FA-DETR: Frequency-Aware for Aerial Small Object Detection

[1] Shuailiang Song
Dalian University of Technology
Dalian, China
songshuailiang@mail.dlut.edu.cn

[2] Fan Wang*
Dalian University of Technology
Dalian, China
wangfan@dlut.edu.cn

[3] Xiaopeng Hu
Dalian University of Technology
Dalian, China
huxp@dlut.edu.cn

[4] Xinrong Wu
Dalian University of Technology
Dalian, China
wuxinrong@mail.dlut.edu.cn

***Abstract*—Aerial small object detection is a challenging task, primarily attributed to the small size of targets and lack of discriminative texture. Existing methods rely on the preservation of shallow feature maps and global context modeling to boost detection performance. However, these methods still suffer from inefficient shallow detail preservation and high-level semantic redundancy, thus compromising their effectiveness. To enhance detailed information representation and semantic modeling, we propose FA-DETR, an efficient Transformer-based framework with frequency awareness for aerial small object detection. Specifically, we introduce a Wavelet-Transform Mamba Enhancement Module (WTMEM), which applies a wavelet transform to decompose shallow features into low-frequency and high-frequency subbands and models them separately. WTMEM enhances high-frequency detail features while capturing low-frequency global context, thereby preserving key target details under complex background interference. To suppress semantic redundancy, we design a Large-Kernel Convolutional Self-Attention (LKCSA) module to enable more effective semantic interaction. We also propose a Cross-Stage Information Reactivation Module (CSIRM) to alleviate the degradation of small object detail information during feature fusion, thereby improving the quality of features fed into the detection heads. Extensive experiments on VisDrone and SIMD datasets demonstrate that FA-DETR achieves competitive performance (+2.2 AP on VisDrone and +1.9 AP on SIMD) while maintaining low computational complexity.**

***Index Terms*—Aerial Small Object Detection, Transformer, Frequency-Aware, Mamba, Large Kernel Convolution**

I. Introduction

Aerial object detection plays an important role in intelligent transportation, urban governance, security surveillance, and disaster assessment. Compared with ground-view imagery, aerial images typically feature smaller targets and more cluttered scenes, making accurate detection more challenging. Classical detectors [1-3] are mostly built on convolutional neural networks (CNNs). In recent years, Transformer-based detectors [4-10] have leveraged the self-attention mechanism to enable unrestricted global context modeling, thus achieving remarkable performance in universal object detection tasks. Nevertheless, in the aerial scenes, this global modeling advantage fails to be directly transformed into effective representations for sparse targets. In aerial scenes, targets are extremely small while the background dominates, so target regions can be overwhelmed by background areas, causing small object details to be more easily lost during downsampling and cross-scale interactions. Maintaining shallow feature maps helps preserve details, but it incurs substantial computational and memory costs in Transformer frameworks. Dome-DETR [11] leverages density responses from shallow features to guide window attention toward foreground regions, but introduces heavy computational burdens. In contrast, RT-DETR [9] and its lightweight variants [12-14] achieve real-time performance yet rely more on deep, low-resolution modeling, which tends to reduce sensitivity to small objects. Therefore, an effective detection framework is needed to balance efficiency and fine-grained representation capability.

Beyond feature representation, transformer-based detectors employ global self-attention on high-level features for semantic interaction. In aerial scenes, the background dominates most regions. Therefore, global self-attention may assign similar weights to many features, reducing the contrast between salient targets and background regions. This further lowers sensitivity to critical cues. Recently, some models [5-7] adopt an enhanced global deformable self-attention mechanism that adaptively samples regions of interest to improve feature extraction efficiency. However, under complex backgrounds it remains vulnerable to redundant information, and attention can still be dispersed over background regions.

