Journal of Beijing Institute of Technology, 2026, Vol. 35, No. 2

Small Target Detection for UAVs Based on High-Low Frequency Attention and Local Affine Modeling

Junming Gao[1], Yanshan Zhang[1]✉, Yuanzhang Fan[1], Bao Tian[1], Yinhui Xu[2]

(1. School of Computer Science, Zhengzhou University of Aeronautics, Zhengzhou 450046, China;
2. Zhengzhou Research Institute, Beijing Institute of Technology, Zhengzhou 450000, China)

Abstract: Unmanned aerial vehicle (UAV) aerial images often feature rapidly changing perspectives, extremely small target scales, and significant occlusions and background interference, posing dual challenges to accuracy and stability in real-time detection. To address these issues, this paper proposes frequency-affine-lightweight detection (FALDet) within the you only look once version 8 (YOLOv8) framework, systematically improving detection through three main approaches: First, replacing spatial pyramid pooling-fast (SPPF) with intra-scale feature interaction-high-low frequency attention (AIFI-HiLo) to model high-frequency local and low-frequency global attention in parallel, balancing edge details and long-range semantics while maintaining low computational overhead. Second, it replaces part of C2f with local affine deformable convolution (LA-DCN), introducing a unified local affine sampling grid to reduce degrees of freedom and enhance stability against rotational, scaling, and translational deformations. Third, it designs lightweight cross-scale dynamic detection head (LiteX-DyHead), which effectively improves recall and localization consistency for dense small objects through lightweight preprocessing, dynamic/deformable alignment, and multi-scale fusion. Using VisDrone2019 as the primary evaluation dataset, ablation and comparative experiments were conducted under unified training strategies and input resolutions. Results demonstrate that FALDet achieves stable improvements over YOLOv8s in both mAP@0.5 and mAP@0.5:0.95 while maintaining high frames per second, validating the effectiveness and practicality of the proposed method. The method's effectiveness was further validated on the SIMD dataset.

Keywords: unmanned aerial vehicle (UAV); small object detection; local affine deformable convolution (LA-DCN); dynamic detection head

1 Introduction

In recent years, with the widespread application of unmanned aerial vehicles (UAVs) in urban security, traffic monitoring, disaster response, and smart agriculture, visual perception tasks based on UAV platforms have become a frontier in computer vision research [1–3]. Among these, object detection, as a fundamental component, directly impacts the performance and stability of downstream tasks such as object tracking, cross-camera recognition, and behavior understanding [4]. Compared to ground-level perspectives, UAV aerial imagery presents distinct challenges: significant variations in viewpoint, wide imaging scale spans, minute target dimensions, frequent occlusions and overlaps, and complex backgrounds

Manuscript received Oct. 4, 2025; revised Nov. 14, 2025; accepted Nov. 28, 2025. The associate editor coordinating the review of this manuscript was Dr. Lijuan Jia. This work was supported by the Graduate Student Workstation (No. 2025YJSJD4), the Key Research and Development Project of Henan Province (No. 231111212000), the Scientific Research Team Plan of Zhengzhou University of Aeronautics (No. 23ZHTD01005), the Key Scientific and Technological Project of Henan Province (Nos. 242102210150, 242102220048, 252102220053), the Key Research Project of the Henan Provincial Higher Education (No. 24A520051),the Henan Province Talent Support Program (No. 254000510003), and Program for Science and Technology of Henan Province of China (No. 242300421411).
✉ Corresponding author. Email: yanshan@zua.edu.cn
DOI: 10.15918/j.jbit1004-0579.2025.069

with strong interference. These characteristics impose a triple challenge of "small, dense, and variable" on detection algorithms [5]: insufficient edge and semantic representation for small targets, significant interference between bounding boxes due to dense distribution, and frequent geometric and pose variations that are difficult to align [6].

Among numerous detection paradigms, the you only look once (YOLO) series has gained widespread adoption in industrial deployment and academic research due to its end-to-end, single-stage, and low-latency advantages [7]. As the latest representative, YOLOv8 achieves a favorable balance in structural design, training strategies, and inference efficiency, demonstrating strong transferability [8]. However, directly applying YOLOv8 to UAV aerial photography scenarios, while demonstrating good performance, suffers from insufficient hierarchical attention to local-global information in its SPPF, limited adaptation to geometric deformations in C2f, and potential recall deficiencies in dense small-object scenarios with its native detector heads [9].

To address these challenges, this paper proposes frequency-affine-lightweight detection (FALDet): enhancing detail-context coupling via frequency-division attention, boosting geometric robustness through local affine deformable convolutions, and strengthening cross-scale interactions with a lightweight dynamic detector head. Key contributions include:

1) Intra-scale feature interaction-high-low frequency attention (AIFI-HiLo) frequency-split attention replaces spatial pyramid pooling-fast (SPPF): Drawing inspiration from frequency-split concepts, we design parallel high-frequency (local) and low-frequency (global) attention branches to respectively enhance edge texture and long-range dependency representation, fusing outputs in the channel dimension. This approach balances detail sensitivity and semantic modeling with low additional overhead, effectively expanding the effective receptive field.

2) Local affine deformable convolution (LA-DCN): Local affine deformable convolution replacing C2f: Generates a unified convolution sampling grid via local affine parameters, offering more controllable degrees of freedom and more stable training compared to pointwise-shifted DCN. Enhances alignment and robustness under geometric deformations like rotation, translation, and scaling while maintaining lightweight architecture.

3) LiteX-DyHead cross-scale dynamic detection head: Introduces dynamic channel selection and multi-scale information exchange pathways to enhance response consistency for small objects across different scales in both classification and regression tasks, significantly mitigating missed detections and localization fluctuations in dense scenes.

2 Related Work

2.1 Attention and Frequency-Separated Modeling

Attention mechanisms in convolutional networks have evolved along two primary pathways: one category comprises lightweight attention modules directly integrated as convolutional neural network (CNN) plugins to enhance feature discriminative power, such as SE-Net's channel relabeling [10], CBAM's channel-spatial cascading [11], ECA-Net's local cross-channel interactions [12], and SKNet's selectable receptive fields [13]; The other category involves Transformer backbones, which employ windowing or hierarchical strategies to mitigate computational bottlenecks of global self-attention in high-resolution scenarios. Examples include Swin Transformer, which achieves hierarchical processing and local-global balance through sliding windows [14]. Addressing the characteristics of drone aerial images—small, dense objects with complex backgrounds. the HiLo Attention approach splits the attention head into high-frequency (local, detail) and low-frequency (global, semantic) compo-

nents. This design balances small object edges with long-range dependencies while offering superior throughput efficiency on GPUs/CPUs [15]. The AIFI-HiLo module in this paper adopts this high/low-frequency parallel attention paradigm to enhance detail-context collaborative modeling with minimal additional overhead.

