Digit. Signal Process. 174 (2026) 105952

Contents lists available at ScienceDirect

Digital Signal Processing

journal homepage: www.elsevier.com/locate/dsp

EMFNet: An efficient multi-scale fusion network for UAV small object detection

Mingquan Wang [a,b], Huiying Xu [a,b,*], Yiming Sun [c], Hongbo Li [d], Zeyu Wang [a,b], Yi Li [a,b], Ruidong Wang [a,b,e], Xinzhong Zhu [a,b]

[a] Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, 321004, Jinhua Zhejiang, China
[b] School of Computer Science and Technology, Zhejiang Normal University, 321004, Jinhua Zhejiang, China
[c] School of Automation, Southeast University, 210096, Nanjing Jiangsu, China
[d] Beijing Geekplus Technology Co, 100101, Ltd Beijing, China
[e] School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, 210044, Nanjing Jiangsu, China

ARTICLE INFO

Keywords:
Object detection
Small object
Unmanned aerial vehicles
Feature extraction

ABSTRACT

Object detection in UAV aerial images holds significant application value in traffic monitoring, precision agriculture, and other fields. However, this task faces numerous challenges, including significant variations in object sizes, complex background interference, high object density, and class imbalance. Additionally, processing high-resolution aerial images involves disturbances such as uneven lighting and weather variations. To address these challenges, we propose an EMFNet model. This model effectively solves the problems in object detection in drone aerial images by enhancing the response to object areas under different lighting and weather conditions, suppressing interference from complex backgrounds, and improving adaptability to changes in image object size. Specifically, firstly, the lightweight vision transformer architecture RepViT is innovatively used as the backbone of EMFNet, combined with **D**ual **C**ross-Stage **P**artial **A**ttention (**DCPA**) to optimize multi-scale feature fusion and background suppression, thereby enhancing small object feature extraction under varying lighting and weather conditions. Second, we propose the **C**ontext **G**uided **D**ownsample **B**lock (**CGDB**) to improve the downsampling process and mitigate feature information loss. Finally, the DyHead detection head utilizing the three-level attention mechanism receives three appropriately located prediction heads for classification and localization, thus improving the detection accuracy of dense and rare objects. Experiments on the VisDrone and UAVDT datasets demonstrate that EMFNet, with 6.76M parameters, achieves AP improvements of 7.5% and 15.2% over the baseline models, respectively.

1. Introduction

Unmanned Aerial Vehicles (UAVs) have been widely applied in diverse sectors such as traffic monitoring, precision agriculture, and infrastructure inspection due to their advantages of high mobility, low cost, and scalable deployment. Unlike ground-based surveillance, UAV-based aerial imagery is characterized by a high bird's-eye view, covering extensive geographic areas. However, this broad perspective introduces inherent challenges: images often exhibit high spatial resolution but contain objects of extremely small scales relative to the background. Furthermore, UAVs frequently operate in unstructured environments, leading to unpredictable variations in lighting, weather conditions, and viewing angles. These factors, combined with uneven object distribution and severe background clutter, constitute significant bottlenecks in current UAV-based object detection technology.

Deep learning techniques, particularly Convolutional Neural Networks (CNNs) [1–3], have revolutionized computer vision by replacing traditional hand-crafted features with self-learned hierarchical representations. These networks can adaptively learn global semantic information and local details, thereby alleviating the difficulties of joint optimization in object localization and classification. Despite these advancements, applying generic CNN-based detectors directly to UAV aerial imagery often yields suboptimal results due to the domain gap between natural scene images such as MS COCO and PASCAL VOC and aerial views.

Existing detection frameworks primarily fall into two technical paradigms: single-stage and two-stage detectors. Two-stage models, represented by architectures like Faster R-CNN [4], employ a Region Proposal Network (RPN) to generate candidate boxes followed by a refinement stage. While they typically offer superior detection accuracy,

* Corresponding author.
E-mail addresses: 2359720950@qq.com (M. Wang), xhy@zjnu.edu.cn (H. Xu), sunyiming@seu.edu.cn (Y. Sun), Jason.li@geekplus.com (H. Li), 14797857499@163.com (Z. Wang), leeyee@zjnu.edu.cn (Y. Li), iswangrd@zjun.edu.cn (R. Wang), zxz@zjnu.edu.cn (X. Zhu).

https://doi.org/10.1016/j.dsp.2026.105952

Available online 27 January 2026

especially for small objects, their complex multi-stage pipeline results in substantial computational overhead and high latency. This makes them difficult to deploy on resource-constrained UAV edge devices that demand real-time feedback. Conversely, single-stage detectors, exemplified by the YOLO series and SSD [5], treat detection as a regression problem, achieving faster inference speeds. However, they historically struggle with the "small object dilemma": as the network depth increases to capture semantic information, the spatial resolution of feature maps decreases due to successive downsampling operations. Consequently, the feature representations of tiny objects-often occupying only a few pixels-are prone to vanishing, leading to high missed detection rates.

In the specific context of UAV surveillance, these limitations are exacerbated. First, standard downsampling mechanisms like strided convolution or max pooling often discard critical spatial details required to distinguish small targets from complex backgrounds. Second, conventional backbone networks lack the dynamic adaptability to handle the drastic intra-class variance caused by changing weather conditions, including fog and strong sunlight, as well as diverse flying altitudes. Third, dense object occlusion in scenarios such as traffic congestion confuses the detection head, resulting in inaccurate bounding box regression.

To address these challenges effectively, we propose EMFNet, an efficient multi-scale fusion network tailored for UAV small object detection. We innovate the architecture from three dimensions: backbone, neck, and head. Firstly, acknowledging the need for a lightweight yet powerful feature extractor, we adopt the RepViT architecture as the backbone. By integrating a dynamic feature optimization mechanism and cross-block Squeeze-and-Excitation (SE) [6] modules, RepViT enhances the model's sensitivity to object areas under varying lighting conditions while maintaining a compact parameter size. Secondly, to tackle the information loss during feature pyramid construction, we propose two novel modules in the neck: the Dual Cross-Stage Partial Attention (DCPA) and the Context Guided Downsample Block (CGDB). The DCPA module captures both local details and global spatial contexts to suppress background interference, while the CGDB replaces standard downsampling layers to preserve fine-grained features of small objects. Finally, we redesign the detection head by introducing a Dynamic Head (DyHead) equipped with a three-level attention mechanism comprising scale-aware, spatial-aware, and task-aware components. This allows the model to selectively focus on optimally positioned prediction heads, significantly improving localization accuracy for dense and rare objects.

The main contributions of this paper are summarized as follows:

1. We integrate RepViT as the backbone to balance computational efficiency with robust feature extraction, enabling the model to adapt to small objects under diverse lighting and weather conditions.
2. We propose the DCPA and CGDB modules. DCPA optimizes multi-scale fusion and suppresses background noise via attention mechanisms, while CGDB minimizes feature information loss during downsampling.
3. We innovate the detection head by incorporating DyHead with a cascaded attention architecture. By selectively feeding three optimally positioned prediction heads, we enhance the model's adaptability to scale variations and dense occlusions.
4. Experiments on VisDrone and UAVDT datasets show EMFNet achieves AP improvements of 7.5% and 15.2% over baselines, outperforming state-of-the-art models with only 6.76M parameters.

2. Related work

2.1. Traditional object detection

Deep learning-driven object detection has emerged as a pivotal technology in computer vision, significantly surpassing traditional manual feature extraction methods in complex scenarios. CNNs excel at automatically extracting high-level semantic features from raw pixels, facilitating accurate object localization and identification. These frameworks are broadly categorized into two-stage and single-stage paradigms. Two-stage detectors, such as R-CNN [7], Fast R-CNN [8], and Faster R-CNN [4], prioritize higher precision by generating candidate regions before refinement. However, their computational intensity often renders them unsuitable for real-time UAV applications. Conversely, single-stage detectors, exemplified by SSD [5], RetinaNet [9], and the YOLO series [10–15], offer rapid inference by predicting bounding boxes and class probabilities in a single forward pass, though they historically struggled with dense small objects due to the lack of fine-grained feature preservation.

Since its inception, the YOLO framework has evolved continuously to balance speed and accuracy. YOLOv8 [16] represents a significant advancement by incorporating an anchor-free head and C2f modules to enhance feature extraction capabilities. Subsequent iterations, including YOLOv9 [17], YOLOv10 [18], and YOLOv11 [19], have introduced progressive improvements in gradient optimization and semantic understanding, while the recent YOLOv12 [20] focuses on attention efficiency to further refine the trade-off between performance and latency. Beyond these CNN-based architectures, the Real-Time DEtection TRansformer (RT-DETR) [21] has emerged as a powerful contender, leveraging a hybrid encoder and efficient attention mechanisms to eliminate the need for Non-Maximum Suppression (NMS), thereby achieving competitive inference speeds with high accuracy.

However, despite these architectural breakthroughs in generic object detection, applying these models directly to UAV aerial imagery often yields suboptimal results. A critical gap remains: these detectors, primarily optimized for natural scene images such as MS COCO, frequently struggle with the inherent challenges of aerial surveillance, such as weak global context modeling for tiny objects, drastic scale variations, and complex environmental interference.

2.2. UAV object detection

Recent UAV image object detection studies have seen notable advancements. Chen et al. [22] proposed the Shallow & Deep Fovea Vision Block, a structure that mimics the double fovea structure of an eagle, enabling efficient feature extraction through a large receptive field. Liao et al. [23] incorporated a MHSA-darknet module into the backbone to better handle scale variance, and a specialized detection head to improve the recognition of dense small objects. Du et al. [24] introduced CEASC, optimizing detection heads via sparse convolution to balance accuracy and resource use. Chen et al. [25] created an RGB-IR fusion framework with adaptive feature alignment, enhancing cross-modal consistency. Tian et al. [26] developed MECA and MFR to boost feature representation and fusion. Zhou et al. [27] designed UAVDNet, leveraging multi-source interaction for better feature fusion. Zhang et al. [28] proposed UAV-DETR, utilizing Transformers for improved multi-path fusion. Huang et al. [29] focused on feature enhancement to reduce detection errors. Liu et al. [30] improved YOLOv8s-P2 with slice assistance, while Huang et al. [31] addressed small-object detection inefficiency. Li et al. [32] introduced density map modeling for clustered objects. Wang et al. enhanced UAV-YOLOv8 with feature fusion and attention mechanisms. Zeng et al. [33] proposed YOLOv7-UAV with improved feature capture. LIU et al. [34] bolstered YOLOv5 with shallow feature detection and FEBlock. Li et al. [35] optimized YOLOv5s for better occluded object detection. Zhang et al. [36] enhanced model robustness with multi-source feature alignment. Similarly, Bakirci [37,38] leveraged diverse multi-altitude datasets to validate YOLOv8 in ITS, demonstrating that the C2f module, decoupled head, and DIoU loss optimizations significantly enhance detection performance across various illumination and high-density traffic scenarios. From different perspectives, these researches are dedicated to improving the accuracy and efficiency of object detection in UAV aerial photography scenarios by optimizing feature extraction, feature fusion, and network architecture, providing new ideas and methods for the development of this field.

