LEA-DETR: A Lightweight and Efficient Attention-Enhanced Model for UAV Object Detection

Haohao Ma , Mingliang Zuo , Qiqi Ge *

Posted Date: 29 October 2025

doi: 10.20944/preprints202510.2237.v1

Keywords: UAV; small object detection; deep learning; attention mechanism; feature fusion

Article

LEA-DETR: A Lightweight and Efficient Attention-Enhanced Model for UAV Object Detection

Haohao Ma, Mingliang Zuo and Qiqi Ge*

Department of Automation, Shanghai Jiao Tong University, Shanghai, 200240, China

* Correspondence: geqq@sjtu.edu.cn

Highlights

What are the main findings?

- An adaptive attention fusion (AAF) module enhances feature precision by integrating sparse and dense attention.
- EMS-FPN improves multi-scale fusion efficiency and small object detection.
- A novel PG-Backbone reduces parameters and computation while preserving accuracy.

What are the implications of the main findings?

- Enhanced accuracy and robustness in small-object detection for remote sensing imagery.
- A lightweight design with reduced parameters for UAV-based monitoring in complex backgrounds.

Abstract

Object detection in UAV scenarios has emerged as a highly significant yet challenging research topic. UAV-captured images often suffer from issues such as tiny object sizes with indistinct features, dense object distributions, and frequent occlusions. Additionally, UAV platforms are constrained by limited onboard computing power and storage capacity, which necessitates lightweight and efficient detection algorithms. In this paper, we propose LEA-DETR, a lightweight and efficient UAV object detection model based on the DETR framework. Specifically, we design an adaptive sparse-dense attention fusion module to enhance the model's ability to selectively attend to critical features. Furthermore, we introduce an efficient multi-scale convolutional FPN to effectively integrate low-level semantic information, particularly for small object detection. To further improve local representation capability, we design a novel gated inverted bottleneck convolution module, which forms the core of our newly proposed PG-Backbone for feature extraction. We conduct extensive ablation and comparison experiments on the VisDrone2019 dataset. The ablation results demonstrate that the proposed LEA-DETR-S and LEA-DETR-M models significantly reduce the number of parameters and computational overhead while achieving notable improvements in accuracy. Comparative evaluations show that LEA-DETR achieves the best overall performance among all evaluated models, making it a practical and efficient solution for UAV-based object detection in real-world scenarios.

Keywords: UAV; small object detection; deep learning; attention mechanism; feature fusion

1. Introduction

In recent years, the rapid advancement of Unmanned Aerial Vehicle (UAV) technology has enabled a wide range of applications in complex environments, owing to their compact size, high mobility, operational flexibility, and ease of deployment. UAVs have shown great potential in various fields such as traffic monitoring[1], agricultural inspection[2], and modern military

 doi:10.20944/preprints202510.2237.v1

reconnaissance[3]. A common requirement across these applications is the ability to detect and track multiple targets from high-altitude aerial views, which is essential for effective situation awareness and automated decision making.

Despite these advantages, UAV-based vision systems still face several significant challenges. Due to the limited payload capacity of UAVs, onboard computing devices often lack the ability to run large-scale deep learning models[4]. Moreover, UAVs typically operate at relatively high altitudes to expand their observation range, causing targets in the captured images to appear extremely small, which poses a substantial challenge for accurate detection. In addition, UAV imagery frequently involves complex scenes with dense object distributions and occlusions, making the detection task even more difficult[5].