To enhance target detail feature representation and semantic modeling, we propose **FA-DETR**, an end-to-end framework for aerial small object detection. It achieves enhanced small object detection performance based on frequency-domain awareness and information compensation mechanisms. To efficiently preserve key shallow detail features, we introduce the Wavelet-Transform Mamba Enhancement Module (WTMEM). WTMEM leverages a frequency-aware mechanism to strengthen structural detail representations, improving the discriminability of small-object features with minimal computational overhead. To enhance the target semantic representation of high-level feature maps, we design a Large-Kernel Con-

* Corresponding authors.

volutional Self-Attention (LKCSA) module. LKCSA enables efficient semantic interaction via large-receptive-field convolutions, and enhances target-related semantic aggregation while avoiding the redundant computation of global attention. In addition, to alleviate the progressive attenuation of small object details during multi-scale feature fusion, we propose a Cross-Stage Information Reactivation Module (CSIRM). CSIRM establishes a recalibration pathway between the pre- and post-cross-scale fusion features, thereby recovering weakened small object responses and enabling the detection heads to receive representations that better preserve both semantic context and fine-grained details. In summary, our contributions are outlined as follows:

- We propose FA-DETR, an end-to-end framework for aerial small object detection. This framework aims to improve the detection performance of small targets in aerial scenes through frequency-domain awareness and information compensation mechanisms.
- We introduce the WTMEM to enhance shallow fine-grained feature representations via frequency-domain decoupling, and construct the CSIRM to compensate for detail loss during multi-scale fusion, thereby stabilizing the input representations for detection.
- We design the LKCSA to improve high-level semantic aggregation for small targets while suppressing redundant background interference.
- We validate the effectiveness of our proposed model through extensive experiments on the UAV dataset VisDrone and the remote-sensing vehicle dataset SIMD.

II. Related Work

A. Small Object Detection

Small object detection remains a long-standing challenge. Recent studies have explored various optimization strategies to enhance small object representation, such as high-pass filtering, feature fusion, and selective attention mechanisms. FBRT-YOLO [15] enhances localization by fusing semantic and spatial information, FFCA-YOLO [16] strengthens feature representation via multi-branch and dilated convolutions, and HS-FPN [17] preserves high-frequency details of small objects in feature pyramid interaction using frequency-domain high-pass filtering. However, most existing methods primarily focus on spatial-domain feature extraction and contextual modeling, while the potential of frequency-domain information remains largely underexplored.

B. Mamba

In linear time-invariant systems, state space models (SSMs) process 1D signals using state variables and ODEs. Mamba [18] enhances this but faces limitations with 2D visual data. The 2D Selective Scan (SS2D) mechanism [19] expands image patches in four directions to create independent sequences. These sequences are processed with the Selective Scan Space State Sequential Model (S6) [18] and then aggregated to reconstruct the 2D feature map. Wave-Mamba [20] has explored the application of wavelet transform integrated with Mamba in the field of low-light image enhancement, while we aim to investigate the potential of their combination in object detection.

III. Method

In this section, we propose the FA-DETR, an aerial small object detection model based on D-FINE-N. As shown in Fig. 1, FA-DETR consists of three core modules: WTMEM, LKCSA and CSIRM. WTMEM enhances fine-grained small object features by performing frequency-domain decoupling of low- and high-frequency components. LKCSA improves semantic modeling at the S5 layer with large-receptive-field convolutions, while reducing background interference. CSIRM compensates for the attenuation of small object responses during multi-scale feature fusion by cross-stage feature recalibration, improving the quality of the features fed into the detection heads.

A. Wavelet-Transform Mamba Enhancement Module

To preserve shallow fine-grained details, we replace the original HG block in Stage 2 of the backbone with the WTMEM. Given the input feature map $X \in \mathbb{R}^{C\times H\times W}$ at Stage 2 (S2), WTMEM first applies the Discrete Wavelet Transform (DWT) [20] to decompose the shallow features into low-frequency and high-frequency sub-bands.

$$(X_{LL}, X_{LH}, X_{HL}, X_{HH}) = \mathrm{DWT}(X), \tag{1}$$

where X_{LL} represents the low-frequency component of the input feature map X, and $\{X_{LH}, X_{HL}, X_{HH}\}$ represent the high-frequency components, all of which are in $\mathbb{R}^{C\times \frac{H}{2}\times \frac{W}{2}}$.