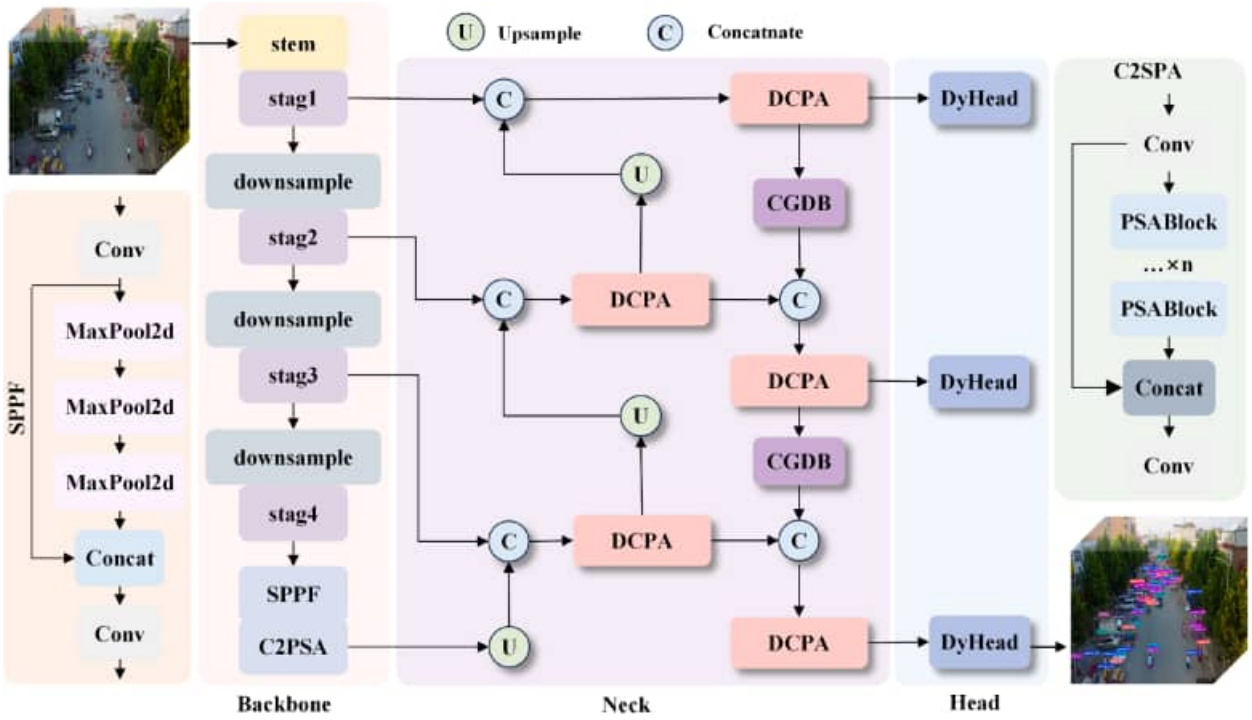

Fig. 1. Overall architecture of proposed EMFNet.

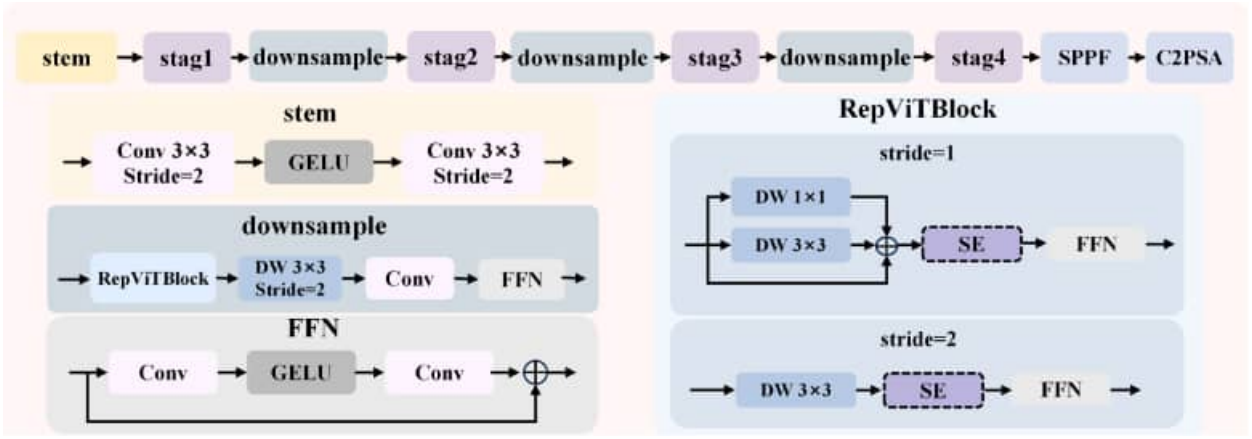

Fig. 2. The RepViT backbone architecture is at the top. Beneath it are components like the stem, downsample, FFN, and RepVitBlock. Notably, the RepVitBlock is available in two versions: one with a stride of 1 Notably, the RepVitBlock is available in two versions: one with a stride of 1 and another with a stride of 2.

3. Method

3.1. Overview

To address the challenges posed by UAV aerial images we propose EMFNet, whose structure is shown in Fig. 1. EMFNet is optimized as follows: first, it innovates RepViT as the backbone network to enhance small object feature extraction capabilities; second, it proposes DCPA to further improve multi-scale feature fusion; third, it proposes CGDB as the downsampling method to reduce feature information loss; finally, it adopts the DyHead detection head to process prediction heads P_2, P_3, and P_4, significantly improving both small object detection accuracy and overall model robustness.

3.2. Lightweight vision transformer backbone

Standard backbones often struggle to capture dense, small objects or handle occlusions in UAV imagery, while also lacking adaptability to environmental interferences such as noise and illumination changes. To address these limitations while balancing efficiency, we innovated RepViT [39] as the backbone for EMFNet, with its structure shown in the Fig. 2. RepViT leverages structural reparameterization to combine the representational power of multi-branch topologies during training with the speed of a single-path architecture during inference. This design, augmented by Cross-block Squeeze-and-Excitation (SE) modules [6], dynamically recalibrates channel weights to enhance feature responsiveness to small and occluded objects.

The architecture begins with a Stem module optimized for inference stability and low latency. Replacing standard patchify operations, it utilizes two stacked 3×3 convolutions (stride = 2) to progressively expand input channels to 48:

$$x_{stem} = \mathrm{Conv}^{C=24}_{3\times3,\mathrm{stride}=2}\left(\mathrm{Conv}^{C=48}_{3\times3,\mathrm{stride}=2}(x_{input})\right) \tag{1}$$

Following the Stem, the network extracts hierarchical features across four stages with resolutions ranging from $H/4$ to $H/32$. The core

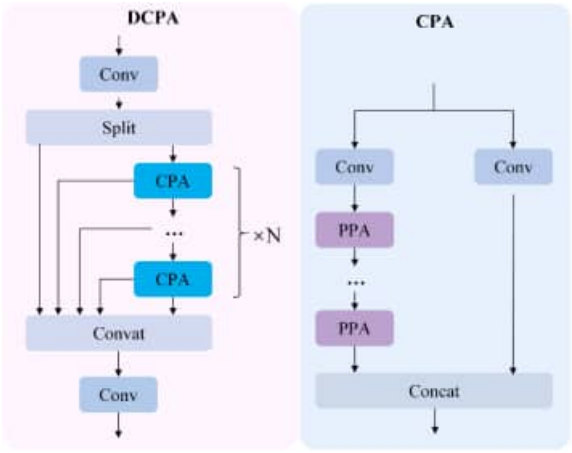

Fig. 3. On the left is the architecture of the DCPA module, and on the right is the architecture of the CPA module.

RepViTBlocks decouple token and channel mixing. During inference, the parallel branches-Depthwise Separable Convolution (DWConv), Pointwise Convolution, and Identity-are mathematically merged into a single kernel to eliminate computational redundancy. The reparameterized weights are calculated as:

$$W_{final} = W_{\text{DWConv}} + W_{1\times1} + W_{identity} \tag{2}$$

To minimize information loss during resolution reduction, we employ a sandwich-style downsampling layer. This layer preserves multi-scale contextual information by sandwiching a 1×1 channel-adjusting convolution between a DWConv (stride = 2) and a Feed-Forward Network (FFN), thereby improving robustness against complex backgrounds:

$$x_{down} = \text{FFN}\left(\text{Conv}_{1\times1}\left(\text{DWConv}_{3\times3,\text{stride}=2}(x)\right)\right) \tag{3}$$

3.3. Dual cross-stage partial attention

The traditional C3k2 [19] bottleneck, constrained by fixed 3×3 kernels and channel compression, is ill-suited for the drastic scale variations and complex backgrounds typical of UAV imagery. This design often leads to the loss of high-frequency details and feature over-smoothing during compression, significantly reducing the signal-to-noise ratio for small objects. Furthermore, the absence of cross-level fusion and attention mechanisms hinders the model from capturing global context or adaptively focusing on critical regions, resulting in high false detection rates in complex scenarios.