To address these issues—namely, the limited computational resources and the difficulty of small object detection in UAV scenarios—this paper explores a lightweight object detection approach based on the DEtection TRansformer (DETR) framework. Our work focuses on enhancing the performance of small object detection while maintaining computational efficiency, making it more suitable for deployment on resource-constrained UAV platforms. The main contributions of this paper are as follows:

- We propose an Adaptive Attention Fusion (AAF) module that adaptively integrates sparse and dense attention through a pre-fusion mechanism. This design enhances the model's ability to extract global contextual features, thereby significantly improving the detection performance on small objects.
- To further enhance the detection capability for small objects, we introduce a novel feature fusion pyramid network, termed EMS-FPN. This module enlarges the receptive field and strengthens detail preservation through globally heterogeneous multi-scale convolution, all while maintaining low model complexity in terms of parameters and computational cost.
- We design a Lightweight Attention-Gated Module (LAGM) based on gated inverted bottleneck convolution to achieve locally dynamic perception. LAGM improves feature modeling in dense scenes and for small objects, while preserving the overall efficiency and lightweight nature of the model.

The remainder of this paper is organized as follows. Section 2 reviews the background of object detection and recent advances in UAV-based object detection. Section 3 presents the detailed architecture of the proposed LEA-DETR, including the Adaptive Attention Fusion module, the Efficient Multi-scale Convolutional Feature Pyramid Network, and the backbone enhancements via gated mechanisms. Section 4 provides extensive experiments, including ablation studies and comparisons with state-of-the-art methods, to demonstrate the effectiveness of LEA-DETR. Finally, Section 5 concludes the paper with a summary of our contributions and future directions.

2. Related Work

Object detection has made significant strides in recent years, driven largely by advances in deep learning. Based on the detection pipeline, existing methods can be broadly categorized into two types: two-stage and one-stage detectors. Two-stage detectors, such as the R-CNN series[6–8], first generate region proposals and then perform classification and bounding box regression. While offering high accuracy, they often suffer from lower inference speed. In contrast, one-stage detectors reformulate object detection as a regression task, directly predicting object locations and categories from raw images. These models, exemplified by the YOLO (You Only Look Once) series[9–15], are faster and more suitable for real-time applications due to their streamlined architecture.

The YOLO series has undergone rapid iterations and innovations. YOLOv1[9], introduced in 2016, pioneered one-stage object detection by dividing the input image into an S×S grid, with each cell responsible for predicting bounding boxes and class probabilities. Despite achieving 45 FPS, its grid-based structure limited the accuracy of small object detection. YOLOv3[10], released in 2018, marked a significant improvement by introducing the DarkNet 53 backbone and employing Feature

 doi:10.20944/preprints202510.2237.v1

Pyramid Networks (FPN) for multi-scale prediction, enhancing the model's ability to detect overlapping and small-scale objects.

YOLOv5[12], reimplemented in PyTorch, introduced novel designs such as the Focus slicing structure for information-preserving downsampling, the CSPNet-based lightweight backbone, and a dual pyramid structure combining FPN and PANet. It also incorporated automatic anchor box generation, mixed-precision training, and Exponential Moving Average (EMA) weight updates, making it one of the most widely adopted detectors in industry.

YOLOv8[15], developed by Ultralytics, replaced the traditional CSP structure with a new C2f module, adopted an anchor-free head, and integrated dynamic convolution mechanisms. It supports multi-task training for detection, instance segmentation, and pose estimation within a unified framework, achieving an mAP of 56.8 on the COCO dataset.

Several specialized YOLO variants have also emerged. YOLOv4[11] introduced a comprehensive "Bag of Freebies" training paradigm by combining Mosaic augmentation, the CIoU loss, and PAN-based feature fusion. YOLOv7[14] proposed a unified framework for multi-task learning with innovations in dynamic label assignment and gradient flow optimization, balancing accuracy and speed. Deployment-focused versions like YOLOv6[13] and YOLO-NAS[15] utilize re-parameterization and quantization-aware training to enable efficient inference on embedded devices.

In contrast to YOLO's regression-based approach, the DEtection TRansformer (DETR)[16] framework redefined object detection as a set prediction problem. By leveraging an encoder–decoder transformer architecture, DETR eliminated the need for anchors and Non-Maximum Suppression (NMS). The encoder uses self-attention to capture global context, while the decoder employs cross-attention to align learned object queries with image features. DETR achieves an mAP of 42.0 on COCO, but suffers from high computational cost, slow convergence, and suboptimal small-object performance.