The low-frequency component X_{LL} is fed into the Low-Frequency State Space Module (LFSSM) for global long-range dependency modeling, where, in contrast, the high-frequency components $\{X_{\mathrm{LH}}, X_{\mathrm{HL}}, X_{\mathrm{HH}}\}$ are input to the Omnidirectional Perception Enhancement Module (OPEM) to capture fine-grained features. Finally, the Inverse Discrete Wavelet Transform (IDWT) is utilized to reconstruct the features, as shown in Fig. 1.

*1) **LFSSM**:* To efficiently capture long-range dependencies without incurring the computational cost of global self-attention, we integrate the recently popular VSSM [19] structure into the low-frequency branch. This integration leverages the selective modeling capabilities of its core Mamba module to avoid redundant interference and efficiently capture contextual information. Specifically, given the low-frequency input feature $X_{\mathrm{LL}} \in \mathbb{R}^{C\times \frac{H}{2}\times \frac{W}{2}}$, we first apply Layer Normalization (LN) to the input and then pass it into the VSSM for contextual modeling, resulting in X_{vssm}. Next, we apply LN to X_{vssm}, and then feed it into the Feed-Forward Network (FFN) for further processing, resulting in X_{local}. Finally, we add X_{vssm} and X_{local} through a residual connection to obtain the final low-frequency feature representation X_{lfssm}.