To overcome these limitations, we propose an efficient feature extraction architecture named Dual Cross-Stage Partial Attention (DCPA), illustrated in Fig. 3. The core of DCPA lies in the cascaded optimization of the Parallelized Patch-Aware Attention (PPA) module [40], whose structure is detailed in Fig. 4. Specifically, given an input feature $x \in \mathbb{R}^{H\times W\times C}$, DCPA first applies a 1×1 convolution to generate a baseline feature map $x_{\text{skip}} = \text{Conv}(x)$. Subsequently, the features undergo multi-branch fusion and enhancement through parallel PPA modules, which integrate local-global attention and multi-level convolutions. The final output is generated via residual connections as follows:

$$x_{out}^{(i)} = \text{PPA}(x_{in}^{(i)}) + x_{in}^{(i)} \tag{4}$$

The PPA module extracts multi-scale features through three parallel branches: a local branch (p = 2), a global branch (p = 4), and a cascaded convolution branch. For the input feature tensor $\mathbf{F} \in \mathbb{R}^{H'\times W'\times C}$, it first adjusts the channel dimension through pointwise convolution to obtain $\mathbf{F}' \in \mathbb{R}^{H'\times W'\times C'}$. The local and global branches utilize unfold and reshape operations to partition $\mathbf{F}'$ into non-overlapping patches ($p \times p, H'/p, W'/p, C'$), then dynamically weight relevant features through channel averaging, linear transformation, and feature selection mechanisms:

$$\hat{t}_i = P \cdot \sin(t_i, \xi) \cdot t_i \tag{5}$$

where $\xi \in \mathbb{R}^{C'}$ is the task embedding vector, $P \in \mathbb{R}^{C'\times C'}$ is the learnable parameter matrix, and $\sin(\cdot,\cdot)$ denotes the cosine similarity function. The cascaded convolution branch generates multi-level features $F_{\text{conv1}}, F_{\text{conv2}}$, and F_{conv3} through three consecutive 3×3 convolutional layers, which are then fused through summation to produce F_{conv}. The three branch features are added to obtain the preliminary fused feature $\hat{F}$, which is further optimized through efficient channel attention module [41] and spatial attention module [42]:

$$\mathbf{F}_c = \mathbf{M}_c(\hat{\mathbf{F}}) \otimes \hat{\mathbf{F}}, \mathbf{F}_s = \mathbf{M}_s(\mathbf{F}_c) \otimes \mathbf{F}_c \tag{6}$$

where $\otimes$ denotes element-wise multiplication, $M_c \in \mathbb{R}^{1\times1\times C'}$ and $M_s \in \mathbb{R}^{H'\times W'\times1}$ represent channel and spatial attention weights respectively. The final output undergoes ReLU and batch normalization to obtain $F'' \in \mathbb{R}^{H'\times W'\times C'}$. In DCPA, the outputs of multiple PPA modules are fused into high-dimensional feature representations through channel concatenation and 1×1 convolution:

$$y = \text{Conv}(\text{Concat}(x_{out}^{(1)}, x_{out}^{(2)}, \ldots, x_{out}^{(n)})) \in \mathbb{R}^{H\times W\times C'} \tag{7}$$

This proposed DCPA effectively addresses the multi-scale feature loss problem caused by the single convolutional kernel limitation in traditional C3k2 modules through local-global multi-granularity feature interaction, adaptive attention guidance, and lightweight depthwise separable convolution design. It significantly improves the signal-to-noise ratio and localization accuracy of small objects in complex backgrounds.

3.4. Context guided downsample block

Drawing inspiration from the Context Guided mechanism in CGNet [43], we construct the Context Guided Downsample Block (CGDB), illustrated in Fig. 5, to address the loss of fine-grained details inherent in standard downsampling-a critical issue for small object detection in UAV imagery. Unlike conventional approaches that decouple downsampling from context modeling, CGDB adopts a "Downsample-Context-Refine" paradigm. It integrates context perception directly into the downsampling stage via parallel local and dilated convolutions to balance detail preservation with context awareness, further augmented by a global channel attention mechanism for dynamic feature recalibration.

The workflow proceeds as follows: Given an input feature map $F \in \mathbb{R}^{H\times W\times C}$, we employ a learnable strategy over non-parametric pooling. The local feature extractor $f_{loc}(\cdot)$ utilizes a standard 3×3 convolution to capture intrinsic texture and edge information:

$$F_{\text{loc}} = f_{\text{loc}}(F) = \text{Conv}_{3\times3}(F) \tag{8}$$

Simultaneously, the surrounding context extractor $f_{sur}(\cdot)$ employs a 3×3 dilated convolution (rate r) to expand the receptive field without increasing parameters. This branch compensates for spatial information loss by maintaining the semantic relationship between the object and its environment:

$$F_{\text{sur}} = f_{\text{sur}}(F) = \text{DilatedConv}_{3\times3}^{r}(F) \tag{9}$$

Subsequently, the joint feature extractor $f_{joi}(\cdot)$ concatenates these features, fusing local details with global context through Batch Normalization (BN) and PReLU activation:

$$F_{\text{joi}} = \text{PReLU}(\text{BN}(\text{Concat}(F_{\text{loc}}, F_{\text{sur}}))) \tag{10}$$

A 1×1 convolution then reduces the channel dimension by half to generate compact features $F_{red} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$:

$$F_{\text{red}} = \text{Conv}_{1\times1}^{2C\to C}(F_{\text{joi}}) \tag{11}$$

Finally, in the global context enhancement stage, the extractor $f_{glo}(\cdot)$ generates channel attention weights $W_{glo} \in \mathbb{R}^{1\times1\times C}$ using global average pooling (GAP) followed by a two-layer MLP:

$$W_{\text{glo}} = \sigma(\text{MLP}(\text{GAP}(F_{\text{red}}))) \tag{12}$$

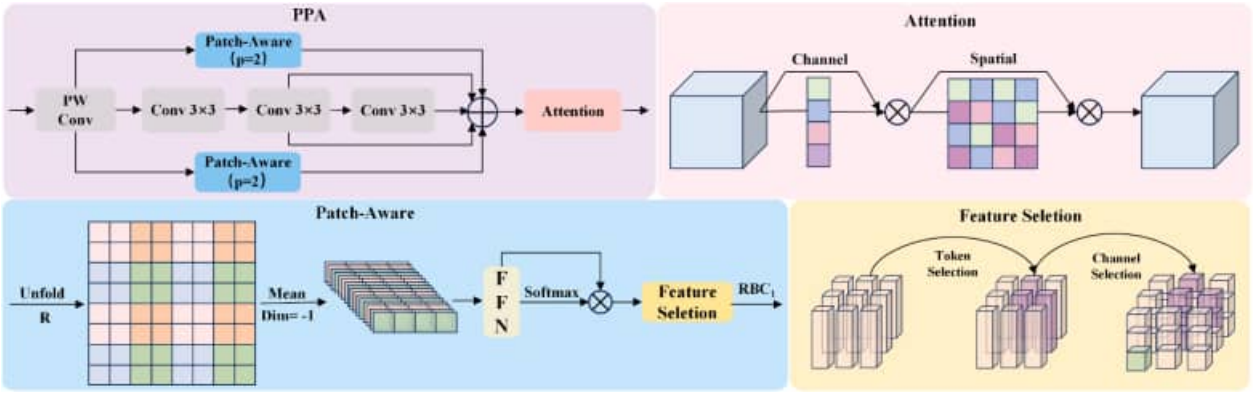

Fig. 4. Above left is the PPA block architecture. The Patch-Aware block and Attention block are the two components of PPA. The Patch-Aware block includes a subpart called Feature Selection.

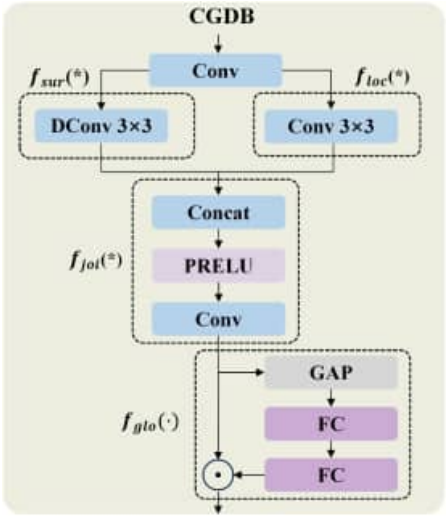

Fig. 5. The architecture of CGDB module.

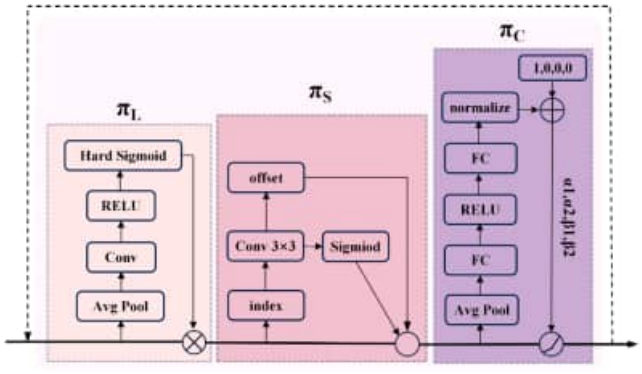

Fig. 6. The architecture of DyHead module.

The final output is obtained by weighting the enhanced features:

$$F' = F_{\text{red}} \odot W_{\text{glo}} \tag{13}$$

3.5. Efficient dynamic detection head

Aerial images are characterized by small, distant objects, severe occlusion, and complex backgrounds, often exacerbating detection difficulties due to extreme scale variations and class imbalance. To overcome the limitations of traditional detection heads in these scenarios, we introduce the DyHead [44], as illustrated in Fig. 6, which integrates three synergistic attention mechanisms. Specifically, Scale-Aware Attention adapts to object size to prevent large objects from overshadowing smaller ones; Spatial-Aware Attention utilizes deformable convolution and sparse sampling to enhance localization in dense, occluded scenes; and Task-Aware Attention dynamically balances task requirements to improve rare category identification. The core modules-scale-aware π_L, spatial-aware π_S, and task-aware π_C-cascade to realize multi-dimensional feature augmentation:

$$W(\mathcal{F}) = \pi_C(\pi_S(\pi_L(\mathcal{F}) \cdot \mathcal{F}) \cdot \mathcal{F}) \cdot \mathcal{F} \tag{14}$$

The scale-aware module π_L addresses the challenge of significant object scale variations by employing a cross-layer feature pyramid interaction strategy, generating dynamic scale weights through global average pooling and lightweight convolution:

$$\pi_L(\mathcal{F}) \cdot \mathcal{F} = \sigma(f(\frac{1}{SC}\sum_{s,c}\mathcal{F})) \cdot \mathcal{F} \tag{15}$$

where $\sigma(\mathbf{x}) = \max(0, \min(1, (\mathrm{x}+1)/2))$ serves as the nonlinear activation function, mitigating feature misalignment issues in multi-scale objects by suppressing extreme response values. The spatial-aware module π_S addresses the deformation and occlusion characteristics of small objects by introducing deformable convolution and sparse sampling mechanisms:

$$\pi_S(\mathcal{F}) \cdot \mathcal{F} = \frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{K} w_{l,k} \cdot \mathcal{F}(l; p_k + \Delta p_k; c) \cdot \Delta m_k \tag{16}$$

Through self-learned spatial offsets Δp_k and importance weights Δm_k, the model can adaptively focus on key regions in low-resolution feature maps, enhancing perception capability for spatial distribution of tiny objects. The task-aware module π_C employs a dynamic threshold mechanism to decouple classification and regression tasks:

$$\begin{aligned}&\pi_C(\mathcal{F}) \cdot \mathcal{F} \\ &= \max\left(\alpha^1(\mathcal{F}) \cdot \mathcal{F}_c + \beta^1(\mathcal{F}), \alpha^2(\mathcal{F}) \cdot \mathcal{F}_c + \beta^2(\mathcal{F})\right)\end{aligned} \tag{17}$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2$ are learnable parameters that dynamically activate along the channel dimension to suppress interference noise between tasks, achieving collaborative optimization of classification and localization objectives. Through a three-stage attention cascade incorporating scale awareness, spatial consistency, and task orientation, this module effectively fuses global contextual information with local details across multi-scale feature maps, significantly enhancing the model's adaptability to densely distributed objects, complex background interference, and extreme scale variations.

