domain Branch. This branch exploits the global receptive field of the Fourier Transform to capture long-range dependencies that are difficult to model spatially, while simultaneously isolating high-frequency cues essential for small object detection. Given input features $\hat{F}_i' \in \mathbb{R}^{H\times W\times C}$, we apply the Two-Dimensional Discrete Fourier Transform (2D DFT) to transition into the frequency domain, yielding the spectral feature f_i, as formulated in (3):

$$f_i(U,V) = \sum_{x=0}^{H-1}\sum_{y=0}^{W-1} \hat{F}_i'(x,y) e^{-j2\pi\left(\frac{Ux}{H}+\frac{Vy}{W}\right)}, \tag{3}$$

where (x, y) and (U, V) denote spatial and frequency coordinates, respectively. To decouple background context from object edge details, we employ binary masks M_{low} and M_{high}. The low-pass mask M_{low} isolates the central low-frequency region, whereas M_{high} operates inversely to extract high-frequency components representing structural edges, as defined in (4):

$$f_i^{low} = f_i \otimes M_{low}, \quad f_i^{high} = f_i \otimes M_{high}, \tag{4}$$

To handle complex traffic backgrounds, a learnable global filter $N \in \mathbb{R}^{H\times W\times C}$ is introduced to adaptively modulate the low-frequency component f_i^{low}. The modulated background is then recombined with the preserved high-frequency details

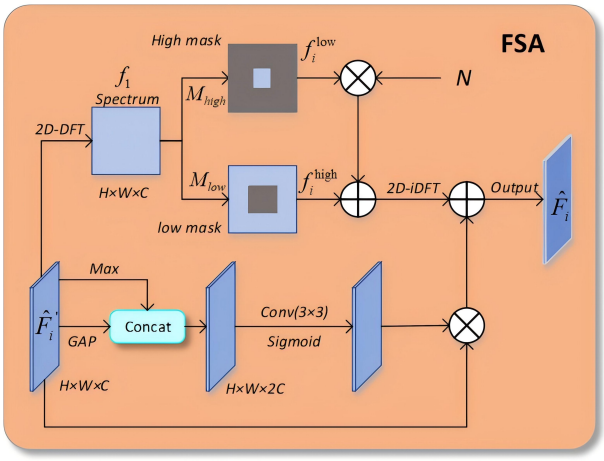

Fig. 4. The structure of FSA. 2D-DFT stands for 2D Discrete Fourier Transform, and 2D-iDFT stands for 2D Inverse Discrete Fourier Transform.

and reverted to the spatial domain via the Inverse 2D DFT, resulting in feature $\hat{F}_i''$, as calculated in (5):

$$f_i' = f_i^{high} \oplus (f_i^{low} \otimes N),$$
$$\hat{F}_i''(x,y) = \frac{1}{HW} \sum_{U=0}^{H-1} \sum_{V=0}^{W-1} f_i'(U,V) e^{j2\pi\left(\frac{Ux}{H} + \frac{Vy}{W}\right)}, \tag{5}$$

This design optimizes global background modeling via adaptive filtering while explicitly preserving high-frequency edge information to prevent the submersion of small objects.

Spatial Attention Branch. Concurrently, to enhance salient target localization, input features $\hat{F}_i'$ undergo channel-wise Max Pooling and Global Average Pooling. The resulting descriptors are concatenated and processed via a 3×3 convolution followed by a Sigmoid activation to generate an attention map, which is element-wise multiplied with $\hat{F}_i'$ to yield the enhanced output $SA(\hat{F}_i')$.