$$X_{vssm} = \mathrm{VSSM}(\mathrm{LN}(X_{LL})) + X_{LL}, \tag{2}$$

$$X_{local} = \mathrm{FFN}(\mathrm{LN}(X_{vssm})), \tag{3}$$

$$X_{lfssm} = X_{vssm} + X_{local}. \tag{4}$$

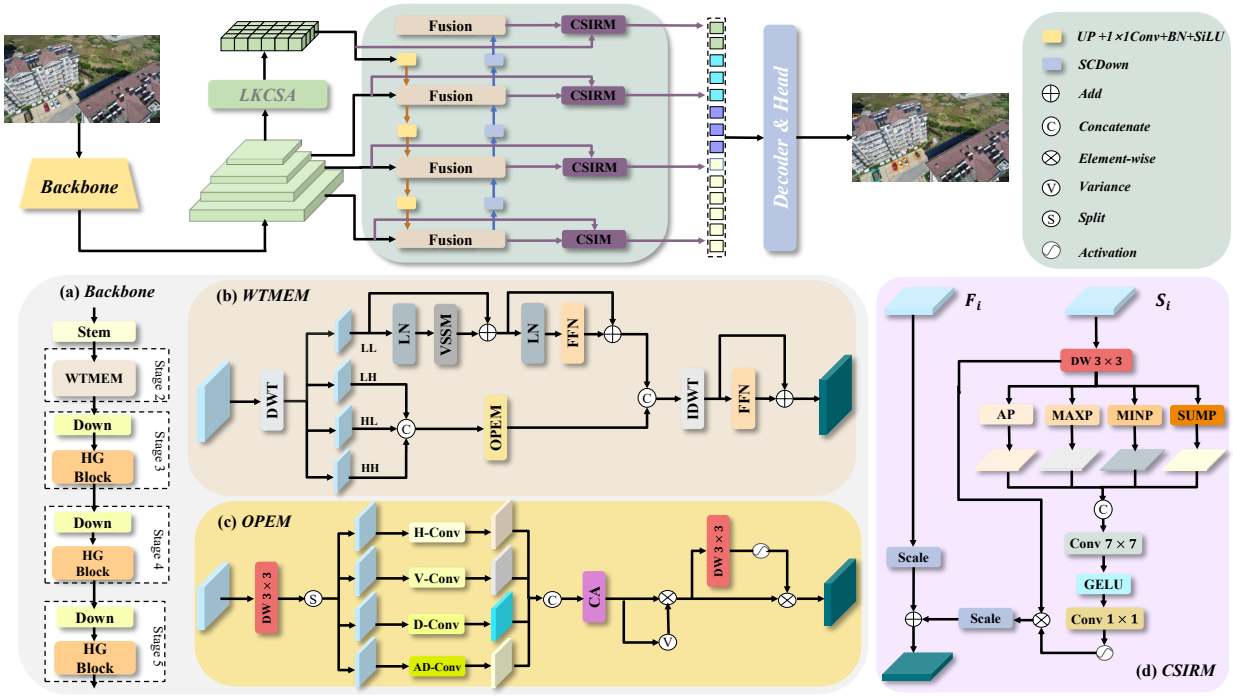

Fig. 1. The overall network architecture: the input image is fed into the backbone to extract four-level feature maps, which are then processed by intra-scale interaction and cross-scale interaction before being passed to the decoder. (a) shows the backbone structure, including the proposed WTMEM (b) and the OPEM (c). (d) presents the structure of the proposed CSIRM.

*2) **OPEM**:* In high-frequency processing, without explicit directional modeling and an effective filtering mechanism, strong responses from complex background textures can be amplified together with target details, thereby degrading the fine-grained representation of small objects. To this end, we introduce the OPEM in the high-frequency path of the WTMEM to model and denoise the fine-grained structures from $\{X_{LH}, X_{HL}, X_{HH}\}$.

Specifically, we first apply a 3×3 depth-wise convolution to the high-frequency input X_h, where X_h is concatenated from three high-frequency components along the batch dimension, and then split it into four sub-branches along the channel dimension with an emphasis on equal division:

$$X_{h1}, X_{h2}, X_{h3}, X_{h4} = \text{Split}(\text{DWConv}_{3\times3}(X_h)), \quad (5)$$

The four branches are processed using fixed 3x3 edge convolution kernels [21] in the horizontal, vertical, main-diagonal, and anti-diagonal directions to cover edges and texture responses of different orientations. The resulting features are then concatenated to obtain the aggregated representation.

$$\begin{aligned} X_{cat} = \text{Concat}(&\text{HConv}(X_{h1}), \text{VConv}(X_{h2}), \\ &\text{DConv}(X_{h3}), \text{ADConv}(X_{h4})), \end{aligned} \quad (6)$$

This parallel directional modeling captures high-frequency variations of small-object contours across orientations without introducing significant overhead.

To suppress noisy channels and emphasize features containing target-related information, we generate the channel attention map of X_{cat} using simple average pooling followed by an activation function, which is then element-wise multiplied with X_{cat}. To further refine the feature representation, we compute the variance across all channels for each spatial location, resulting in a variance modulation map. The output is then passed through a Sigmoid activation function and element-wise multiplied with the channel-modulated features to preserve the informative high-variance details containing target-related information.

$$X_{\text{va}} = \sigma\big(\mathcal{V}\big(\text{CA}(X_{\text{cat}})\big)\big) \otimes \text{CA}(X_{\text{cat}}), \quad (7)$$

where $\mathcal{V}$ denotes the variance calculation, and $CA(\cdot)$ represents the channel attention operation. The features after modulation are processed by a depth-wise gating mechanism to yield the refined high-frequency components.

$$X_{\text{opem}} = \sigma(\text{DWConv}_{3\times3}(X_{\text{va}})) \otimes X_{\text{va}}, \quad (8)$$

Finally, the refined low-frequency and high-frequency feature components are remapped back to the spatial domain using the Inverse Discrete Wavelet Transform (IDWT), so as to restore the complete feature representation.

$$X_{wtmem} = \text{IDWT}(X_{lfssm}, X_{opem}). \quad (9)$$

B. Large-Kernel Convolutional Self-Attention

To reduce redundant background interference with high-level semantics, we introduce a 13×13 large receptive field convolution, as shown in Fig. 2, to effectively capture the contextual information of the target region by expanding

TABLE I
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART DETECTORS ON THE VISDRONE2019 VALIDATION SET. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY.