2.2 Deformable Convolution and Geometric Modeling

Standard convolutions compute on fixed sampling grids, struggling to adapt to rotations, scale changes, and non-rigid deformations. Deep convolutional networks (DCNs) learn offset parameters to enable variable sampling positions, significantly enhancing geometric adaptability [16]. DCNv2 further incorporates modulation mechanisms to boost spatial selectivity and representational power [17]. However, in scenarios with dense small objects and high resolution, per-point offsets introduce high degrees of freedom, unstable training, and deployment complexity. Consequently, some approaches shift toward affine/low-degree-of-freedom deformation modeling to improve stability and efficiency. Examples include generating kernel sampling grids via unified affine transformations [18], which preserve geometric expressiveness while reducing parameter and offset divergence risks. The proposed LA-DCN extends this approach: predicting a single local affine transformation per spatial location to uniformly generate sampling grids for convolutional kernels. This enhances alignment robustness under rotation/scaling/translation while controlling computational overhead, making it suitable for edge-side drone deployment [19].

2.3 Multi-Scale Fusion and Dynamic Detection Head

Multi-scale representation is particularly critical for small UAV targets. Feature pyramid network (FPN) pioneered top-down with lateral connections to construct pyramid features [20], while PANet further enhanced bottom-up localization information through path aggregation [21]. DyHead unifies multiple attentions "across layers, spaces, and channels" into a dynamic head framework that dramatically enhances scale awareness and task coherence [22]; VFNet improves candidate ranking through IoU-aware classification scores (IACS) [23], making dense detection more reliable; PAA mitigates the mismatch between positive/negative sample distribution and test targets through probabilistic anchor allocation [24]. Additionally, the end-to-end transformer detector (DETR) replaces NMS with ensemble prediction [25], while RT-DETR maintains the end-to-end paradigm by introducing an efficient hybrid encoder and high-quality query selection, achieving accuracy-speed tradeoffs comparable to or even better than the YOLO series in real-time scenarios [26]. The proposed LiteX-DyHead introduces dynamic channel selection and cross-scale interaction paths under lightweight constraints. Combined with FPN/PAN (path aggregation network) necks, it enhances recall and localization consistency for extremely small/dense objects.

3 Improved Algorithm

FALDet extends YOLOv8's "backbone-neck-detection head" paradigm: Replacing SPPF with AIFI-HiLo in the backbone to concurrently model high-frequency local and low-frequency global information through frequency-division parallel modeling; In the neck, it replaces part of C2f with LA-DCN, unifying the sampling grid with local affine transformations to enhance geometric robustness and multi-scale alignment; In the detection head, it adopts LiteX-DyHead, which robustly improves small object recall and localization consistency through dynamic channel selection and cross-scale interactions. The overall structure is shown in Fig. 1. maintaining tensor scale and deployment friendliness comparable to YOLOv8.

3.1 AIFI-HiLo

In the standard intra-scale feature interaction (AIFI) module, multi-head self-attention (MSA)

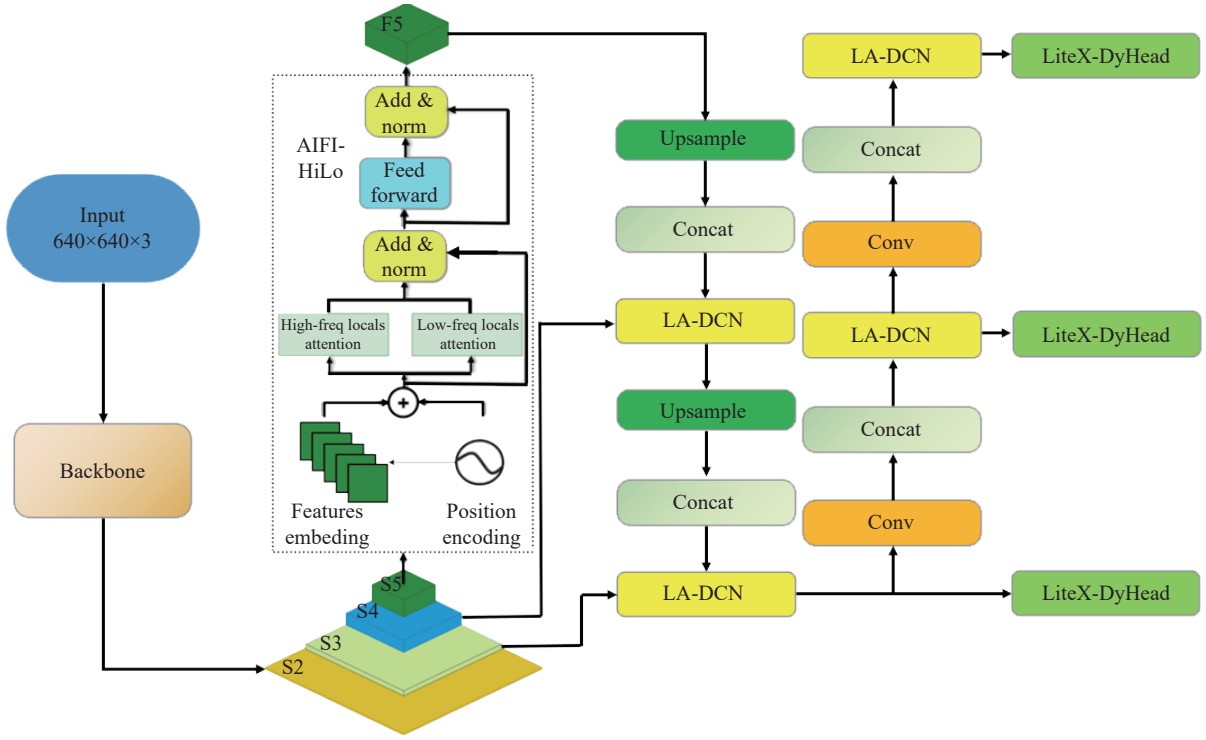


Fig. 1 Schematic diagram of the overall FALDet architecture

serves as the core computational unit, enhancing feature expression through global dependency modeling. However, MSA incurs excessive computational overhead in high-resolution or object-dense scenarios and tends to waste computational resources on irrelevant background regions. To reduce complexity while maintaining modeling capability, this paper integrates HiLo Attention into AIFI, constructing the novel AIFI-HiLo module.