Feature Fusion. As formulated in (6), the final module output $\hat{F}_{out}$ is synthesized via the element-wise addition of the frequency-domain reconstruction $\hat{F}_i''$ and the spatially enhanced features $SA(\hat{F}_i')$. This dual-domain strategy synergizes global spectral awareness with local saliency focusing, ensuring comprehensive feature representation.

$$\hat{F}_{out} = \hat{F}_i'' \oplus SA(\hat{F}_i'). \tag{6}$$

D. Dynamic Hypergraph Fusion

To mitigate the inherent limitations of CNNs in modeling complex "multi-to-multi" correlations, we design the DHF module. Positioned at the core of the encoder, DHF is designed to mine latent high-order semantic correlations through adaptive hypergraph computation. By synergizing this with an efficient local perception branch, the module achieves comprehensive cross-scale and cross-location feature enhancement. as illustrated in Fig. 1.

Dynamic Hypergraph Fusion. as illustrated in the left panel of Fig. 5. To synthesize multi-scale context, the DHF module aligns hierarchical feature inputs to a unified spatial resolution via resampling operations. The aligned features are concatenated and integrated through a 1×1 convolution before bifurcating into two parallel streams: the High-Order Branch and the Low-Order Branch.

High-Order. Centered on the C3 Adaptive Hypergraph (C3AH) module, this branch transcends traditional fixed-threshold graph construction by implementing data-driven *Adaptive Hyperedge Generation* (see the right panel of Fig. 5). Specifically, context vectors derived from global pooling operations generate dynamic offsets to modulate global prototypes, yielding content-aware hyperedge prototypes tailored to the input image. A continuous incidence matrix A is then computed based on the similarity between pixel features and these prototypes. Subsequently, *Hypergraph Convolution* executes a two-stage "vertex-hyperedge-vertex" message passing process. This mechanism effectively models complex, non-local semantic co-occurrences—such as implicit correlations between vehicles and infrastructure—to achieve high-order feature enhancement.Aggregation: DHF aggregates contextual information from the upstream FSA module and multi-level features. Distribution: The features enhanced by both high-order and low-order processing are redistributed to various encoder stages via the Full-Pipeline Aggregation-Distribution (FullPAD) Tunnel. Following channel adjustment via 1×1 convolution, these enhanced features undergo attention-based weighted fusion with the outputs of the backbone's RepNCSPELAN4 module and the SCDown downsampled features.

Low-Order Branch. To complement the global semantic modeling of the hypergraph, this parallel branch incorporates the DS-C3k module to capture local texture details. Utilizing a Cross Stage Partial (CSP) architecture based on depthwise separable convolutions, it efficiently extracts fine-grained features within local receptive fields. This design ensures the robust preservation of geometric contours crucial for small object detection while minimizing computational overhead.

Full-Pipeline Distribution. As illustrated in Fig. 1, the DHF module is integrally coupled with the entire Encoder hier-

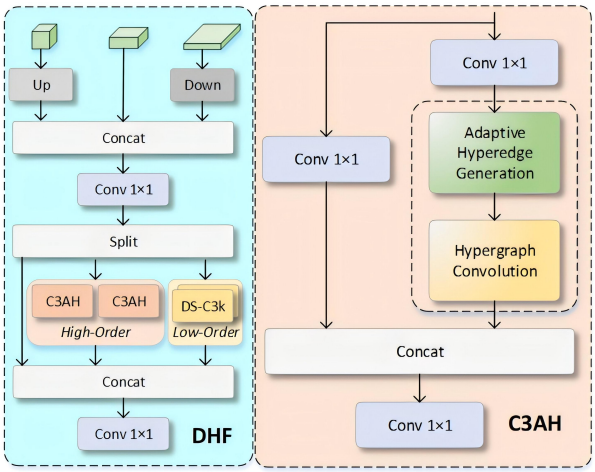

Fig. 5. The detailed structure of DHF.

TABLE I
COMPARISON WITH STATE-OF-THE-ART REAL-TIME DETECTORS ON MS COCO DATASET.