To address these limitations, several DETR[16] variants have been proposed. Deformable DETR[17] introduces deformable attention to reduce computational overhead by focusing on sparse key sampling points. DN-DETR[18] applies a denoising training scheme by injecting controlled noise into ground-truth boxes to enhance query discrimination. RT-DETR[19], a recent real-time transformer-based detector, combines YOLO-style lightweight designs with DETR's attention-based modeling, employing a hybrid encoder and IoU-aware query selection. It achieves 108 FPS and 53.1 mAP on COCO with a single 1080Ti GPU, making transformer-based detection feasible for industrial deployment.

Given the real-time constraints in UAV-based applications, one-stage detectors have become the preferred choice. Zhai et al. enhanced YOLOv8 by introducing high-resolution detection heads and integrating SPD-Conv and GAM attention modules, which improved small-object detection while reducing model complexity[20]. Sun et al. proposed RSOD, a YOLOv3-based framework tailored for UAV scenarios, incorporating shallow feature localization, multi-scale FPN fusion, adaptive weighting, and improved channel attention mechanisms[21]. Wang et al. developed UAV-YOLOv8 with a new WIoUv3 loss function, BiFormer attention, and a Focal FasterNet module for better multi-scale feature fusion and reduced false negatives[22]. Zeng et al. introduced SCA-YOLO based on YOLOv5, employing hybrid attention and an improved SEB module to enhance small-object feature representation and foreground–background separation[23]. Zhao et al. modified YOLOv7 with an additional prediction head for tiny targets and a simple non-parametric attention module, demonstrating robust performance in maritime UAV imagery[24].

In addition to YOLO-based improvements, DETR-based models for UAV detection have also gained traction. Zhang et al. proposed UAV-DETR, which incorporates multi-scale frequency-domain feature fusion, downsampling, and semantic alignment modules to enhance generalization and accuracy[25]. Kong et al. developed Drone-DETR with a lightweight ESDNet backbone and EDF-FAM attention module, utilizing dynamic competitive learning for better small-object detection[26]. Liu et al. introduced ESO-DETR, featuring gated single-head attention blocks, multi-head self-attention across multiple scales, and a novel ESO-FPN combining large-kernel convolutions with

 doi:10.20944/preprints202510.2237.v1

dual-domain attention. They also proposed the EMASlideVariFocal loss for dynamic weighting, significantly improving detection precision[27].

In summary, while recent UAV object detection approaches have shown promising accuracy improvements, most of them overlook the practical constraints of UAV platforms, such as limited onboard computational resources and storage. Therefore, it is crucial to develop detection frameworks that strike a balance between high accuracy and lightweight design for UAV deployment.

3. Methods

3.1. Overview

To address the challenges of small object detection while maintaining a lightweight model structure, we propose a novel detection framework named LEA-DETR. This model is specifically designed to improve detection accuracy in diverse UAV-based scenarios. The overall architecture of LEA-DETR is illustrated in Figure 1. The framework consists of three key components. First, the PG-Backbone serves as the feature extraction network, generating multi-scale feature representations that capture both spatial details and semantic context. Second, the Attention-Augmented Fusion (AAF) module enhances the semantic expressiveness of deep features at the P5 level by adaptively refining attention across spatial and channel dimensions. Third, the Efficient Multi-Scale Feature Pyramid Network (EMS-FPN) is employed to effectively fuse shallow features at the P2 level, preserving fine-grained details crucial for detecting small objects. Finally, the refined multi-scale features are processed by a transformer-based decoder to generate the final detection results.