The schematic of HiLo Attention is shown in Fig. 2. Its core idea is to divide features into two frequency components: High-frequency branch: Performs local attention at native resolution, focusing on capturing edge textures and small object details. Low-frequency branch: Downsamples input features before applying global attention, acquiring cross-regional long-range dependencies at lower computational cost. Through this combination of "local precision and global context", AIFI-HiLo achieves both sensitivity to small objects and comprehensive scene understanding.

Let the input features from the backbone be

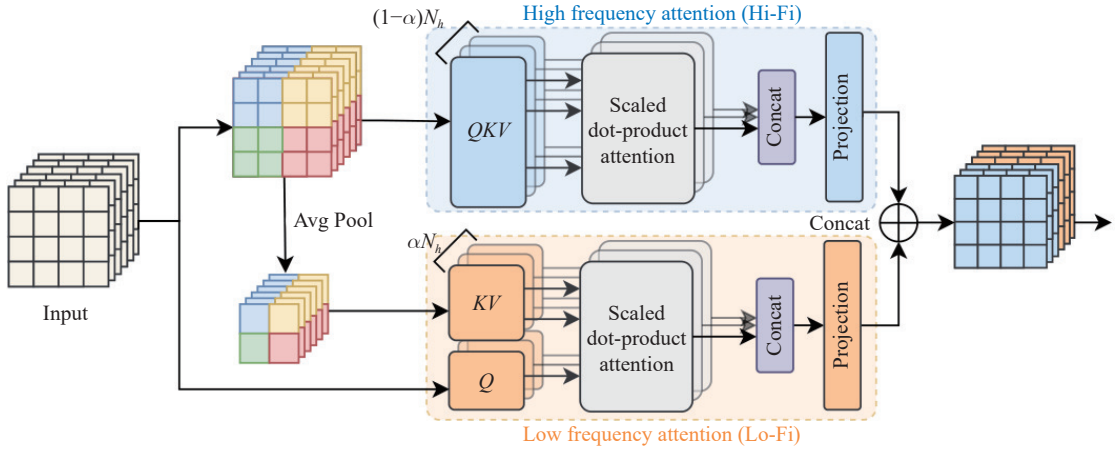


Fig. 2 Schematic diagram of the HiLo attention structure

$S_5 \in \mathbb{R}^{H\times W\times C}$. These are flattened into a sequence of length $N = H \times W$.

First, the flattened sequence is mapped into query, key, and value.

$$\boldsymbol{Q} = \boldsymbol{S}_5\boldsymbol{W}_Q, \boldsymbol{K} = \boldsymbol{S}_5\boldsymbol{W}_K, \boldsymbol{V} = \boldsymbol{S}_5\boldsymbol{W}_V \quad (1)$$

where $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V \in \mathbb{R}^{C\times d}$ is a learnable projection matrix, and$d$ represents the single-head dimension (total channels obtained via multi-head concatenation).

Divide multi-head attention into high/low frequency components (head ratio as $h_h : h_l$). The high-frequency branch applies a local window/sparse mask M_{local} to the high-frequency heads (Q_h, K_h, V_h) at the original resolution to constrain the attentiondomain and emphasize details:

$$\text{Attn}_{\text{high}} = \text{Softmax}\left(\frac{\boldsymbol{Q}_h\boldsymbol{K}_h^\top}{\sqrt{d}} + M_{\text{local}}\right)\boldsymbol{V}_h \quad (2)$$

where M_{local} is a block diagonal/circular window mask (elements outside the window are set to$-\infty$), ensuring attention remains within the neighborhood.

The low-frequency branch first downsamples the input by a stride of$s$ $\overline{S}_5 = \text{Down}_s(S_5)$ and maps it to $\overline{Q}_l$, $\overline{K}_l$, $\overline{V}_l$. Global attention is performed at this lower resolution before upscaling back to the original scale:

$$\text{Attn}_{low} = \text{Softmax}\left(\frac{\boldsymbol{Q}_l\boldsymbol{K}_l^\top}{\sqrt{d}}\right)\boldsymbol{V}_l. \quad (3)$$

This approach captures long-range dependencies while significantly reducing the quadratic computational cost of global attention.

Concatenate outputs from high- and low-frequency branches, then restore spatial geometry via output mapping:

$$\boldsymbol{F}_5 = \text{Reshape}\left(\text{Concat}\left(\text{Attn}_{\text{high}}, \text{Attn}_{\text{low}}\right)\boldsymbol{W}_O\right) \quad (4)$$

where $\boldsymbol{W}_O \in \mathbb{R}^{C\times C}$ is the output projection matrix.

AIFI-HiLo replaces fixed pooling with parallel high/low-frequency attention, balancing local details and global context while expanding the effective receptive field without altering the backbone resolution or channel layout. This design enhances sensitivity to minute targets, improves stability in complex backgrounds, and maintains low additional computational overhead.

3.2 LA-DCN

To further enhance the geometric modeling capability of detection heads in complex scenes, this paper proposes local affine deformable convolution (LA-DCN). Unlike traditional DCN, which independently predicts offsets at each sampling point within the kernel, LA-DCN learns a unified local affine transformation at each spatial location. This approach reduces the number of parameters while preserving modeling flexibility and enhances the stability of geometric transformations.

In conventional convolutional operations, the sampling positions of the kernel are fixed. For example, in a 3×3 convolution, the kernel samples 9 points on a regular grid of the input feature map, as shown in Fig. 3. This operation can be represented as

$$y(\boldsymbol{p}_0) = \sum_{k=1}^{K} w_k \cdot x(\boldsymbol{p}_0 + \boldsymbol{p}_k) \quad (5)$$

where p_0 denotes the center position of the convolution, p_k represents the coordinates of the regular sampling points of the kernel, and w_k is the kernel weight. K indicates the number of sampling points (e.g., K=9 corresponds to a 3×3 convolution).

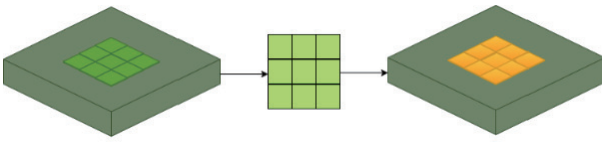


Fig. 3 Conventional operation flow

This approach is efficient and stable but lacks the ability to model geometric deformations, making it difficult to adapt to targets with significant rotation, scaling, or deformation.