Method	Publication	Input size	AP	AP_{50}	AP_s	Param	FLOPs
CNN-Based							
YOLOv11-S [22]	-	640 × 640	22.4	38.4	12.8	9.4M	21.3G
YOLOv11-M [22]	-	640 × 640	26.7	44.2	17.1	20.4M	67.7G
YOLOv12-S [3]	NeurIPS 2025	640 × 640	22.6	38.3	13.3	26.4M	88.9G
YOLOv12-M [3]	NeurIPS 2025	640 × 640	25.6	42.7	15.9	20.0M	67.7G
FBRT-YOLO-S [15]	AAAI 2025	640 × 640	25.9	42.4	-	2.9M	22.9G
FBRT-YOLO-M [15]	AAAI 2025	640 × 640	28.4	45.9	-	7.2M	58.7G
Transformer-Based							
Deformable DETR [5]	ICLR 2020	1300 × 800	27.1	42.2	-	40.0M	173.0G
Sparse DETR [8]	ICLR 2022	1300 × 800	27.3	42.5	-	40.9M	121.0G
RT-DETR [9]	CVPR 2024	640 × 640	26.5	44.2	18.5	20.0M	60.0G
UAV-DETR [12]	IROS 2025	640 × 640	**29.5**	**48.6**	20.5	21.3M	72.5G
D-FINE-N(4scale) [10]	ICLR 2025	640 × 640	28.7	46.6	**21.3**	4.5M	21.3G
FA-DETR(Ours)	-	640 × 640	**30.9**	**50.2**	**23.2**	4.6M	23.9G

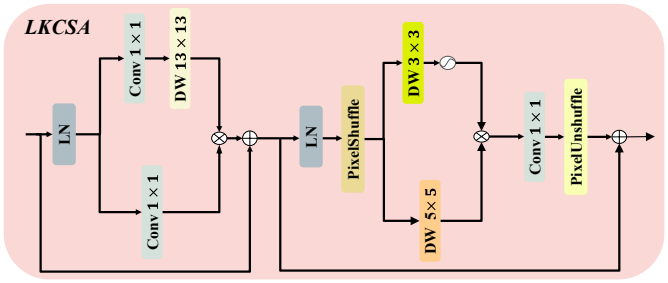

Fig. 2. Overall architecture of LKCSA. LKCSA replaces traditional self-attention

the receptive field. Subsequently, element-wise multiplication is applied to fuse the convolved features with the original features, which helps to emphasize the semantic information of the target region and reduce background interference, thereby improving the sensitivity to the semantics of sparse small objects. To further enhance the feature extraction capability for small objects, the LKCSA module incorporates a Dual-Scale Feed-Forward Network (DS-FFN). DS-FFN first upsamples the input feature map by a factor of 2 using a 1×1 convolution followed by PixelShuffle. It then processes the features in two parallel branches with 5×5 and 3×3 depthwise convolutions, respectively, and finally merges them via element-wise gating.

C. Cross-Stage Information Reactivation Module

To mitigate the loss of target detail information during multi-scale fusion, we introduce the CSIRM at the pyramid fusion level. CSIRM uses the corresponding pre-fusion feature X_{S_i} as a reference and establishes a re-calibrated branch on the fused feature X_{F_i}, thereby restoring and enhancing the response of small objects. Specifically, CSIRM aggregates multi-dimensional pooling information to extract detail cues from the pre-fusion feature X_{S_i}, and ultimately injects the recalibrated responses into the fused feature X_{F_i} through weighted fusion, thereby improving the quality of features after multi-scale fusion. Compared to single pooling, multi-dimensional pooling can complementarily capture target information, thereby effectively enhancing the expression of features. Given X_{S_i} and X_{F_i}, we first calibrate X_{S_i} with a 3×3 depth-wise convolution.

$$X_i^{cal} = \mathrm{DWConv}_{3\times3}(X_{S_i}), \tag{10}$$

We then apply average, min, max, and sum pooling to X_i^{cal} and concatenate them along the channel dimension.

$$\begin{aligned} X_i^{pool} = \mathrm{Concat}(&\mathrm{AP}(X_i^{cal}), \mathrm{MINP}(X_i^{cal}), \\ &\mathrm{MAXP}(X_i^{cal}), \mathrm{SUMP}(X_i^{cal})), \end{aligned} \tag{11}$$

This multi-dimensional pooling aggregation introduces complementary statistics, effectively enhancing target detail responses while reducing sensitivity to background fluctuations.