4. Experiments

In this section, we present a comprehensive evaluation of the proposed EMFNet on the VisDrone [45] and UAVDT [46] datasets to vali-

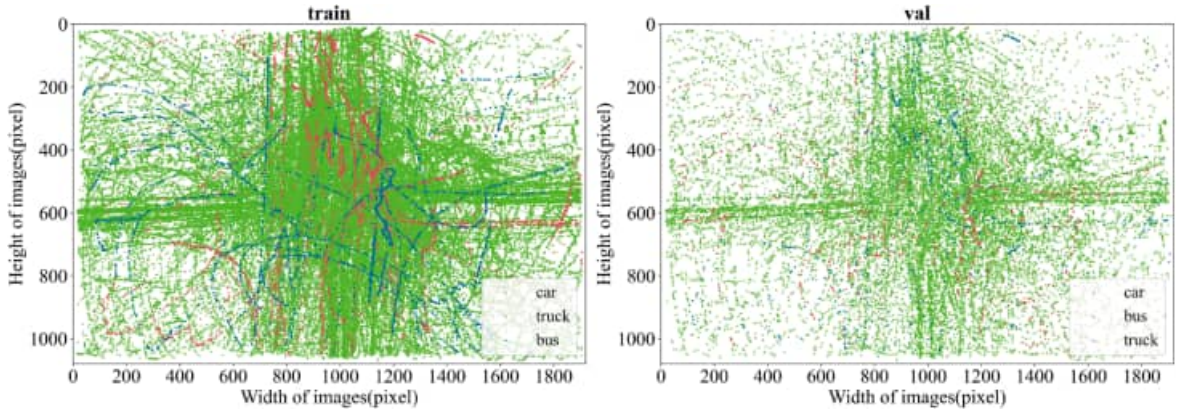

Fig. 7. On the left is the category distribution of the UAVDT training set and on the right is the category distribution of the UAVDT validation set.

date its effectiveness and generalization capability. To ensure a fair and rigorous comparison, particularly regarding inference speed and computational cost, all models included in our experiments were retrained from scratch on our local server using the same hardware environment. The only exception is RT-DETR++ [47], for which we cite the results directly from the original publication due to the unavailability of its source code.

4.1. Dataset

We conducted experiments using two distinct datasets, VisDrone and UAVDT, to validate the generalization capability of our model. The VisDrone dataset is specifically designed for object detection in daily scenes captured by drones. It contains 8899 static images, with 6471 for training, 548 for validation, and 1610 for testing, covering various scenarios such as urban areas, rural regions, and traffic environments, presenting high complexity and challenges. Compared to traditional image datasets, VisDrone features diverse shooting angles, significant scale variations of objects, and substantial influences from occlusion, motion blur, and lighting changes, thereby posing considerable challenges to existing object detection algorithms. Our primary task involves training and prediction on 10 categories within this dataset: pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor.

The UAVDT dataset is a large-scale public dataset specifically designed for object detection and tracking tasks from a drone's perspective. We randomly selected 7433 images from the UAVDT training set and divided them into training and validation sets at an 8:2 ratio. This dataset includes three categories: car, truck, and bus, with the class distributions of the divided training and validation sets illustrated in Fig. 7. This dataset is primarily used to evaluate the detection performance of UAVs in complex contexts. Additionally, it provides challenging samples under varying lighting, weather, and occlusion conditions, effectively testing the model's robustness and adaptability in real-world applications.

4.2. Hyperparameter settings

The experiments were conducted on an Ubuntu 22.04.5 LTS system equipped with 128GB RAM and an Intel Xeon(R) CPU E5-2683 V4 processor. The experimental environment was configured as follows: Python 3.10.14, PyTorch 2.2.2, and CUDA 12.1, with all models trained and tested on four NVIDIA GeForce RTX 2080 Ti GPUs. To mitigate the class imbalance issue mentioned in Section 5.1, we employed the Mosaic augmentation strategy during the training phase. By stochastically synthesizing composite samples, this technique serves as an implicit oversampling mechanism for under-represented categories. Furthermore, combined with Mixup and random geometric transformations, this data-centric approach effectively smooths the label distribution and prevents the model from overfitting to majority classes. The detailed hyperparameter settings for the training are presented in Table 1.

Table 1
Hyperparameter configuration details.

Parameter	Value
epoch	300
batch size	32
lr	0.01
imgsz	640
momentum	0.937
weight decay	0.0005
optimizer	SGD
lr schedule	Linear Decay
early stopping	False

4.3. Performance evaluation

In object detection tasks, evaluating model performance is crucial. This paper employs evaluation metrics proposed by MS COCO [48] to measure the effectiveness of the proposed method, including AP, AP_{50}, AP_{75}, AP_{small}, AP_{medium}, AP_{large} and parameters. AP refers to the average precision calculated with IoU thresholds ranging from 0.50 to 0.95 (in steps of 0.05). AP_{50} and AP_{75} represent the average precision at IoU thresholds of 0.5 and 0.75, respectively. MS COCO categorizes objects into three size groups: small objects (area $< 32 \times 32$), medium objects (area between 32×32 and 96×96), and large objects (area $> 96 \times 96$), with AP calculated separately for each category.

To further ensure the reliability of our experimental results and verify that the improvements are not due to random fluctuations, we conducted a statistical significance analysis. Specifically, we performed 5 independent training runs for both the Baseline and EMFNet on the VisDrone and UAVDT validation sets using different random seeds. On the VisDrone dataset, EMFNet achieved an average AP of 25.5% (±0.3%), compared to 18.0% (±0.4%) for the Baseline. Similarly, on the UAVDT dataset, EMFNet demonstrated superior stability with an average AP of 73.4% (±0.2%), significantly outperforming the Baseline's 58.2% (±0.5%). Paired t-tests conducted on both datasets yielded p-values < 0.01, confirming that the performance improvements of EMFNet are statistically significant at the 99% confidence level across different scenarios.

4.4. Ablation experiment

We conducted ablation experiments on the VisDrone dataset to validate the impact of each module on the baseline model and investigate

Table 2
Ablation experiments on the VisDrone dataset, P_x indicates that P_x prediction heads were used, and a $\surd$ below the module indicates that it was used.

Model	*RepViT*	*DCPA*	*CGDB*	*DyHead*	*AP*	AP_{50}	AP_{75}	AP_{small}	AP_{medium}	AP_{large}	*Paramters*
Baseline					18	31.1	17.9	9.6	28	34.7	**2.58**
$\{P_3, P_4, P_5\}$	$\surd$				20.3	35	20.5	11.4	31.3	41.7	6.43
$\{P_3, P_4, P_5\}$		$\surd$			18.2	31.4	18.3	9.8	28	37.6	2.99
$\{P_3, P_4, P_5\}$			$\surd$		18.2	31.6	18	9.6	28.1	36.2	2.99
$\{P_3, P_4, P_5\}$				$\surd$	19.8	33.7	20	10.6	30.7	39.4	3.1
$\{P_3, P_4, P_5\}$	$\surd$	$\surd$			20.5	35.1	20.7	11.7	31.4	43	6.78
$\{P_3, P_4, P_5\}$	$\surd$	$\surd$	$\surd$	$\surd$	21.7	36.9	21.7	12.3	33.2	40.8	7.71
$\{P_2, P_3, P_4\}$	$\surd$	$\surd$	$\surd$	$\surd$	**25.5**	**42.6**	**25.9**	**17**	**36**	**41.8**	6.76

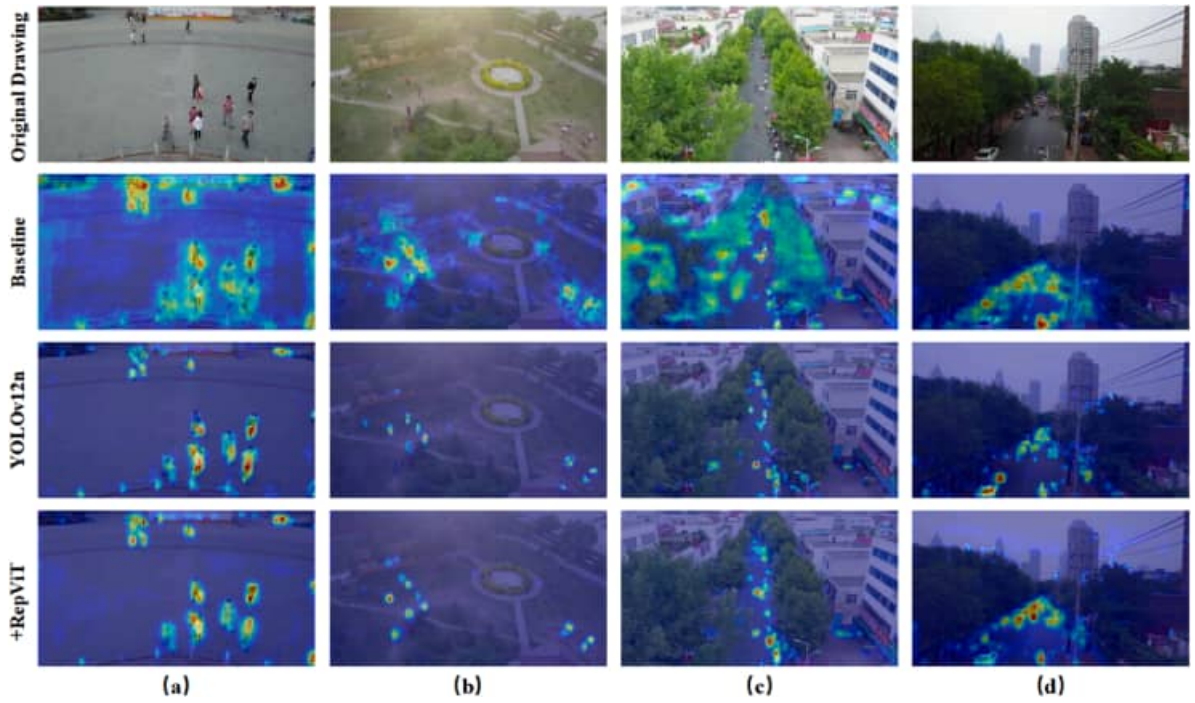

Fig. 8. The heatmaps on Visdrone, (a), (b), (c), and (d) belong to different scenarios, and the fourth row represents the introduction of RepVit on the baseline model.

the combined effects of module integration, with detailed results presented in Table 2.