To address the aforementioned issues,

deformable convolution (DCN) introduces a learnable offset Δp_k at each sampling point location. This allows the convolution kernel to adapt to the target shape rather than being constrained to a fixed grid. The DCN convolution process is illustrated in Fig. 4.

$$y(\boldsymbol{p}_0)=\sum_{k=1}^{K} w_k \cdot x(\boldsymbol{p}_0+\boldsymbol{p}_k+\Delta \boldsymbol{p}_k) \tag{6}$$

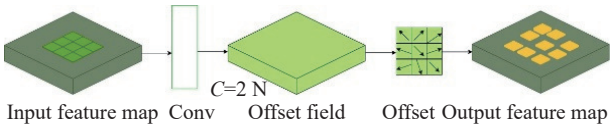


Fig. 4 DCN convolution operation flow

Although DCN significantly enhances modeling capabilities, independently learning offsets at each sampling point can lead to training instability and divergent sampling point distributions. Additionally, the number of parameters increases linearly with kernel size.

Standard convolutions sample on fixed regular grids, struggling to adapt to target geometric deformations. While DCN enhances flexibility by predicting independent offsets for each sampling point, its degrees of freedom increase linearly with kernel size, and independent point movements cause sampling instability. To address this, we propose LA-DCN: predicting a single local affine transformation at each spatial location to uniformly generate the entire sampling grid. This enables coordinated deformation of sampling points, significantly improving stability and parameter efficiency while preserving geometric expressiveness. The LA-DCN convolution pipeline is illustrated in Fig. 5.

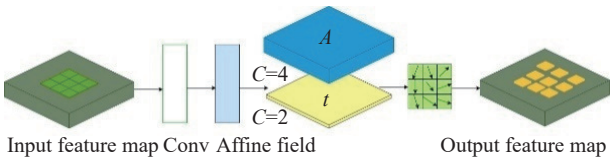


Fig. 5 LA-DCN convolution operation flow

Specifically, for a center position, we first predict a 2D affine matrix and a translation vector. Theseare then used to transform the uniformly sampled points.

$$\boldsymbol{q}_k=\boldsymbol{A}\cdot\boldsymbol{p}_k+\boldsymbol{t} \tag{7}$$

The convolution operation is thus rewritten as

$$y(\boldsymbol{p}_0)=\sum_{k=1}^{K} w_k \cdot x\left(\boldsymbol{p}_0+\boldsymbol{q}_k\right) \tag{8}$$

Here, the offset of all sampling points is determined by unified $\boldsymbol{A}$ and t. Compared to DCN convolution, this not only reduces the number of parameters from 18 in 3×3 convolution to 6 but also ensures geometric consistency within the convolution kernel. This allows sampling points to maintain stable spatial relationships under variations such as rotation, scaling, and translation.

3.3 LiteX-DyHead

To enhance feature representation for multi-scale object detection, this paper proposes a lightweight cross-scale dynamic detection head (LiteX-DyHead). This module takes the three feature maps $\{P_3, P_4, P_5\}$ output from the neck network as input, fusing information from adjacent layers at each feature level to obtain enhanced feature representations. Its overall structure is shown in Fig. 6, comprising the following steps.

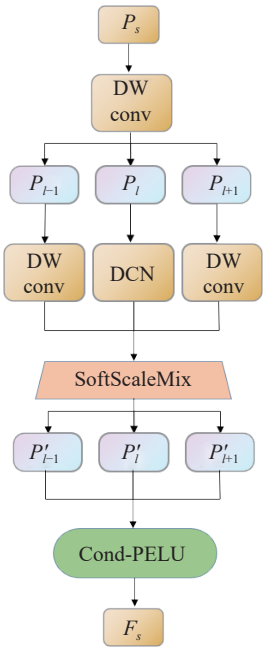


Fig. 6 LiteX-DyHead workflow

First, to reduce computational complexity and enhance feature compactness, the input multi-scale feature maps undergo deep-separable convolution (DWConv) processing, yielding lightweight intermediate features

$$\boldsymbol{P}_l = \mathrm{DWConv}(\boldsymbol{P}_s), s \in \{3,4,5\} \quad (9)$$

Deformable convolution (DCN) is introduced on the current layer features to model spatial offsets and enhance geometric modeling capabilities

$$\widehat{\boldsymbol{P}}_l = \mathrm{DCN}(\boldsymbol{P}_l, \Delta), \Delta = \phi(\boldsymbol{P}_l) \quad (10)$$

where $\phi(\cdot)$ denotes the shared offset predictor.

Simultaneously, features from adjacent layers undergo DWConv processing, followed by downsampling and upsampling operations respectively to align their resolution with the current layer

$$\widehat{\boldsymbol{P}}_{l-1} = \mathrm{Down}\left(\mathrm{DWConv}\left(\boldsymbol{P}_{l-1}\right)\right), \widehat{\boldsymbol{P}}_{l+1} = \mathrm{Up}(\mathrm{DWConv}(\boldsymbol{P}_{l+1})) \quad (11)$$

If the current layer is P_3, no lower-level features exist, defining $\widehat{P}_{l-1} = 0$. If the current layer is P_5, no higher-level features exist, defining $\widehat{P}_{l+1} = 0$.

After scale alignment, concatenate low-level, current-level, and high-level features. Apply a 1×1 convolution to predict per-pixel softmax weights, then perform weighted fusion:

$$\boldsymbol{W} = \mathrm{softmax}(\psi([\widehat{\boldsymbol{P}}_{l-1}, \widehat{\boldsymbol{P}}_l, \widehat{\boldsymbol{P}}_{l+1}]) \quad (12)$$

$$\boldsymbol{F}_l = \boldsymbol{W}_1 \odot \widehat{\boldsymbol{P}}_{l-1} + \boldsymbol{W}_2 \odot \widehat{\boldsymbol{P}}_l + \boldsymbol{W}_3 \odot \widehat{\boldsymbol{P}}_{l+1} \quad (13)$$

where $\psi(\cdot)$ denotes the 1×1 convolution operation, and $\odot$ represents element-wise multiplication. Through the SoftScaleMix module, the model can adaptively select the fusion ratio of multi-scale information.

To further enhance the nonlinear expressive power of features, a conditional PELU activation function is introduced. First, the fused feature F_l undergoes global average pooling (GAP) to obtain channel statistics and generate dynamic parameters a and b:

$$[a, b] = \eta(\mathrm{GAP}(\boldsymbol{F}_l)) \quad (14)$$

Subsequently, the features undergo nonlinear transformation via conditional PELU:

$$\boldsymbol{F}_s = a \cdot \sigma\left(\frac{\boldsymbol{F}_l}{b + \epsilon}\right) \quad (15)$$

where $\sigma(\cdot)$ denotes the SiLU activation function andϵ represents the numerical stability constant. By dynamically adjusting parameters, Cond-PELU adaptively modifies activation patterns based on input features, thereby enhancing discriminative capability.