The multi-dimensional pooled feature X_i^{pool} undergoes a 7×7 convolution, GELU activation, 1×1 convolution, and Sigmoid normalization to obtain the spatial gating map A_i.

$$A_i = \sigma(\mathrm{Conv}_{1\times1}(\mathrm{GELU}(\mathrm{Conv}_{7\times7}(X_i^{pool})))), \tag{12}$$

We use A_i to modulate the preliminary calibrated feature X_i^{cal}, obtaining the final re-calibrated feature X_i^{rea} through element-wise multiplication, thereby focusing on key regions and suppressing background interference.

$$X_i^{rea} = X_i^{cal} \otimes A_i, \tag{13}$$

Finally, we calibrate the fused feature with a learnable scaling factor α_i and control the injection strength of the re-calibrated branch with another learnable factor β_i, where $\alpha_i, \beta_i \in \mathbb{R}^{C\times1\times1}$ denote channel-wise modulation factors. We then weight sum these two branches, reactivating the responses of the target regions, yielding the output X_{P_i}.

$$X_{P_i} = \alpha_i X_{F_i} + \beta_i X_i^{rea}. \tag{14}$$

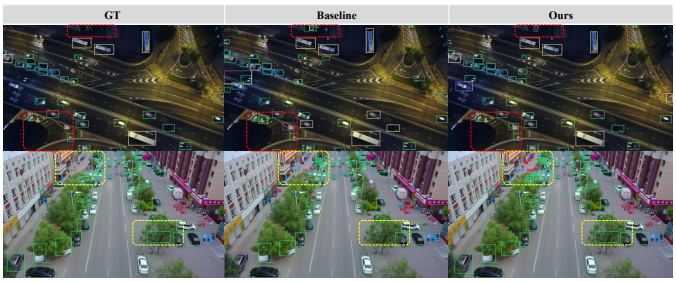

Fig. 3. Qualitative comparison of detection results on the VisDrone dataset. The first column shows the annotated images (ground truth), the second column presents the predictions of the baseline model D-FINE-N, and the third column shows the results of the proposed FA-DETR.

IV. Experimental Results and Analysis

A. Datasets

*1) **VisDrone**:* VisDrone [23] is a UAV-based aerial object detection benchmark with 6,471 training images, 548 validation images, and 3,190 test images, covering 10 object categories and over 2.5 million bounding boxes.

*2) **SIMD**:* SIMD [24] is a satellite imagery benchmark for multi-scale vehicle detection, containing 5,000 RGB images of size 1024×768 collected from 79 locations (4:1 train-test split). It provides 45,096 annotations across 15 categories, including vehicles, aircraft, and boats.

B. Implementation Details

All experiments are conducted on an NVIDIA GeForce RTX 3090 GPU. Our method is built upon the D-FINE-N baseline model, with the input resolution fixed at 640×640 for both training and testing phases. The model is trained for 300 epochs using the AdamW optimizer, with the hyperparameter configuration as follows: momentum coefficient of 0.9, weight decay of 0.0001, batch size of 8, and initial learning rate of 0.0008. We evaluate performance primarily using the COCO-style Average Precision (AP), along with additional AP scores at different thresholds and object scales, including **AP**, **AP**$_{50}$, **AP**$_s$.

C. Comparison Experiments

The experimental results in Table I demonstrate the superior performance of FA-DETR. Compared with the baseline D-FINE-N(4scale), FA-DETR improves AP from 28.7% to 30.9% (+2.2%), boosts AP$_{50}$ from 46.6% to 50.2% (+3.6%), and increases AP$_s$ from 21.3% to 23.2% (+1.9%). In terms of efficiency, the parameter count only slightly increases from 4.5M to 4.6M (+0.1M), while the computational complexity rises from 21.3G to 23.9G (+2.6 GFLOPs). With comparable computational costs, FA-DETR achieves the best overall performance among all comparison algorithms, striking an excellent balance between detection accuracy and computational efficiency.