4.4.1. Ablation experiments with RepViT

We innovated RepViT as the backbone of the EMFNet model for image feature extraction, achieving remarkable performance improvements: AP_{50} increased by 3.9%, AP_{75} by 2.6%, and AP by 2.3%. Furthermore, detection accuracy was enhanced across all object size categories. These gains are attributed to RepViT's global average pooling operation and unique sandwich-style downsampling architecture. Despite the growth in model parameters, we believe that this increase in parameters is completely worthwhile and is perfectly acceptable on edge computing devices such as drones.

To further validate the effectiveness of RepViT, we generated visualization results through heatmaps. As shown in Fig. 8, when the RepVit backbone is not adopted, the model is distracted and unfocused, which makes it difficult to accurately capture the key objects, and there is an obvious misdetection phenomenon. YOLOv12n is improved compared to baseline, but there is still the phenomenon of misdetection, whereas after adopting the RepVit backbone, the model is able to significantly improve the ability of focusing on the key objects, effectively reduce the interference of the complex background, and accurately localize the object area, which significantly improves the feature differentiation of dense small objects, reduces the phenomenon of misdetection, and improves the detection accuracy.

4.4.2. Ablation experiments with DCPA

In our pursuit of the optimal balance between parameter efficiency and feature representation capability, we innovated the C3k2 block by evaluating various state-of-the-art modules including iRMB [49], MogaBlock [50], LFE [51], and PPA as replacements for the standard bottleneck. As illustrated in Table 3, a distinct trade-off was observed. While lightweight modules like iRMB and LFE successfully reduced the parameter count, they concurrently led to a degradation in detection accuracy where AP dropped to 17.2% and 17.7% respectively. This indicates that over-compressing the network compromises the extraction of fine-grained details essential for UAV imagery. Conversely, MogaBlock increased computational cost without any performance gain. In contrast, the PPA module achieved the most favorable trade-off. Despite a marginal increase in parameters of approximately 0.36M compared to Baseline, it delivered the highest accuracy gains and boosted the overall AP to 18.2%.

The DCPA with a PPA offers significant improvements in two key aspects: strengthened multi-scale feature fusion capability for effectively handling UAV objects of varying sizes, and the introduced efficient channel attention and spatial attention mechanism notably improved the model's robustness against dynamic backgrounds. Experimental results demonstrate that implementing DCPA in the neck section yielded performance gains over the baseline model: AP_{50} increased by 0.3%, AP_{75} by 0.4%, AP by 0.2%, AP_{small} by 0.2%, and AP_{large} by 2.9%.

Table 3
Ablation experiments on the C3k2 module conducted on the VisDrone dataset. "+X" indicates that the Bottleneck of C3k2 is replaced with component X.

Model	*AP*	AP_{50}	AP_{75}	AP_{small}	AP_{medium}	AP_{large}	*Paramters*	*FPS*
Baseline	18	31.1	17.9	9.6	28	34.7	2.58	1
+IRMB	17.2	30.1	17.1	9.2	26.6	35.2	**2.42**	1
+MogaBlock	17.2	30.1	17	8.2	26.5	34.9	2.6	1
+LFE	17.7	30.9	17.2	9.6	26.9	36.7	2.51	1
+PPA	**18.2**	**31.4**	**18.3**	**9.8**	**28**	**37.6**	2.94	1

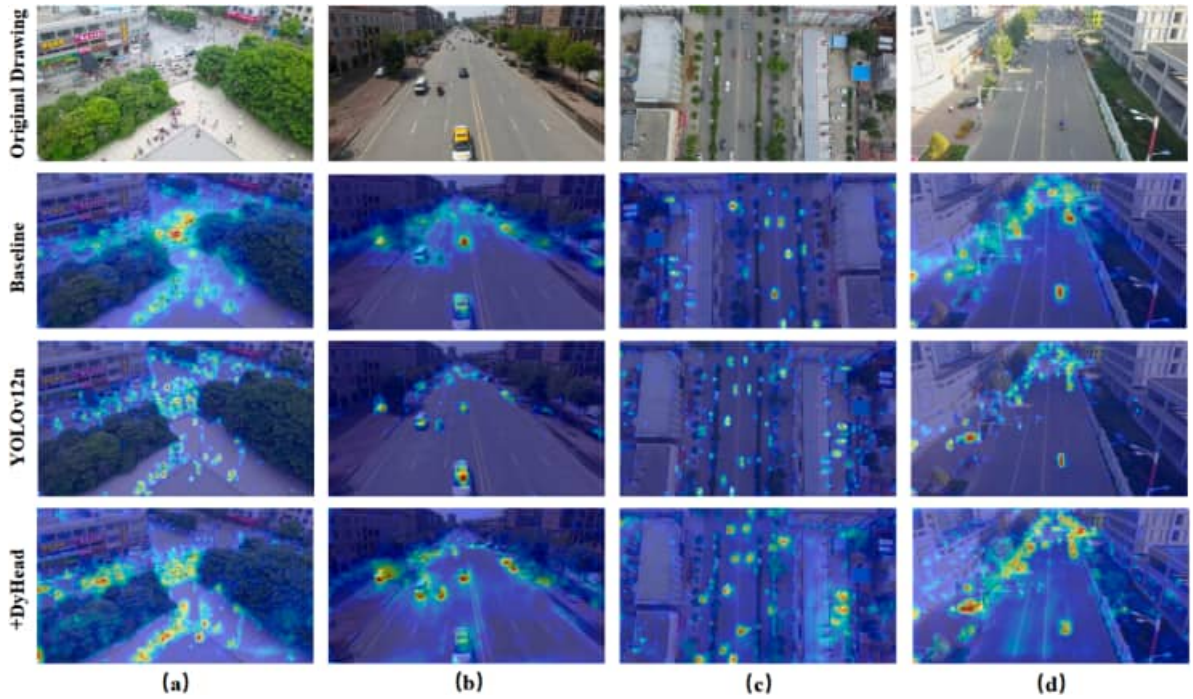

Fig. 9. The heatmaps on Visdrone, (a), (b), (c), and (d) belong to different scenarios, and the fourth row represents the introduction of DyHead on the baseline model.

4.4.3. Ablation experiments with CGDB

In the neck section of baseline, we replaced traditional downsampling methods with CGDB. Through its multi-scale feature collaborative extraction mechanism, CGDB effectively overcomes the limitations of conventional downsampling. It employs parallel local convolution and dilated convolution branches to simultaneously capture detailed features and large-scale contextual information. After channel compression and cross-dimensional interaction, a global attention mechanism dynamically selects key feature channels, thereby preserving multi-granularity semantic information while reducing feature map dimensions. Experimental results demonstrate that incorporating CGDB improves AP_{50} by 0.5%, AP_{75} by 0.1%, AP by 0.2%, AP_{medium} by 0.9%, and AP_{large} by 1.5%.

4.4.4. Ablation experiments with DyHead

The DyHead detection head, based on a three-level coupled feature enhancement architecture with dynamic multi-head attention mechanisms, demonstrates significant advantages in UAV image object detection tasks. Experimental results show that this detection head achieves an AP of 19.8%, representing a 1.8% improvement over the baseline model, with AP_{50} and AP_{75} increasing by 2.6% and 2.1% respectively. Additionally, AP_{small} improved by 1%, AP_{medium} by 2.7%, and AP_{large} by 4.7%, validating the effectiveness of its three-level attention cascade mechanism. Through joint optimization of feature enhancement, spatial adaptation, and task coordination, this mechanism significantly enhances the model's adaptability to multi-scale objects and dense occlusion scenarios in UAV top-view imagery while maintaining real-time inference speed.

To further validate the effectiveness of DyHead, we generated visualization results through heatmaps. As shown in Fig. 9. The model equipped with DyHead demonstrates broader attention to key regions in feature maps, accurately distinguishing small objects from backgrounds while significantly reducing missed detections and increasing the number of detected objects, thereby comprehensively improving object detection performance.

4.4.5. Ablation experiments with prediction head

As presented in Tables 4 and 5, we investigated the impact of different prediction head configurations on detection accuracy and parameter efficiency using the VisDrone and UAVDT datasets. The initial baseline, employing standard prediction heads $\{P_3, P_4, P_5\}$, achieved an AP of 21.7% on VisDrone and 60.9% on UAVDT, with a parameter count of 7.71M.

Recognizing that small objects in UAV imagery rely heavily on fine-grained spatial details often lost in deeper layers, we first introduced a high-resolution prediction head (P_2). This four-head configuration ($\{P_2, P_3, P_4, P_5\}$) significantly boosted performance, increasing AP to 24.6% on VisDrone and 73.2% on UAVDT. These results confirm that the high-resolution features from P_2 are critical for precise localization and feature alignment of tiny objects. However, this addition inevitably increased the model size to 7.89M parameters.

To further optimize the trade-off between efficiency and accuracy, we re-evaluated the necessity of the deep-layer head P_5. In the

Table 4
Ablation experiments on different prediction heads performed on the VisDrone dataset. P_x indicates that P_x prediction heads were used.

Prediction Head	*AP*	AP_{50}	AP_{75}	AP_{small}	AP_{medium}	AP_{large}	*Paramters*
$\{P_3, P_4, P_5\}$	21.7	36.9	21.7	12.3	33.2	40.8	7.71
$\{P_2, P_3, P_4, P_5\}$	24.6	41	24.9	16.1	35.3	41.3	7.89
$\{P_2, P_3, P_4\}$	**25.5**	**42.6**	**25.9**	**17**	**36**	**41.8**	**6.76**

Table 5
Ablation experiments on different prediction heads performed on the UAVDT dataset. P_x indicates that P_x prediction heads were used.