Method	Input Size	Params(M)	FLOPs(G)	FPS	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
RT-DETR-R18	640	10.28	26.9	231	0.438	0.609	0.482	0.282	0.480	0.595
DEIM-s	640	10.24	25.3	272	0.445	0.613	0.484	0.288	0.486	0.597
D-FINE-s	640	10.27	26.5	244	0.446	0.612	0.485	0.285	0.484	0.597
YOLOv10-s	640	9.21	20.7	394	0.433	0.597	0.479	0.278	0.480	0.580
YOLOv11-s	640	9.46	21.7	376	0.441	0.606	0.483	0.281	0.482	0.584
YOLOv12-s	640	9.43	21.6	405	0.439	0.603	0.481	0.280	0.480	0.589
YOLOv13-s	640	10.13	23.1	359	0.443	0.604	0.483	0.284	0.481	0.591
Ours	640	**10.26**	**25.9**	**280**	**0.457**	**0.621**	**0.491**	**0.301**	**0.486**	**0.599**

archy via the FullPAD Tunnel mechanism. Features enhanced by both high-order and low-order processing are redistributed to various encoder stages. This full-pipeline aggregation-distribution mechanism transcends the unidirectional information flow limitations of traditional FPNs, enabling the entire encoder network to share the global high-order semantics mined by DHF.

IV. EXPERIMENTS

A. Experimental Setting

We evaluated WH-Mamba on MS COCO 2017 and BDD100K, utilizing standard Average Precision (AP) metrics. BDD100K was specifically selected to assess robustness under diverse weather and illumination conditions. All models were implemented in PyTorch on dual NVIDIA RTX 3090 GPUs and trained for 300 epochs with a global batch size of 16 and an input resolution of 640×640. We employed the AdamW optimizer (initial learning rate 1×10^{-3} with cosine decay) and standard data augmentations, including Mosaic and Mixup, following established protocols.

B. Comparison with Other Methods

We benchmarked WH-Mamba against leading real-time object detectors, including the CNN-based YOLO series and Transformer-based models. In these experiments, we adopt the small variant to evaluate the architecture's efficiency and effectiveness under strict parameter constraints.

Performance on MS COCO. As summarized in Table I, WH-Mamba achieves superior performance on the COCO validation set. With an AP of 45.7%, our model outperforms the latest YOLOv13-s by 1.4% and the Transformer-based DEIM-s by 1.2%, while maintaining a competitive inference speed of 280 FPS. Crucially, in the small object regime (AP_S), WH-Mamba attains a score of 30.1%, surpassing the previous best, D-Fine-s, by a significant margin of 1.6%. This empirical evidence strongly supports our hypothesis that the WCMB effectively preserves high-frequency details that are typically lost in conventional downsampling operations.

Performance on BDD100K. As detailed in Table II, The results on the BDD100K dataset highlight the model's robustness in complex traffic scenarios. WH-Mamba achieves an AP of 43.2%, exceeding YOLOv13-s and DEIM-s by 1.8% and 1.7%, respectively. Notably, the high AP_{75} (48.5%) indicates that our model generates tighter, more accurate bounding boxes, which is critical for collision avoidance in autonomous driving. The performance gains in this unstructured environment verify the efficacy of the DHF module in modeling implicit, high-order correlations among traffic agents.

TABLE II
COMPARISON WITH STATE-OF-THE-ART REAL-TIME DETECTORS ON BDD100K DATASET.

Method	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
RT-DETR-R18	0.401	0.601	0.469	0.234	0.445	0.501
DEIM-s	0.415	0.607	0.477	0.240	0.451	0.514
D-FINE-s	0.408	0.603	0.475	0.239	0.453	0.511
YOLOv10-s	0.400	0.592	0.459	0.223	0.435	0.496
YOLOv11-s	0.409	0.597	0.462	0.225	0.439	0.506
YOLOv12-s	0.411	0.603	0.470	0.227	0.441	0.509
YOLOv13-s	0.414	0.605	0.473	0.231	0.446	0.513
Ours	**0.432**	**0.618**	**0.485**	**0.259**	**0.465**	**0.521**

C. Ablation Study

To systematically quantify the contribution of each proposed component, we conducted extensive ablation studies on the BDD100K dataset.