As shown in Fig. 3, we select two challenging examples with complex backgrounds to compare our method with the baseline on small object detection. In the first image, the red dashed box highlights a typical false positive of the baseline, where traffic signs and wall textures in an unlabeled region are mistakenly recognized as vehicles. In contrast, our method effectively suppresses such background-induced confusion and avoids misclassification. In the second image, the yellow dashed boxes indicate missed detections in both an unlabeled region and a labeled region, suggesting that the baseline struggles to detect occluded objects. By comparison, our method more consistently detects the occluded targets and provides more accurate localization.

To further verify the generalization capability of FA-DETR, we evaluate the method on the SIMD dataset, and the experimental results are presented in Table II. Compared with the baseline D-FINE-N(4scale), our method improves AP from 65.1% to 67.0% (+1.9%) and boosts AP$_{50}$ from 79.2% to 81.2% (+2.0%). Its accuracy performance is comprehensively superior to that of other comparison algorithms.

TABLE II
Performance Comparison of Different SOTA Detectors on the SIMD Dataset

Pool	AP	AP$_{50}$	Param	FLOPs
RT-DETR-R18 [9]	63.7	78.6	19.8M	57.0G
YOLOv9-M [2]	62.2	76.6	20.0M	76.3G
Deformable DETR [5]	59.7	75.6	40.0M	196.0G
EMSD-DETR [14]	64.3	79.4	18.4M	68.3G
HPS-DETR [13]	63.5	79.8	15.5M	68.3G
D-FINE-N(4scale) [10]	65.1	79.2	4.4M	21.3G
FA-DETR(ours)	**67.0**	**81.2**	4.6M	23.9G

D. Ablation Studies

The ablation experiment results in Table III demonstrate that the performance improvement of the FA-DETR model on the VisDrone dataset stems from the synergistic gain effect of each module. After replacing the baseline module with WTMEM, the AP, AP$_{50}$, and AP$_s$ of the model are increased by 1.2%, 1.9%, and 1.1%, respectively. This improvement can be attributed to the effective capture of edge details of small objects by WTMEM. Following the further introduction of LKCSA, the AP, AP$_{50}$, and AP$_s$ of the model are further improved by 0.5%, 0.8%, and 0.3%, respectively. Among these, the improvement margin of AP$_s$ is limited. The reason is that LKCSA focuses more on modeling high-level semantic information with larger receptive fields, which can effectively avoid the loss of global information but has relatively insufficient sensitivity to small objects with sparse feature distribution. Finally, after integrating CSIRM, the various metrics of the model are increased by an additional 0.5%, 0.9%, and 0.5%, respectively. Through its information compensation mechanism, this module further optimizes the overall detection performance of the model.

To further verify the effectiveness of the multi-dimensional pooling in CSIRM, we conduct comparative experiments on the VisDrone dataset. As shown in Table IV, multi-dimensional pooling achieves a significant overall performance improvement compared with single average pooling (AP). We

TABLE III
ABLATION STUDY OF OUR PROPOSED METHOD ON THE VISDRONE DATASET

WTMEM	LKCSA	CSIRM	AP	AP_{50}	AP_s	Param	FLOPs
×	×	×	28.7	46.6	21.3	4.5M	21.3G
✓	×	×	29.9	48.5	22.4	4.6M	22.7G
×	✓	×	29.7	48.2	22.2	**4.4M**	**21.2G**
×	×	✓	29.9	48.6	22.1	4.6M	22.5G
✓	✓	×	30.4	49.3	22.7	4.6M	22.6G
✓	✓	✓	**30.9**	**50.2**	**23.2**	4.6M	23.9G

attribute this gain to its ability to capture complementary fine-grained details across multiple dimensions, providing a stronger information basis for re-activating target regions.