LiteX-DyHead reconstructs the detection head, strengthening inter-scale information flow and consistency between classification/regression. While maintaining the same interface as the baseline, it reduces high-score misclassifications and stabilizes boundary predictions for dense and small objects.

4 Experiments and Results Analysis

4.1 Datasets and Experimental Setup

To validate the effectiveness of the proposed FALDET network in detecting small objects from drone aerial photography, this paper conducts a systematic experimental evaluation on the publicly available VisDrone2019 dataset. Fig. 7 shows some data examples. During training, several key parameter configurations were employed to enhance the dataset's generalization capability. Specifically, data augmentation methods included image scaling, horizontal flipping, concatenation, and image translation. The parameters listed in Tab. 1 represent the probabilities at which these methods were applied during training. These measures not only effectively improve model robustness but also enhance performance across diverse scenarios. Detailed configuration of model training parameters is provided in Tab. 1.

4.2 Evaluation Metrics and Implementation Details

In this study, we employ precision (P), recall

Fig. 7 VisDrone2019 dataset: (a) sparse; (b) shaded; (c) night; (d) sunny; (e) dense; (f) night+dense; (g) very small target; (h) complex background; (i) night+dense+occlusion

Tab. 1 Training Setup Parameters

Parameters	Setup
Epochs	200
Batch size	4
Optimizer	SGD
NMS IoU	0.7
Initial learning rate	0.01
Final learning rate	0.01
Momentum	0.937
Weight decay	0.0005
Image scale	0.5
Image flip left- right	0.5
Mossaic	1.0
Image translation	0.1
Close Mosaic	0

(R), and mean average precision (mAP) as evaluation metrics for model performance. AP measures detection accuracy for individual classes, while mAP is calculated as the average AP across all classes to assess overall detection performance. Specifically, mAP0.5 denotes the mAP value at an intersection over union (IoU) threshold of 0.5. IoU measures the overlap between predicted and ground-truth bounding boxes. During model evaluation, these metrics collectively reflect the detector's precision, recall, and overall performance.

$$P = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} \tag{16}$$

$$R = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{17}$$

$$\mathrm{AP} = \int_0^1 P(R)\,\mathrm{dR} \tag{18}$$

$$\mathrm{mAP} = \frac{1}{N}\int_0^1 P(R)\mathrm{dR} \tag{19}$$

In model performance evaluation, TP (true positive) denotes the number of samples correctly identified as positive by the model. Meanwhile, FP (false positive) represents the number of negative samples incorrectly classified as positive by the model. Furthermore, FN (false negative) indicates the number of positive samples

the model failed to correctly identify, i.e., the number of positive samples misclassified as negative.

4.3 Ablation Studies

To investigate the actual contribution of each module in the FALDET network to overall performance, this paper designed a series of ablation experiments. These experiments involved sequentially removing and combining the AIFI-HiLo module, LA-DCN module, and LiteX-DyHead detection head for analysis. YOLOv8s serves as the baseline architecture for comparison, with its mAP@0.5 of 39.3 on the VisDrone2019 test set established as the foundational reference value. Subsequently, by introducing individual modules sequentially, the impact of each enhancement on detection accuracy and inference efficiency was observed. Additionally, two or three modules were combined to evaluate synergistic enhancement effects. All experiments were conducted under unified datasets, training strategies, and evaluation metrics to ensure comparability and objectivity of results, as shown in Tab. 2.

Analysis of ablation experiment results demonstrates that the proposed method significantly impacts target detection in remote sensing images, not only enhancing detection performance but also reducing model parameters and computational complexity to a certain extent.

Fig. 8 presents the comparison curve of experimental results between the original and improved algorithms on the VisDrone dataset. As shown, with increasing training epochs, the mAP0.5 and mAP0.95 of the improved algorithm gradually surpass those of the original YOLOv8 algorithm.

Fig. 9 presents visualization results of YOLOv8s and the improved model on the VisDrone detection dataset. Comparison clearly demonstrates that the proposed model significantly reduces false positives and false negatives, achieving higher recognition accuracy.

4.4 Comparison with State-of-the-Art Methods

To comprehensively demonstrate the advantages of the proposed method, we conducted a series of comparative experiments on the VisDrone dataset using multiple representative state-of-the-

Tab. 2 Ablation experiment results

Methods	AIFI-HiLo	LA-DCN	LiteX-DyHead	mAP@0.5	mAP@0.5:0.95	Model size (MB)	Params (MB)	FPS
1	×	×	×	39.3	23.5	22.0	11.1	84
2	√	×	×	41.5	24.8	21.5	10.8	86
3	×	√	×	43.1	25.6	21.0	10.6	87
4	×	×	√	40.2	24.0	20.3	10.3	86
5	√	√	×	44.0	26.1	20.1	10.1	91
6	√	√	√	45.1	28.3	19.8	9.8	92

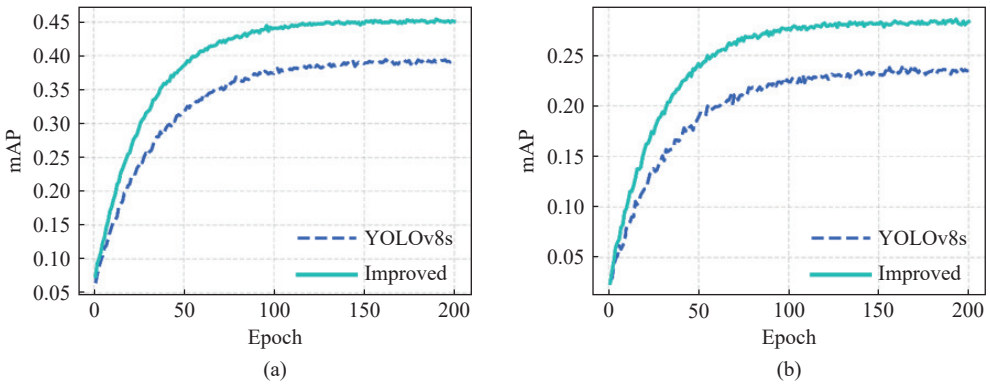
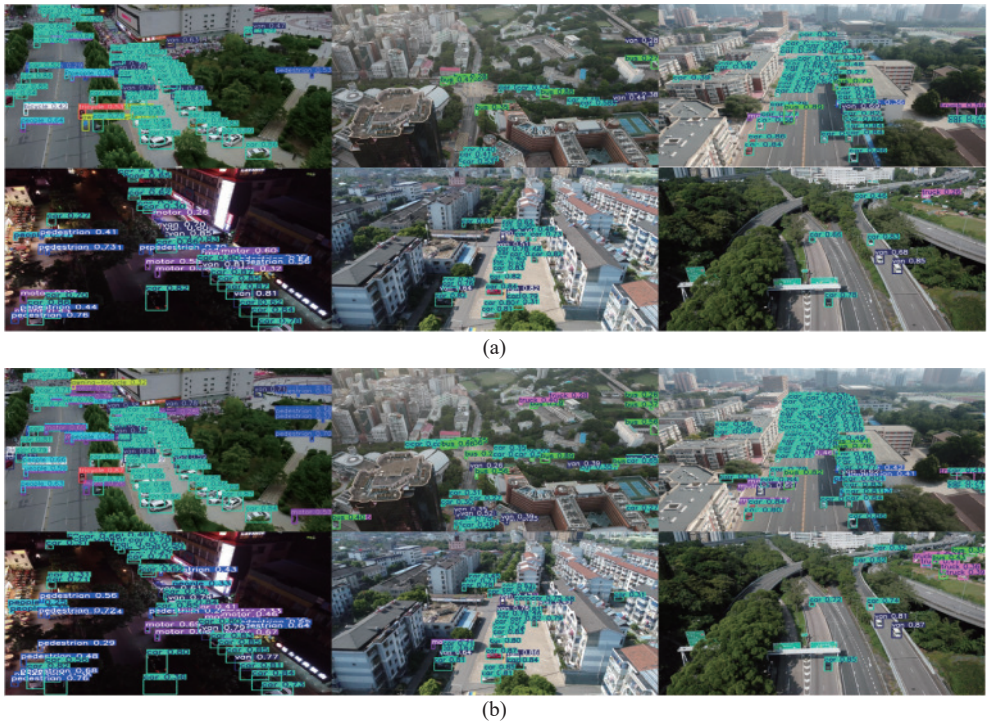