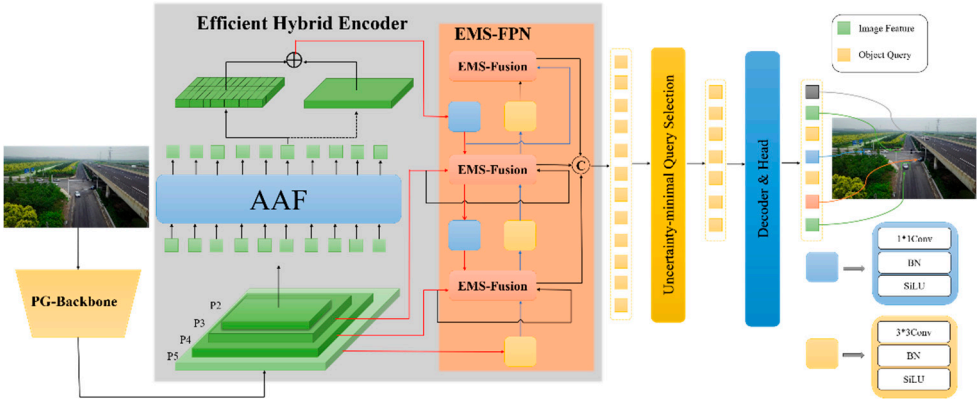


Figure 1. Figure 1 illustrates the overall architecture of the proposed LEA-DETR model. The PG-Backbone extracts multi-scale features, which are enhanced by the AAF module and subsequently fused by the EMS-FPN to achieve accurate detection of small objects.

3.2. Adaptive Attention Fusion (AAF)

Zhou et al. proposed the Adaptive Sparse Self-Attention (ASSA) module, which introduces a sparse self-attention mechanism based on the squared ReLU activation[28]. This mechanism aims to suppress features with low query-key similarity scores that could negatively impact the model's performance. To mitigate excessive sparsity, a parallel dense attention branch is incorporated, allowing the model to retain informative features while reducing noise and redundancy.

However, the original ASSA module employs a static fusion strategy to combine sparse and dense attention branches, lacking an adaptive selection mechanism. Furthermore, the module only models spatial dependencies while neglecting inter-channel relationships. This static design may result in suboptimal feature selection or omission, particularly in complex scenes or scenarios involving small object detection, where both precise spatial localization and semantic channel-wise awareness are crucial. Consequently, the representational power of the ASSA module may be limited under such challenging conditions.

 doi:10.20944/preprints202510.2237.v1

To address the aforementioned limitations, we propose an Adaptive Attention Fusion (AAF) mechanism as a lightweight alternative to the AIFI module in RT-DETR. The AAF module adaptively fuses sparse and dense attention based on a simple gating strategy that computes fusion weights from the original input features. This pre-fusion mechanism enables the model to dynamically adapt to the content of the input, making the attention selection purely data-driven.

Furthermore, the AAF module jointly integrates both spatial and channel attention to enhance the expressiveness and discriminative power of feature representations. As illustrated in Figure 2, the module adopts a parallel structure that combines a Squeeze-and-Excitation (SE) block with Sparse Self-Attention (SSA) and Dense Self-Attention (DSA).

Specifically, the input feature X is first split into three branches to generate the Query (Q), Key (K), and Value (V) components. The Q and K branches are initially fused and then split into two sub-branches, which are fed into the SSA and DSA modules, respectively, for spatial modeling. The outputs of SSA and DSA are adaptively fused using a learned weight α derived from the input. This fused spatial attention is then combined with the V component. In parallel, channel-wise attention is modeled using the SE module. Finally, the outputs from the spatial and channel attention paths are merged along the feature dimension to produce a more informative representation.