Prediction Head	*AP*	AP_{50}	AP_{75}	AP_{small}	AP_{medium}	AP_{large}	*Paramters*
$\{P_3, P_4, P_5\}$	60.9	75.7	70.5	28.6	75.7	92.5	7.71
$\{P_2, P_3, P_4, P_5\}$	73.2	91.5	85.9	48.3	84.8	91.7	7.89
$\{P_2, P_3, P_4\}$	**73.4**	**91.6**	**86.1**	**50**	**84.6**	**91.8**	**6.76**

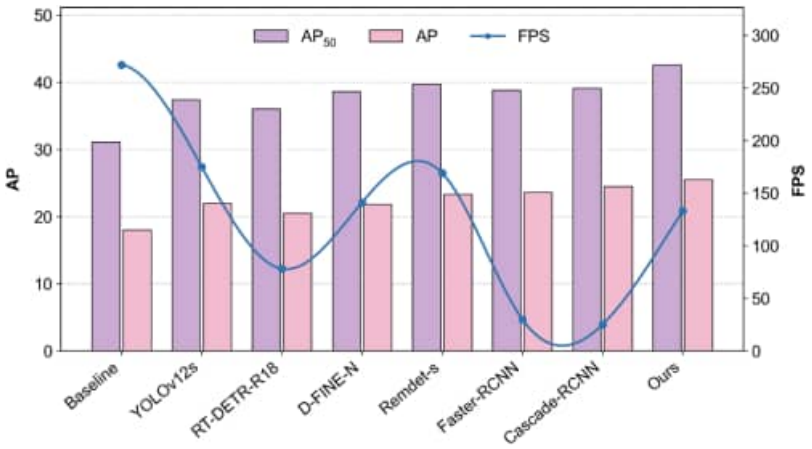

Fig. 10. The trade-off analysis between inference speed and detection accuracy among different state-of-the-art models.

context of UAV small object detection, the large receptive field of P_5 captures coarse semantic information that is largely redundant and prone to introducing background noise. Consequently, we removed P_5 to form the final $\{P_2, P_3, P_4\}$ configuration. This strategic pruning resulted in a highly efficient architecture with 6.76M parameters—representing a 12.3% reduction compared to the original baseline ($\{P_3, P_4, P_5\}$). Despite the reduced parameter count, detection accuracy further improved, reaching a peak AP of 25.5% on VisDrone (+3.8% over baseline) and 73.4% on UAVDT (+12.5% over baseline). These consistent improvements across both datasets demonstrate that substituting the coarse-grained P_5 head with the fine-grained P_2 head effectively filters background interference while enhancing feature representation for dense and rare objects.

4.5. Comparison experiments

We conducted comprehensive comparisons with various object detection models on both VisDrone and UAVDT datasets, including two-stage models (Faster-RCNN and Cascade-RCNN [52]), the Transformer-based DETR [53], single-stage models (SSD and RetinaNet), multiple versions of the YOLO series, emerging real-time Transformer detectors, as well as state-of-the-art drone-specific detection methods.

On the VisDrone dataset, as shown in Table 6, our model outperforms the two-stage models Faster-RCNN-R50 and Cascade-RCNN-R50 in terms of parameter efficiency and achieves a lead in the AP metric. Although DETR removes traditional object detection complexities like anchor generation and non-maximum suppression by harnessing Transformer's global modeling, its large parameter count results in slower training speeds. In contrast, our model achieves comprehensive outperformance over DETR-R50 with significantly fewer parameters. Compared to single-stage models SSD512 and RetinaNet, our model not only requires fewer parameters but also surpasses them in detection accuracy. When compared with drone-specific models Remdet-s [54], EdgeYOLO-s [55], and HicYOLO-s [56], our model outperforms all three in AP and AP_{small} metrics.

To further evaluate the competitiveness of EMFNet against modern architectures, we extended our comparison to include real-time Transformer-based detectors: RT-DETR-R18, D-FINE-N [57], and RT-DETR++. As detailed in Table 6, EMFNet demonstrates exceptional efficiency, achieving an AP of 25.5%, which outperforms RT-DETR-R18 and RT-DETR++ by significant margins of 5.0% and 1.4%, respectively. Notably, EMFNet achieves this with only 6.76M parameters—approximately one-third the model size of RT-DETR. While D-FINE-N utilizes fewer parameters, EMFNet exhibits a distinct performance advantage with a 3.7% higher AP, justifying the modest increase in model size for precision-demanding tasks. In comparisons with the YOLO series, our model's parameter count falls between the lowest and second-lowest parameter models, yet its detection accuracy significantly outperforms all second-lowest parameter versions. These results fully demonstrate our model's superiority in both parameter efficiency and detection performance.

Beyond standard accuracy metrics, we further investigated the balance between inference speed and detection performance by plotting the FPS, AP, and AP_{50} metrics for eight representative models in Fig. 10. While Baseline and YOLOv12s achieve higher raw frame rates, EMFNet operates at 133 FPS (measured at batch size 4 for all models), significantly exceeding the standard threshold for real-time inference. Crucially, our model runs markedly faster than the real-time Transformer model RT-DETR-R18. Most importantly, among all eight compared models, EMFNet achieves the highest AP and AP_{50}. Not only does it

Fig. 11. On the VisDrone dataset, we perform a visual comparison between Baseline, YOLOv12n and our model, where the second row is the ground truth. (a), (b), (c) and (d) in the figure represent different scenarios, and the red boxes mark the regions to focus on.

Table 6
Comparison experiments with different SOTA methods on the VisDrone dataset.

Model	*InputSize*	*AP*	AP_{50}	AP_{75}	AP_{small}	AP_{medium}	AP_{large}	*Paramters*
Faster-RCNN	1344*768	23.6	38.8	25.1	15	35.3	36	41.36
Cascade-RCNN	1344*768	24.5	39.1	26.3	15.5	**36.7**	39.6	69.16
SSD512	512*512	12.8	25.5	11.7	5.6	20.3	28.3	24.68
RetinanetR50	1344*768	13.7	23.7	13.9	5.7	23.4	31.8	36.37
DETR	1333*750	12.6	25.7	11.2	5.3	20.5	32.6	41.56
Remdet-s	640*640	23.3	39.7	23.3	14	34.4	44	11.9
EdgeYOLO-s	640*640	23.5	40.8	23.1	13.8	34.8	**49.1**	9.9
HicYOLO-s	640*640	23.6	40.4	23.4	14.6	33.6	44.6	9.32
RT-DETR-R18	640*640	20.5	36	20.2	12.5	30	39.7	20
D-FINE-N	640*640	21.8	38.6	21	12.9	32.4	47.1	3.73
RT-DETR++	640*640	24.1	40.1	23.6	14.2	34.9	41.2	21.3
YOLOv5n	640*640	8.7	17.9	7.6	5.1	13	18.5	**1.76**
YOLOv5s	640*640	12.8	24.5	11.9	8	18.8	25	7.01
YOLOv8n	640*640	16	28.1	15.6	8.7	24.5	33.9	3
YOLOv8s	640*640	20	33.9	20.4	12.2	29.8	35.2	11.1
YOLOv9t	640*640	17.4	30.3	17.2	8.8	27.6	41.8	2.62
YOLOv9s	640*640	21.3	35.9	21.7	12.4	32.4	44.6	9.6
YOLO11n(baseline)	640*640	18	31.1	17.9	9.6	28	34.7	2.58
YOLO11s	640*640	21.6	36.8	21.8	13	32.5	41.6	9.42
YOLOv12n	640*640	17.5	30.6	17.3	9.2	27.2	38.8	2.56
YOLOv12s	640*640	22	37.4	22.1	13	33	44	9.23
Ours	640*640	**25.5**	**42.6**	**25.9**	**17**	36	41.8	6.76

outperform the UAV object detection model RemDet-s and the Transformer-based D-FINE-N, but it also surpasses two-stage detectors such as Faster R-CNN. This confirms that EMFNet breaks the bottleneck of traditional lightweight models by delivering state-of-the-art accuracy without compromising real-time capability.

As shown in Fig. 12, after 50 epochs, our model EMFNet consistently maintains superior detection accuracy compared to baseline, YOLO11s, YOLOv12n, and YOLOv12s. Not only does it have nearly 3 million fewer parameters than YOLO11s and YOLOv12s, but it also achieves approximately 4% higher AP than YOLO11s. When compared with the YOLOv12n and YOLOv12s, EMFNet still maintains a leading position in AP, fully demonstrating its outstanding performance. Although our model's parameters for 6.76 million exceeds that of baseline, it remains well-suited for UAV object detection tasks.

To demonstrate the superiority of our model, we conducted visual comparisons. As shown in Fig. 11, the illustration includes the original image, ground truth, baseline, YOLOv12n, and EMFNet. We compared not only with the baseline model baseline but also with the YOLOv12n.

Table 7
Comparison experiments with different SOTA methods on the UAVDT dataset.

Model	*AP*	AP_{50}	AP_{75}	AP_{small}	AP_{medium}	AP_{large}	*Paramters*
Faster-RCNN-R50	46.1	64	55.3	41.2	51.6	54	41.36
Cascade-RCNN-R50	59.8	78.7	71.6	53.2	67.7	59	69.16
SSD512	65.6	95	76.8	54	76	84.3	24.68
Retinanet-R50	28.9	40.1	33.5	19.6	37.1	39.1	36.37
DETR-R50	67.6	96.2	79	52.3	78.4	89.7	41.56
Remdet-s	63.3	74.5	71.1	31.8	79.9	**94.3**	11.9
HicYOLO-s	69.8	**98**	83.3	**57**	78.9	89.1	9.32
RT-DETR-R18	68.5	90.8	82.2	49.8	81.8	90.4	20
D-FINE-N	70.7	91.2	84.7	52.1	83	92.2	3.73
YOLOv5n	63.4	90.9	73.9	43.4	75.5	87.2	**1.76**
YOLOv5s	67	94.4	78.5	49.5	77.7	88.2	7.01
YOLOv8n	58.4	74.5	68.4	28.1	74.7	89.7	3
YOLOv8s	61.1	74.7	70.4	30.3	77.4	93.2	11.1
YOLOv9t	56.9	76.8	66.2	23.8	74.9	90.6	2.62
YOLOv9s	61.6	79	71.3	27	80.1	92.1	9.6
YOLO11n(baseline)	58.2	47.8	68	27.5	74.9	90	2.58
YOLO11s	60.5	75	69.9	29.2	77.4	93.7	9.42
YOLOv12n	58.5	74.3	67.6	27.8	75.1	91.2	2.56
YOLOv12s	61.1	74.8	69.9	29.8	78.1	92.9	9.23
Ours	**73.4**	91.6	**86.1**	50	**84.6**	91.8	6.76

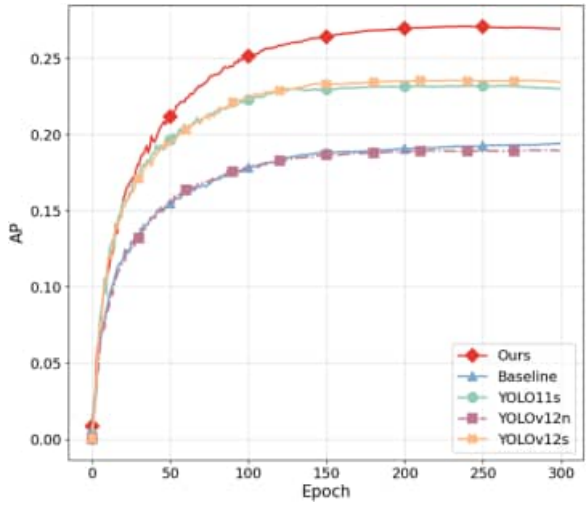

Fig. 12. Comparison curve of AP with epoch transformation within 300 epochs on VisdDrone dataset.