Fig. 8 Comparison curves before and after improvement: (a) mAP0.5; (b) mAP0.5:0.95



(a)

(b)

Fig. 9 Comparison of object detection results on the VisDrone dataset: (a) visualization of detection results for YOLOv8s; (b) visualization of detection results for the improved model

art (SOTA) object detection models. The baseline models include the two-stage detector faster R-CNN, RetinaNet with focus loss mechanism, IterDet based on iterative optimization strategy, and ARSS integrating recurrent subsampling and attention mechanisms. Additionally, the comparison encompasses classic YOLO family models (e.g., YOLOv5 and YOLOv7) and their enhanced variants (including BDH-YOLO and Drone-YOLO (small)). Given the recent success of transformer architectures across visual tasks, transformer-based detectors such as DETR and Swin-transformer are also evaluated.

Tab. 3 summarizes the performance of each model using mean average precision (mAP). Compared to other methods, our algorithm demonstrates superior accuracy, with particularly significant advantages among models of similar parameter scales. Notably, our method achieves 45.1% mAP@0.5, surpassing other lightweight YOLO variants. These results demonstrate our method's significant advantages in drone object detection scenarios. Additionally, Fig. 10 plots a bar chart comparing the mAP performance of all models based on the data in Tab. 3.

Tab. 3 Comparison of Models on VDrone Dataset

Modesls	mAP@0.5	mAP@0.5:0.95	Params (MB)
FasterR-CNN	35.6	19.4	41.1
RetinaNet	26.2	15.7	57.0
IterDet	36.8	20.4	----
ARSS	36.4	23.1	----
YOLOv5s	39.4	23.6	9.1
YOLOv7	40.9	22.3	37.2
YOLOv8s	39.1	23.5	11.1
BDH-YOLO	42.9	26.2	9.39
Drone-YOLO	44.3	27.0	10.9
SwinTransformer	42.5	23.1	38.6
ALDNet	42.8	26.1	15.1
RT-DETR R18	44.8	27.7	20
FALDet (ours)	45.1	28.3	10.3

Fig. 11 presents a comprehensive visualization of evaluation metrics for the proposed FALDet algorithm on the VisDrone dataset. Subfigures (a) to (d) respectively display precision-confidence curves, recall-confidence curves, precision-recall (*P*-*R*) curves, and F1-confidence curves, covering performance across all ten object categories and their category-averaged metrics.

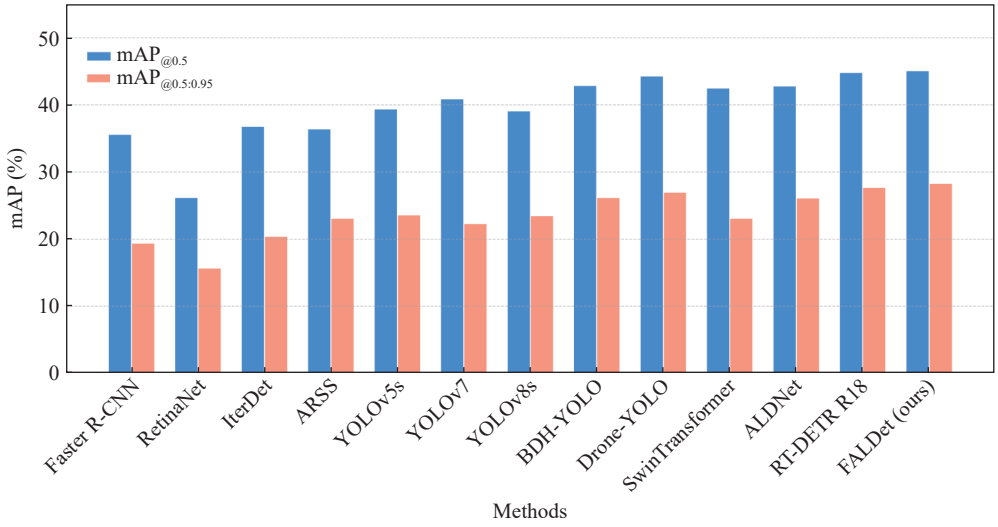


Fig. 10 Bar charts visualizing mAP50 and mAP50:95 accuracy for different algorithms

In Fig. 11(a), the model maintains high accuracy across a broad range of confidence thresholds, particularly excelling in categories such as buses, canopy tricycles, and tricycles. Fig. 11(b) indicates that recall remains relatively stable but decreases as confidence increases, a characteristic behavior attributable to the stricter filtering mechanism.

The *P*-*R* curve in Fig. 11(c) illustrates the trade-off between accuracy and recall, where the area under the curve (AUC) is significantly larger for the bus and canopy tricycle categories. The F1-confidence curve shown in Fig. 11(d) reveals that the F1 score peaks at a confidence threshold of 0.215, with an average value of 0.48 across all categories.