Impact of Individual Modules. As detailed in Table III, the baseline achieves an AP of 41.5%.Integration of WCMB: Replacing the standard backbone stage with our WCMB yields a substantial gain of +0.8% AP. This improvement is primarily attributed to the QDS strategy, which better captures the radial geometry of road scenes compared to standard scanning.Effect of FSA Module: Introducing the Frequency-Spatial Attention (FSA) module further boosts performance to 42.8% AP. Notably, the small object precision (AP_S) sees a marked improvement, validating the FSA's ability to perform "spectral denoising" and recover fine-grained textures in the frequency domain.Contribution of DHF: Finally, incorporating the DHF module elevates the AP to 43.2%. This final leap demonstrates the necessity of modeling high-order, non-local

TABLE III
PERFORMANCE IMPACT OF DIFFERENT COMPONENTS ON WH-MAMBA ON THE BDD100K DATASET.

Method	WCM	FSA	DHF	AP_s	AP	AP_{50}
DEIM-s				0.240	0.415	0.607
+WCM	✓			0.248	0.423	0.611
+FSA	✓	✓		0.255	0.428	0.614
+DHF	✓	✓	✓	**0.259**	**0.432**	**0.618**

TABLE IV
COMPARISON OF DIFFERENT SCANNING STRATEGIES ON BDD100K DATASET.

Strategy	AP	AP_{50}	AP_{75}
Bidirectional	0.416	0.607	0.476
SS2D	0.418	0.408	0.478
QDS	**0.423**	**0.611**	**0.481**

correlations in dense traffic flows, which conventional FPNs fail to capture.

Analysis of Scanning Strategies. A core contribution of this work is the QDS strategy. To validate its superiority over existing 2D serialization methods, we compared it against Bidirectional Scanning and Cross-Scan (SS2D) within the WCM block. As shown in Table IV, the bidirectional strategy yields the lowest performance (41.6% AP) due to its inability to capture vertical dependencies. While the orthogonal SS2D improves this to 41.8%, it still struggles with the diagonal semantic continuity inherent in road perspective lines. By contrast, QDS achieves the highest AP of 42.3%. This confirms that aligning the state-space scanning trajectory with the physical radial topology of the scene is crucial for minimizing information loss during the 2D-to-1D transformation.

V. CONCLUSION

In this paper, we introduced WH-Mamba, a novel framework designed to resolve the critical efficiency-robustness trade-off in intelligent traffic perception. By innovatively synergizing State Space Models with Discrete Wavelet Transforms and Hypergraph theory, we effectively reconciled the conflict between global context modeling and local detail preservation. Specifically, the proposed WCMB, equipped with the QDS strategy, successfully aligns 1D selective scans with the 2D radial topology of road scenes, preventing the loss of spatial structure. Furthermore, the FSA module performs spectral denoising to handle adverse weather conditions, while the DHF module captures complex, non-local semantic correlations beyond the receptive field limits of standard convolutions. Extensive evaluations on the MS COCO and BDD100K benchmarks demonstrate that WH-Mamba significantly outperforms state-of-the-art detectors, including YOLOv13 and DEIM, particularly in small object detection. In future work, we plan to optimize the architecture for deployment on resource-constrained edge devices and explore its scalability to multi-modal perception systems for end-to-end autonomous driving.

REFERENCES

[1] T. Feng, W. Wang, and Y. Yang, "A survey of world models for autonomous driving," 2025, arXiv:2501.11260. [Online]. Available: https://arxiv.org/abs/2501.11260

[2] M. Nikouei *et al.*, "Small object detection: A comprehensive survey on challenges, techniques and real-world applications," *Intelligent Systems with Applications*, vol. 27, p. 200561, Sept. 2025.