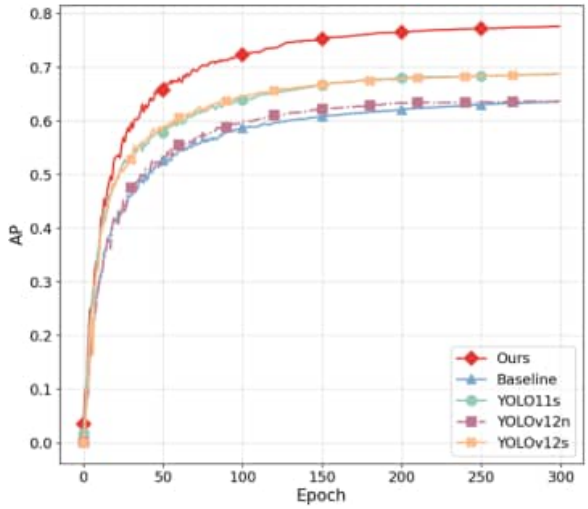

Fig. 13. Comparison curve of AP with epoch transformation within 300 epochs on UAVDT dataset.

The comparison scenarios cover complex environments including daytime, dusk, nighttime, and overexposed conditions. The results show that EMFNet outperforms both baseline and YOLOv12n across all scenarios, fully demonstrating its advantages.

Specifically, in Fig. 11(a), EMFNet successfully detected an unlabeled and occluded car a highly challenging task that neither baseline nor YOLOv12n could accomplish. In Fig. 11(b), under overexposed conditions, baseline misidentified a traffic cone as a pedestrian, while our model overcame this difficulty, effectively reducing false detections. In Fig. 11(c), for the rare category "awning-tricycle" in the dataset, insufficient lighting at dusk caused baseline and YOLOv12n to miss detections, where as EMFNet performed better in reducing missed detections. In Fig. 11(d), in dark conditions, baseline and YOLOv12n have limited detection of unlabeled and densely overlapping pedestrians, while EMFNet detects many unlabeled pedestrians, proving that it learns the object details more deeply.

Since the UAVDT dataset contains only three categories, which are less difficult to detect compared to the Visdrone dataset, the AP metrics of the model are significantly improved. Nevertheless, as shown in Table 7, our model's AP metrics on this dataset are still better than the two-stage models Faster-RCNN-R50 and Cascade-RCNN-R50, and EMFNet achieves a certain lead in AP metrics compared to the single-stage models SSD512 and Retinanet-R50 and DETR-R50. Notably, when benchmarked against the latest real-time Transformer-based detectors, EMFNet demonstrates a superior efficiency-accuracy balance. It surpasses RT-DETR-R18 by 4.9% in AP despite the latter's significantly higher computational cost. Moreover, while D-FINE-N achieves a lower parameter count, EMFNet outperforms it by 2.7% in AP and 1.4% in AP_{75}, verifying that our proposed modules capture fine-grained aerial features more effectively than these generic Transformer architectures. Compared to the UAV object detection model, EMFNet outperforms Remdet-s in all metrics except AP_{large} metrics. Compared to HicYOLO-s, EMFNet has 3.6% higher AP metrics with almost 3 million fewer parameters. Compared to the baseline model, our model AP improves by 11 percentage points; compared to the YOLOv12, the AP of EMFNet is 14% and 8.8% higher than that of YOLOv12n and YOLOv12s, respectively. Fig. 13 shows our advantage even more clearly, with EMFNet's AP consistently outperforming the baseline models YOLO11n, YOLO11s, YOLOv12n, and YOLOv12s after 50 epochs of training.

The visual comparison on the UAVDT dataset is shown in Fig. 14. In the UAVDT dataset, buses and trucks account for an extremely low proportion, making it difficult for models to capture their refined features from a large number of samples. However, our model successfully overcomes this challenge. Whether in daytime or nighttime condi-

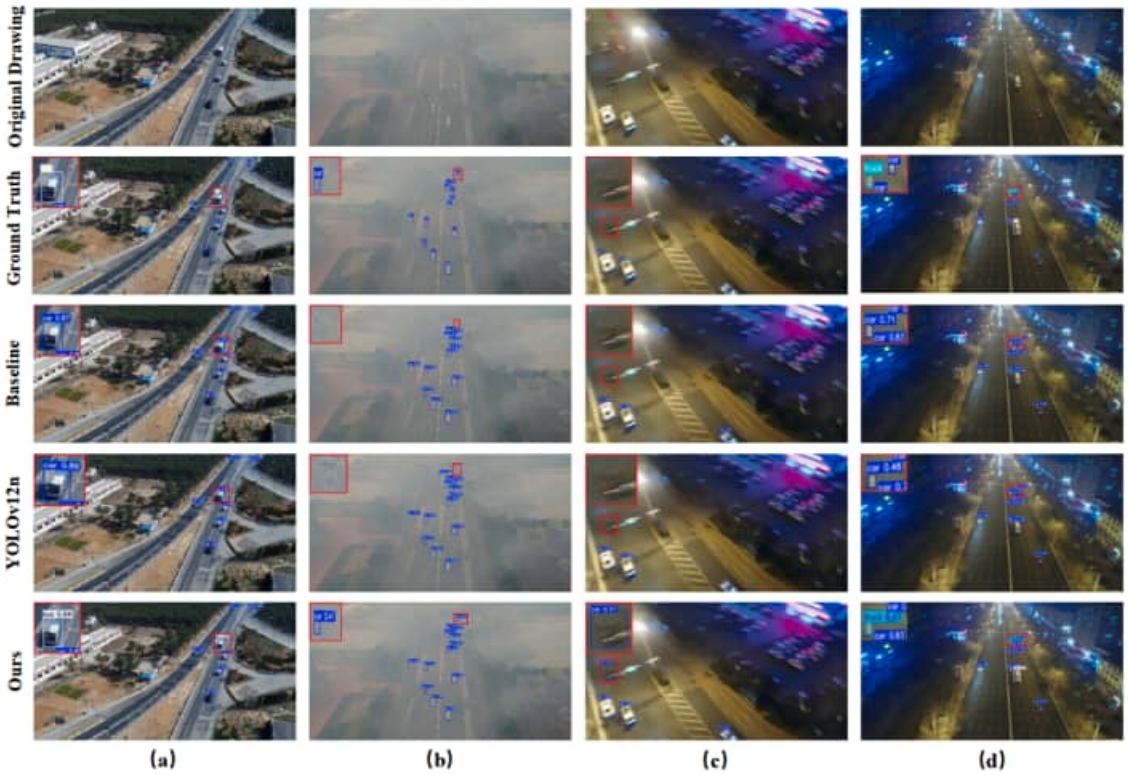

Fig. 14. On the UAVDT dataset, we perform a visual comparison between Baseline, YOLOv12n and our model, where the second row is the ground truth. (a), (b), (c) and (d) in the figure represent different scenarios, and the red boxes mark the regions to focus on.

tions, as shown in Fig. 14(a) and (d), our model demonstrates superior detection performance for buses and trucks compared to baseline and YOLOv12n. In Fig. 14(c), where the image is blurred due to nighttime conditions, the detection task faces significant challenges. Nevertheless, our model not only correctly detects the labeled objects but also successfully identifies more unlabeled objects. Additionally, in Fig. 14(b), under foggy conditions, our model exhibits outstanding performance. Compared to baseline and YOLOv12n, it not only avoids missed detections but also achieves higher confidence in detection results. These results fully demonstrate the superior performance and robustness of our model in handling rare categories and various weather and lighting conditions.

5. Limitation

The current study has certain limitations. As this phase focuses on architectural validation, experiments were conducted on workstation-grade GPUs rather than resource-constrained UAV edge devices. Future research will address these aspects by deploying EMFNet on edge computing platforms, such as the NVIDIA Jetson Orin NX, and applying model quantization techniques. With a lightweight footprint of only 6.76M parameters, EMFNet achieves substantial accuracy gains by improving AP by 7.5% and 15.2% on the VisDrone and UAVDT datasets, respectively. This performance indicates that the model strikes an effective balance between model complexity and detection precision required for UAV tasks.

6. Conclusion

In this study, we presented EMFNet, a lightweight and efficient single-stage detector specifically engineered to address the challenges of inherent challenges of UAV aerial imagery, including extreme scale variations, dense small object distribution, and complex environmental interference. By innovatively integrating the RepViT backbone with our proposed Dual Cross-Stage Partial Attention (DCPA) and Context Guided Downsample Block (CGDB), the framework effectively optimizes multi-scale feature fusion while mitigating the critical information loss typically associated with downsampling. Furthermore, the incorporation of the DyHead with a cascaded attention mechanism significantly enhances the model's adaptability to diverse lighting conditions and severe occlusions. Extensive experiments on the VisDrone and UAVDT datasets show that EMFNet. With only 6.76 million parameters and a real-time inference speed of 133 FPS, our model achieves substantial AP improvements of 7.5% and 15.2% over the baseline models, respectively. These results confirm that EMFNet strikes an optimal balance between model complexity and detection precision, establishing a new benchmark for lightweight UAV object detection. Looking forward, our future research will focus on bridging the gap between theoretical efficiency and practical deployment. First, we aim to implement model quantization and acceleration techniques to deploy EMFNet on resource-constrained edge computing platforms, such as the NVIDIA Jetson series, ensuring robust performance in real-world flight missions. Second, to further enhance robustness against adverse weather conditions, we plan to explore cross-modal data fusion by incorporating thermal or multi-spectral imagery, thereby extending the detection capabilities of UAVs in all-weather scenarios.

Data availability

Data will be made available on request.

Declaration of competing interest

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

Acknowledgment

This work was supported by the National Natural Science Foundation of China (62376252); Key Project of Natural Science Foundation of Zhejiang Province (LZ22F030003); Zhejiang Province Leading Geese Plan (2025C02025, 2025C 01056); Zhejiang Province Province-Land Synergy Program (2025SDXT004-3).