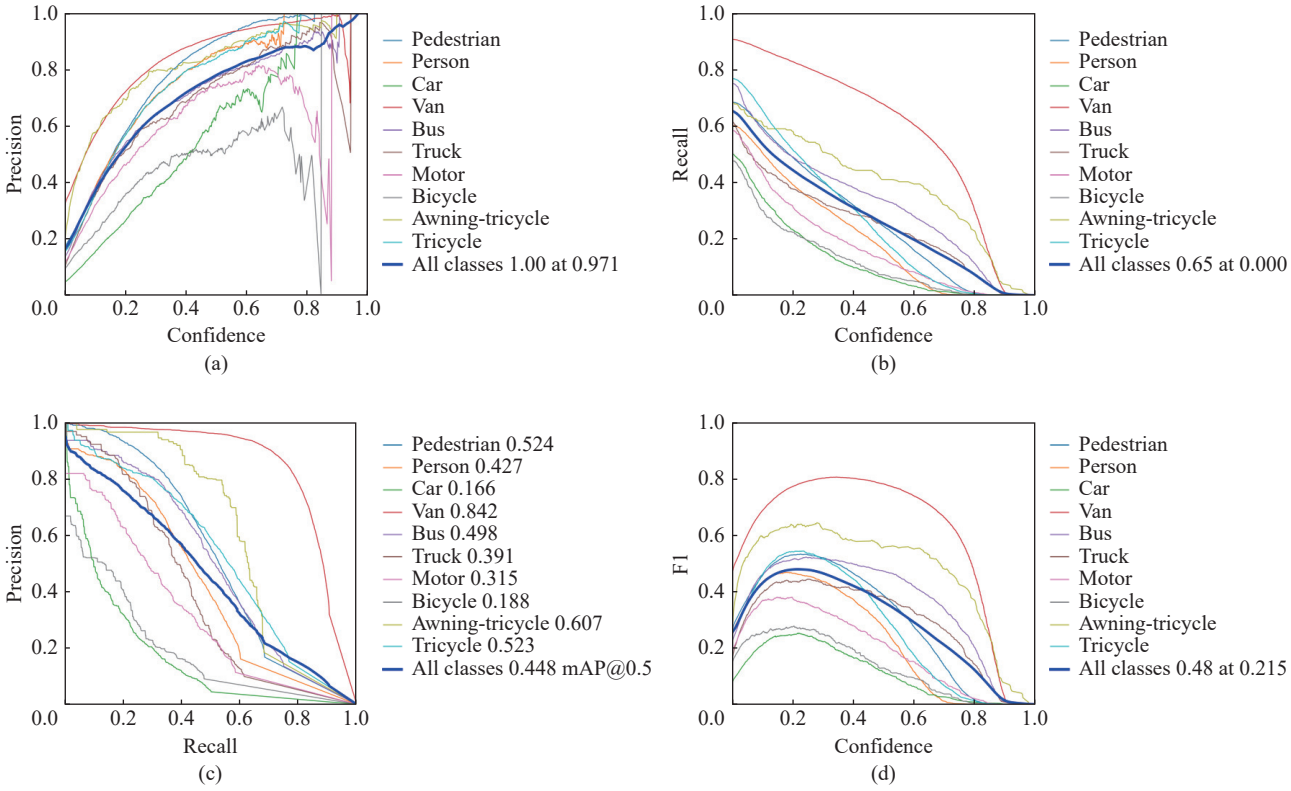


Fig. 11 Visualization of evaluation parameters: (a) *P*-curve; (b) *R*-curve; (c) PR-curve; (d) F1-curve

These visualizations confirm that the improved model not only achieves superior precision and recall on key object categories but also maintains stable detection performance across different confidence thresholds, validating its robustness and generalization capability in drone aerial photography scenarios.

4.5 Validation on Other Datasets

To further evaluate the robustness and generalization capability of the proposed improved algorithm, the representative remote sensing image dataset SIMD (satellite image mini-dataset) was selected for empirical validation. This dataset focuses on typical small object detection tasks in remote sensing images, covering multiple categories with challenging features such as small object sizes, complex backgrounds, and severe occlusions. It comprehensively assesses the adaptability of object detection algorithms in real-world remote sensing scenarios.

During experimentation, the dataset was divided into training and testing sets at a reasonable ratio. The original YOLOv8s model and the improved model were evaluated against each other. Evaluation metrics included mAP@0.5 and mAP@0.5:0.95. Detection results are shown in Tab. 4. Results demonstrate that the proposed algorithm maintains high detection accuracy on this dataset, exhibiting particular robustness in scenarios with dense small objects and strong background interference.

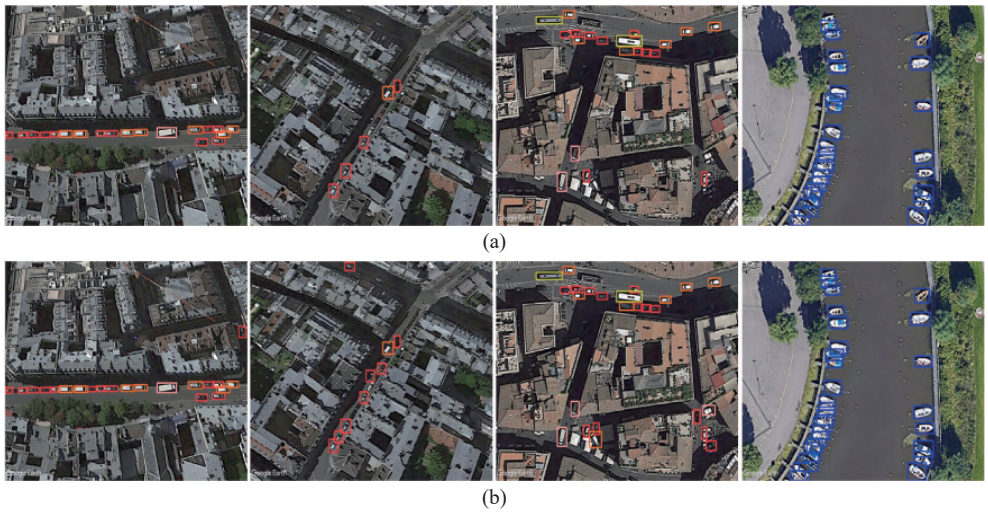
Tab. 4 Comparison of different models on simd dataset

Method	Params	mAP@0.5	mAP@0.5:0.95	F1-scoer
YOLOv11s	9.39	79.8	63.8	0.788
YOLOv10s	8.02	77.4	60.5	0.735
YOLOv8s	11.09	79.1	63.0	0.753
YOLOv7-Tiny	6.01	79.2	60.8	0.764
RTDETR+1	32.8	78.4	56.8	0.742
ReinaNet	586.75	76.1	58.5	-----
FRCNN	28.45	74.3	48.6	-----
YOLOv10-n	2.58	78.9	62.7	0.753
RT-DETR-R18	20	78.4	63.6	0.745
FALet (ours)	23.4	81.9	66.7	0.781

Furthermore, Fig. 12 presents visualizations of detection results on selected test images. Comparisons reveal that the improved model accurately identifies object boundaries across multiple categories while effectively suppressing false positives and false negatives, further validating its generalization capability and practical advantages for small object detection in remote sensing applications.