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT POOLING MECHANISMS IN CSIRM ON THE VISDRONE DATASET

Pool	AP	AP_{50}	AP_s	Param	FLOPs
AP	29.4	47.7	21.6	4.6M	22.5G
Ours	**29.9**	**48.6**	**22.1**	4.6M	22.5G

V. CONCLUSION

In this paper, we propose FA-DETR, a novel end-to-end Transformer-based framework tailored for aerial small object detection. To better preserve shallow fine-grained details and suppress redundant high-level semantics, we propose the Wavelet-Transform Mamba Enhancement Module (WTMEM), Large-Kernel Convolutional Self-Attention (LKCSA), and the Cross-Stage Information Reactivation Module (CSIRM), which together significantly enhance small-object detection performance. Extensive experiments on the VisDrone and SIMD datasets demonstrate that our method can better focus on key target cues and effectively suppress irrelevant background interference in complex aerial scenes, achieving more accurate localization and improved detection performance with only marginal additional computational overhead.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[2] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao, "Yolov9: Learning what you want to learn using programmable gradient information," in *European conference on computer vision*. Springer, 2024, pp. 1–21.

[3] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524*, 2025.

[4] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 16 965–16 974.

[5] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[6] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "Dab-detr: Dynamic anchor boxes are better queries for detr," *arXiv preprint arXiv:2201.12329*, 2022.

[7] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "Dn-detr: Accelerate detr training by introducing query denoising," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 619–13 627.

[8] B. Roh, J. Shin, W. Shin, and S. Kim, "Sparse detr: Efficient end-to-end object detection with learnable sparsity," *arXiv preprint arXiv:2111.14330*, 2021.

[9] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 16 965–16 974.

[10] Y. Peng, H. Li, P. Wu, Y. Zhang, X. Sun, and F. Wu, "D-fine: Redefine regression task in detrs as fine-grained distribution refinement," *arXiv preprint arXiv:2410.13842*, 2024.

[11] Z. Hu, P. Wu, J. Chen, H. Zhu, Y. Wang, Y. Peng, H. Li, and X. Sun, "Dome-detr: Detr with density-oriented feature-query manipulation for efficient tiny object detection," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 101–110.

[12] H. Zhang, H. Zhang, K. Liu, Z. Gan, and G.-N. Zhu, "Uav-detr: efficient end-to-end object detection for unmanned aerial vehicle imagery," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 15 143–15 149.

[13] X. Wang and H. Chen, "Hps-detr: Enhancing small object detection with lightweight feature extraction and transformer integration," *Authorea Preprints*, 2025.

[14] C. Zhang and J. Yang, "Emsd-detr: efficient small object detection for uav aerial images based on enhanced rt-detr model," *The Journal of Supercomputing*, vol. 81, no. 9, pp. 1–33, 2025.

[15] Y. Xiao, T. Xu, Y. Xin, and J. Li, "Fbrt-yolo: Faster and better for real-time aerial image detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 8, 2025, pp. 8673–8681.

[16] Y. Ge, W. Ji, S. Yin, and W. Zhang, "Small object detection in remote sensing based on feature fusion and channel attention," in *2024 8th International Symposium on Computer Science and Intelligent Control (ISCSIC)*. IEEE, 2024, pp. 195–200.

[17] Z. Shi, J. Hu, J. Ren, H. Ye, X. Yuan, Y. Ouyang, J. He, B. Ji, and J. Guo, "Hs-fpn: High frequency and spatial perception fpn for tiny object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 7, 2025, pp. 6896–6904.

[18] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First conference on language modeling*, 2024.

[19] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "Vmamba: Visual state space model," *Advances in neural information processing systems*, vol. 37, pp. 103 031–103 063, 2024.

[20] W. Zou, H. Gao, W. Yang, and T. Liu, "Wave-mamba: Wavelet state space model for ultra-high-definition low-light image enhancement," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1534–1543.

[21] B. Jähne, *Digital image processing*. Springer, 2005.

[22] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024.

[23] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and tracking meet drones challenge," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 11, pp. 7380–7399, 2021.

[24] M. Haroon, M. Shahzad, and M. M. Fraz, "Multisized object detection using spaceborne optical imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3032–3046, 2020.