References

[1] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arxiv:1409.1556 (2014).
[2] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[3] M. Tan, Q. Le, EfficientNetv2: smaller models and faster training, in: International Conference on Machine Learning, PMLR, 2021, pp. 10096–10106.
[4] S. Ren, K. He, R. Girshick, J.Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (6) (2016) pp. 1137–1149
[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: single shot multibox detector, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer, 2016, pp. 21–37.
[6] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[7] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[8] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[10] J. Redmon, S. Divvala, R. Girshick, 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.
[11] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7263–7271.
[12] J. Redmon, A. Farhadi, YOLOv3: an incremental improvement, arxiv:1804.02767 (2018).
[13] A. Bochkovskiy, C.-Y. Wang, H.-Y.M. Liao, YOLOv4: optimal speed and accuracy of object detection, arxiv:2004.10934 (2020).
[14] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, et al., YOLOv6: a single-stage object detection framework for industrial applications, arxiv:2209.02976 (2022).
[15] 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, 2023, pp. 7464–7475.
[16] D. Reis, J. Kupec, J. Hong, A. Daoudi, Real-time flying object detection with YOLOv8, arxiv:2305.09972 (2023).
[17] C.-Y. Wang, I.-H. Yeh, 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.
[18] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, G. Ding, YOLOv10: real-time end-to-end object detection. arXiv 2024, arxiv:2405.14458 (2024).
[19] R. Khanam, M. Hussain, YOLOv11: an overview of the key architectural enhancements, arxiv:2410.17725 (2024).
[20] Y. Tian, Q. Ye, D. Doermann, YOLOv12: attention-centric real-time object detectors, arxiv:2502.12524 (2025).
[21] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, J. Chen, Detrs beat yolos on real-time object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, 16965–16974.
[22] X. Chen, C. Lin, EVMNet: eagle visual mechanism-inspired lightweight network for small object detection in UAV aerial images, Digit. Signal Process. 158 (2025) 104957.
[23] H. Liao, Y. Tang, Y. Liu, X. Luo, ViDoneNet: an efficient detector specialized for target detection in aerial images, Digit. Signal Process. 164 (2025) 105270.
[24] B. Du, Y. Huang, J. Chen, D. Huang, Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13435–13444.
[25] C. Chen, J. Qi, X. Liu, K. Bin, R. Fu, X. Hu, P. Zhong, Weakly misalignment-free adaptive feature alignment for uavs-based multimodal object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 26836–26845.
[26] T. Mengchu, C. Meiji, C. Zhimin, M.A. Yingliang, Y.U. Shaohua, MFR-YOLOv10: object detection in UAV-taken images based on multilayer feature reconstruction network, Chin. J. Aeronaut. 38 (11) (2025) pp.103456 .
[27] L. Zhou, S. Zhao, S. Li, Y. Wang, Y. Liu, X. Zuo, A lightweight object detection method based on fine-grained information extraction and exchange in UAV aerial images, Knowl. Based Syst. 315 (2025) 113253.
[28] H. Zhang, K. Liu, Z. Gan, G.-N. Zhu, UAV-DETR: efficient end-to-end object detection for unmanned aerial vehicle imagery, arxiv:2501.01855 (2025).
[29] S. Huang, S. Ren, W. Wu, Q. Liu, Discriminative features enhancement for low-altitude UAV object detection, Pattern Recognit. 147 (2024) 110041.
[30] S. Liu, M. Zhu, R. Tao, H. Ren, Fine-grained feature perception for unmanned aerial vehicle target detection algorithm, Drones 8 (5) (2024) 181.
[31] Y. Huang, J. Chen, D. Huang, UFPMP-Det: toward accurate and efficient object detection on drone imagery, in: Proc. AAAI Conf. Artif. Intell., 36, 2022, pp. 1026–1033.
[32] C. Li, T. Yang, S. Zhu, C. Chen, S. Guan, Density map guided object detection in aerial images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 190–191.
[33] Y. Zeng, T. Zhang, W. He, Z. Zhang, YOLOv7-UAV: an unmanned aerial vehicle image object detection algorithm based on improved yolov7, Electronics 12 (14) (2023) 3141.
[34] Z. Liu, X. Gao, Y. Wan, J. Wang, H. Lyu, An improved YOLOv5 method for small object detection in UAV capture scenes, IEEE Access 11 (2023) 14365–14374.
[35] S. Li, X. Yang, X. Lin, Y. Zhang, J. Wu, Real-time vehicle detection from UAV aerial images based on improved YOLOv5, Sensors 23 (12) (2023) 5634.
[36] R. Zhang, S. Newsam, Z. Shao, X. Huang, J. Wang, D. Li, Multi-scale adversarial network for vehicle detection in UAV imagery, ISPRS J. Photogramm. Remote Sens. 180 (2021) 283–295.
[37] M. Bakirci, Utilizing YOLOv8 for enhanced traffic monitoring in intelligent transportation systems (ITS) applications, Digit. Signal Process. 152 (2024) 104594.
[38] M. Bakirci, Advanced aerial monitoring and vehicle classification for intelligent transportation systems with YOLOv8 variants, J. Netw. Comput. Appl. 237 (2025) 104134.
[39] A. Wang, H. Chen, Z. Lin, J. Han, G. Ding, RepViT: revisiting mobile CNN from vit perspective, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 15909–15920.
[40] S. Xu, S. Zheng, W. Xu, R. Xu, C. Wang, J. Zhang, X. Teng, A. Li, L. Guo, HCF-Net: hierarchical context fusion network for infrared small object detection, in: 2024 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2024, pp. 1–6.
[41] 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, 2020, pp. 11534–11542.
[42] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, CBAM: convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[43] T. Wu, S. Tang, R. Zhang, J. Cao, Y. Zhang, CGNet: a light-weight context guided network for semantic segmentation, IEEE Trans. Image Process. 30 (2020) 1169–1179.
[44] X. Dai, Y. Chen, B. Xiao, D. Chen, M. Liu, L. Yuan, L. Zhang, Dynamic head: unifying object detection heads with attentions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7373–7382.
[45] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang, et al., VisDrone-DET2019: the vision meets drone object detection in image challenge results, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019,
[46] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, Q. Tian, The unmanned aerial vehicle benchmark: object detection and tracking, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 370–386.
[47] Y. Shufang, RT-DETR++ for UAV object detection, arxiv:2509.09157 (2025).
[48] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.
[49] J. Zhang, X. Li, J. Li, L. Liu, Z. Xue, B. Zhang, Z. Jiang, T. Huang, Y. Wang, C. Wang, Rethinking mobile block for efficient attention-based models, 2023 IEEE, in: CVF International Conference on Computer Vision (ICCV), 2023, pp. 1389–1400.
[50] S. Li, Z. Wang, Z. Liu, C. Tan, H. Lin, D. Wu, Z. Chen, J. Zheng, S.Z. Li, MogaNet: multi-order gated aggregation network, arxiv:2211.03295 (2022).
[51] X. Zhang, H. Zeng, S. Guo, L. Zhang, Efficient long-range attention network for image super-resolution, in: European Conference on Computer Vision, Springer, 2022, pp. 649–667.
[52] Z. Cai, N. Vasconcelos, Cascade R-CNN: delving into high quality object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6154–6162.
[53] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer, 2020, pp. 213–229.
[54] C. Li, R. Zhao, Z. Wang, H. Xu, X. Zhu, RemDet: rethinking efficient model design for UAV object detection, in: Proc. AAAI Conf. Artif. Intell., 39, 2025, pp. 4643–4651.
[55] S. Liu, J. Zha, J. Sun, Z. Li, G. Wang, EdgeYOLO: an edge-real-time object detector, in: 2023 42nd Chinese Control Conference (CCC), IEEE, 2023, pp. 7507–7512.
[56] S. Tang, S. Zhang, Y. Fang, HIC-YOLOv5: improved YOLOv5 for small object detection, in: 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 6614–6619.
[57] Y. Peng, H. Li, P. Wu, Y. Zhang, X. Sun, F. Wu, D-FINE: redefine regression task in DETRs as fine-grained distribution refinement, arxiv:2410.13842 (2024).

Mingquan Wang received the BS degree from Shandong Women's University, in 2024. He is currently pursuing a ME degree in Computer Science at Zhejiang Normal University, focusing on deep learning and computer vision, specifically in object detection.

Huiying Xu received the MS degree from National University of Defense Technology(NUDT), China. She is an associate professor with the School of Computer Science and Technology, Zhejiang Normal University, and also the researcher of Research Institute of Ningbo Cixing Co. Ltd, PR, China. Her research interests include Kernel learning and feature selection, Object Detection, Vision SLAM, Computer vision, Image processing, Pattern recognition, Computer simulation, Deep clustering, Diffusion Model, Multiple Kernel Learning, Learning with incomplete data and their applications. She is a member of the China Computer Federation. She has published papers, including those in highly regarded journals such International Journal of Intelligent Systems, IEEE Transactions on Cybernetics, IEEE Transactions on Multimedia, ACM MM, etc.

Yiming Sun received his MS and PhD degrees from Tianjin University, Tianjin, China, in 2020 and 2024. He received his BS degree from Hefei University of Technology, Hefei, China, in 2017. Now he is working as an assistant professor in the School of Automation, Southeast University, Nanjing, Jiangsu, China. His current research interests include multimodal fusion, object detection, and UAV-based vision.

Hongbo Li received his PhD degree in computer science from Tsinghua University in 2009. Currently, he holds the position of the Chief Technology Officer and Co-founder of Beijing Geek+ Technology Co., Ltd China. In addition, he also serves as the secretary-general of Chinese Intelligent service Society and is an Editorial Board Member of several high-profile journals. His research interests include the design and application of intelligent robots, intelligent information process, and intelligent logistic systems. He has published more than 70 papers in prestigious journals and conference, and has been awarded more than 120 patents, including 46 international invention patents.

Zeyu Wang received the MS degree in Software Engineering from College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua China, in 2025. His current research interests include deep learning, computer vision tasks, Object Detection.

Yi Li received the MS degree in Software Engineering from College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua China, in 2021, where he is currently pursuing the PhD degree in Computer Science and Technology. He is a student member of CFF. His current research interests include deep learning, computer vision tasks, Object Detection, Instance Segmentation, Object Tracking.

Ruidong Wang received the PhD degree in Computer Applied Technology from the Harbin University of Science and Technology, Harbin, China, in 2024. He is currently a lecturer in the School of Computer Science and Technology, Zhejiang Normal University, Jinhua, China. His current research interests include graph data mining, time series analysis, and anomaly detection.

Xinzhong Zhu received the PhD degree from Xidian University and MS degree from National University of Defense Technology (NUDT), China. He is a professor with the School of Computer Science and Technology and dean of Hangzhou research institute of Artificial Intelligence, Zhejiang Normal University, and also the chief scientist of Beijing Geekplus Technology Co., Ltd. and president of Research Institute of Ningbo Cixing Co., Ltd., China. His research interests include Machine learning, Deep clustering, Computer vision, Object detection, Segmentation, Recognition and Tracking, Diffusion Model, Manufacturing informatization, Manufacturing informatization, Robotics and System integration, Laser SLAM, Vision SLAM, Low Quality Data Learning, Multiple Kernel Learning, and Intelligent manufacturing. He is a member of the ACM and certified as CCF distinguished member. Dr. Zhu has published more than 30 peer-reviewed papers, including those in highly regarded journals and conferences such as the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Image Processing, the IEEE Transactions on Multimedia, the IEEE Transactions on Knowledge and Data Engineering, CVPR, NeurIPS, AAAI, IJCAI, ACM MM, etc. He served on the Technical Program Committees of IJCAI 2020 and AAAI 2020.