5 Concusion

This paper proposes FALDet, an improved UAV small object detector based on YOLOv8, addressing challenges like small object size, occlusion, and deformation. By integrating AIFI-HiLo, LA-



(a)

(b)

Fig. 12 Comparisons of object detection results on the SIMD dataset: (a) YOLOv8s; (b) the improved model

DCN, and LiteX-DyHead modules, the model enhances detail representation, geometric alignment, and multi-scale fusion. Experiments show a 5.8% mAP@0.5 gain over YOLOv8s on VisDrone, with strong generalization on SIMD and real-time performance. Future work will focus on adapting to multi-scale targets and optimizing deployment on resource-constrained platforms.

References:

[1] G. Zhang, S. Li, W. Li, M. Wang, M. A. Pagnutti, R. E. Ryan, and V. G. Holekamp, "UAV-based object detection: A survey and benchmark," *Remote Sensing*, vol. 13, no. 13, pp. 2564, 2021.

[2] Y. Chen, J. Wang, X. Chen, Y. Zhang, Z. Liu, S. Li, and L. Wang,"A survey of UAV applications in civil infrastructure," *Automation in Construction*, vol. 134, pp. 104081, 2022.

[3] Y. Liu, Y. Wang, X. Zhang, S. Chen, T. Lin, H. Zhang, and L. Wu,"Recent advances in drone vision: From detection to understanding," *Pattern Recognition*, vol. 135, pp. 109101, 2023.

[4] 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 (CVPR)*, pp. 779-788, 2016.

[5] H. Zhu, L. Wen, D. Du, X. Bian, H. Ling, Q. Hu, H. Peng, and Q. Nie, "VisDrone: A large-scale benchmark for object detection in UAV images, " in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 200-210, 2018.

[6] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, and Y. Zhang, "The unmanned aerial vehicle benchmark: Object detection and tracking, " in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 300-315, 2020.

[7] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection, " *arXiv preprint*, arXiv: 2004.10934, 2020.

[8] G. Jocher, A. Chaurasia, Q. Jing, W. Qiu, J. Fang, and Z. Guo, "Ultralytics YOLOv8: Official release and documentation, " 2023. https://docs.ultralytics.com/.

[9] C. Y. Wang, A. Bochkovskiy, H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, " in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7462-7471, 2023.

[10] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks, " in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132-7141, 2018.

[11] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module, " in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3-19, 2018.

[12] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, "ECA-net: Efficient channel attention for deep convolutional neural networks, " in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11534-11542, 2020.

[13] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks, " in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 510-519, 2019.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows, " in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012-10022, 2021.

[15] R. Li, K. Wang, W. Zuo, R. Timofte, L. Zhang, and L. Zhang,"HiLo attention: Efficient attention for vision transformers," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 480-492, 2022.

[16] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks, " in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 764-773, 2017.

[17] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results, " in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9308-9316, 2019.

[18] T. Lin, C. Wang, J. Liu, S. Chen, T. Y. Lin, K. He, and P. Dollár,"Deformable kernel networks with affine sampling," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 1150-1160, 2021.

[19] S. Chen, Z. Tan, J. Tao, Y. Bin, X. Li, and X. Li,"AffineConv: Towards lightweight and stable geometric modeling," *in Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, pp. 400-408, 2022.

[20] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection, " in *Proceedings of the IEEE Conference on Computer Vision and Pat-*

tern Recognition (CVPR), pp. 2117-2125, 2017.

[21] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, "Path aggregation network for instance segmentation, " in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8759-8768, 2018.

[22] X. Dai, Y. Chen, B. Xiao, D. Chen, M. Liu, L. Yuan, and L. Zhang, "Dynamic head: Unifying object detection heads with attentions, " in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7373-7382, 2021.

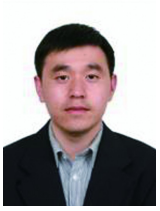
[23] H. Zhang, Y. Wang, F. Dayoub, and N. Sunderhauf, "VarifocalNet: An IoU-aware dense object detector, " in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8514-8523, 2021.

[24] S. Kim, J. Lee, K. Park, and J. Gyeong, "Probabilistic anchor assignment with IoU prediction, " in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 353-367, 2020.

[25] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers, " in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 213-229, 2020.

[26] Y. Xu, J. T. Vogelstein, M. J. Mueller, et al, "RT-DETR: Real-time DETR with efficient attention, " in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1250-1260, 2023.

Junming Gao was born in March 2000. He received the bachelor's degree from Zhengzhou University of Light Industry, in 2023. He is currently pursuing the master's degree with Zhengzhou University of Aeronautics. His primary research focus lies in image processing and computer vision.



Yanshan Zhang was born in Xinxiang, Henan, China, in 1986. He received the B.S. degree in mathematics from Zhengzhou University of Aeronautics, Zhengzhou, China, in 2010. And the Ph.D. degree from Beijing Institute of Technology, Beijing, China, in 2017. Presently, he is an Associate Professor with School of Computer Science, Zhengzhou University of Aeronautics. His research interests include artificial intelligence, digital image processing, fractional calculus.

Yuanzhang Fan received his Ph.D. degree from Beijing Institute of Technology in 2017. Now he is a lecturer and master's supervisor in the School of Computer Science of Zhengzhou University of Aeronautics and Astronautics. His main research interests are environment sensing, automatic driving, and remote sensing image processing.

Bao Tian received his Ph.D. degree from Southwest Petroleum University in 2020. Now he is a lecturer and master's tutor in the School of Computer Science, Zhengzhou University of Aeronautics and Astronautics. His main research interests are machine learning, generative artificial intelligence, etc.

Yinhui Xu was born in Qingdao, Shandong, China, in 1986. He received his B.S. degree in Information Engineering from Beijing Institute of Technology, Beijing, China, in 2009, and his Ph.D. degree from the same university in 2016. Currently, he is an engineer at the Zhengzhou Research Institute, Beijing Institute of Technology, mainly engaged in research on real-time radar signal processing and high-precision echo simulation.