[3] H. Gupta, O. Kotlyar, H. Andreasson, and A. J. Lilienthal, "Robust object detection in challenging weather conditions," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 7508–7517.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 779–788.

[5] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," 2024, arXiv:2410.17725. [Online]. Available: https://arxiv.org/abs/2410.17725

[6] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-centric real-time object detectors," 2025, arXiv:2502.12524. [Online]. Available: https://arxiv.org/abs/2502.12524

[7] T. Lin *et al.*, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 936–944.

[8] N. Carion *et al.*, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 213–229.

[9] Z. Liu *et al.*, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 9992–10002.

[10] S. Huang *et al.*, "DEIM: DETR with improved matching for fast convergence," 2025, arXiv:2412.04234. [Online]. Available: https://arxiv.org/abs/2412.04234

[11] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *Proc. 1st Conf. Lang. Model. (COLM)*, 2024.

[12] Y. Liu *et al.*, "VMamba: Visual state space model," 2024, arXiv:2401.10166. [Online]. Available: https://arxiv.org/abs/2401.10166

[13] Q. Lin, Y. Li, and T. Zhang, "Haar wavelet-enhanced receptive field attention convolution for 3D point cloud object detection," in *Proc. 4th Int. Conf. Artif. Intell., Robot., Commun. (ICAIRC)*, 2024, pp. 753–759.

[14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[15] L. Jiang, G. Ma, W. Guo, and Y. Sun, "YOLO-DH: Robust object detection for autonomous vehicles in adverse weather," *Electronics*, vol. 14, no. 22, Art. no. 4476, 2025.

[16] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, "Hypergraph neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 3558–3565.

[17] A. Wang *et al.*, "YOLOv10: Real-time end-to-end object detection," 2024, arXiv:2405.14458. [Online]. Available: https://arxiv.org/abs/2405.14458

[18] M. Lei *et al.*, "YOLOv13: Real-time object detection with hypergraph-enhanced adaptive visual perception," 2025, arXiv:2506.17733. [Online]. Available: https://arxiv.org/abs/2506.17733

[19] Y. Zhao *et al.*, "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 16965–16974.

[20] Y. Peng *et al.*, "D-FINE: Redefine regression task of DETRs as fine-grained distribution refinement," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025, pp. 44015–44031.

[21] S. Huang, Y. Hou, L. Liu, X. Yu, and X. Shen, "Real-time object detection meets DINOv3," 2026, arXiv:2509.20787. [Online]. Available: https://arxiv.org/abs/2509.20787

[22] O. Siméoni *et al.*, "DINOv3," 2025, arXiv:2508.10104. [Online]. Available: https://arxiv.org/abs/2508.10104

[23] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," 2022, arXiv:2111.00396. [Online]. Available: https://arxiv.org/abs/2111.00396

[24] L. Zhu *et al.*, "Vision Mamba: Efficient visual representation learning with bidirectional state space model," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, 2024, Art. no. 2584.

[25] T. Huang *et al.*, "LocalMamba: Visual state space model with windowed selective scan," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2025, pp. 12–22.

[26] C. Xiao, M. Li, Z. Zhang, D. Meng, and L. Zhang, "Spatial-Mamba: Effective visual state space models via structure-aware state fusion," in *Proc. 13th Int. Conf. Learn. Represent. (ICLR)*, 2025.

[27] Q. Li, L. Shen, S. Guo, and Z. Lai, "WaveCNet: Wavelet integrated CNNs to suppress aliasing effect for noise-robust image classification," *IEEE Trans. Image Process.*, vol. 30, pp. 7074–7089, 2021.

[28] X. Zhao *et al.*, "Multiview hypergraph fusion network for change detection in high-resolution remote sensing images," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 4597–4610, 2024.

[29] Y. Feng *et al.*, "Hyper-YOLO: When visual object detection meets hypergraph computation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 4, pp. 2388–2401, Apr. 2025.