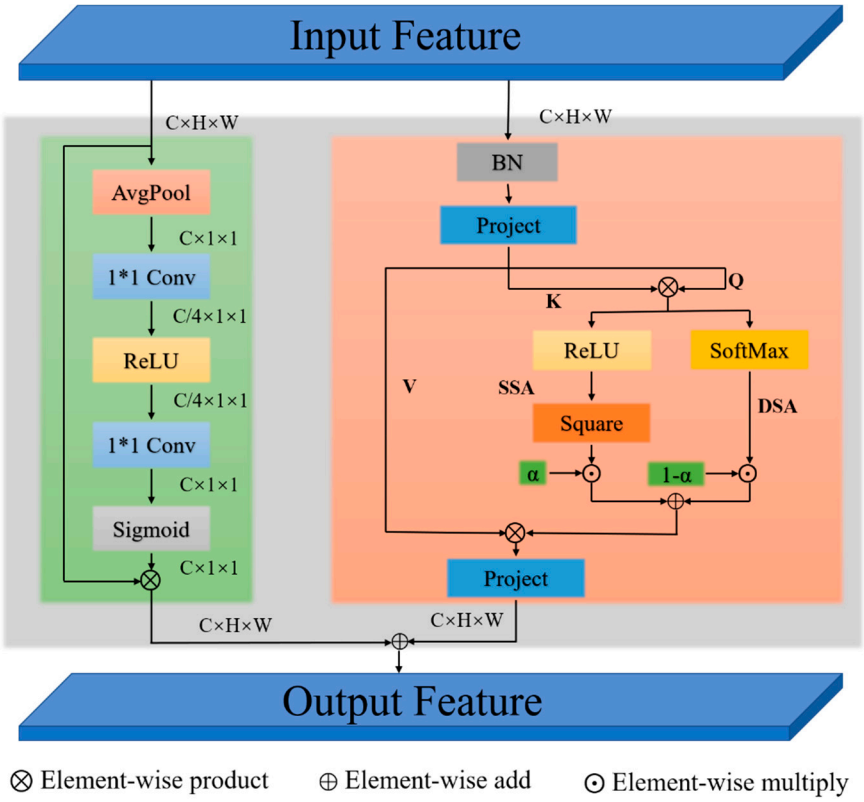


Figure 2. Structure of AAF module.

The entire process can be mathematically formulated as follows:

$$Y = SE(X) + (\alpha \cdot SSA(QK^T) + (1 - \alpha) \cdot DSA(QK^T)) \cdot V, \quad (1)$$

where α is a content-aware weight dynamically computed from the input. The SE block, serving as a lightweight channel attention mechanism, is defined as:

$$SE(X) = AvgPool(Conv(ReLU(Conv(\sigma(X))))) + X, \quad (2)$$

This parallel design maintains the lightweight nature of the module while enabling adaptive selection between sparse and dense attention, as well as complementary modeling of spatial and channel

 doi:10.20944/preprints202510.2237.v1

attention. As a result, the AAF module provides richer and more precise semantic information to benefit the subsequent detection head.

3.3. Efficient Multi-Scale Semantic FPN (EMS-FPN)

To enhance the model's performance in multi-scale object detection tasks, we design an efficient hybrid encoding module based on the Feature Pyramid Network (FPN) for multi-scale feature fusion. However, the traditional FPN architecture only fuses features from the last three layers (P3, P4, P5), neglecting the P2 layer, which contains the richest semantic information crucial for small object detection. This omission significantly limits the model's ability to detect small targets.

A common solution is to directly incorporate the P2 layer into the feature fusion network and integrate it using the same fusion methods as other layers. Nonetheless, since the P2 feature map has a higher spatial resolution, its direct inclusion substantially increases model parameters and computational overhead, thereby impairing real-time detection performance.

To address this, we propose a lightweight and efficient feature fusion architecture termed Efficient Multi-scale Semantic FPN (EMS-FPN), which effectively integrates semantic information from the P2 layer without considerably increasing computation or parameters. As illustrated in Figure 3, (a) depicts the original FPN structure where fusion is performed by a Fusion module excluding P2; (b) shows our proposed EMS-FPN structure. Drawing inspiration from BiFPN's multi-scale weighted fusion mechanism[29], EMS-FPN replaces concatenation operations with element-wise addition to further reduce computation and parameter costs. Additionally, it introduces a feature importance weighting mechanism to enable adaptive selective fusion of multi-scale features, thereby elegantly incorporating the P2 layer into the fusion process.

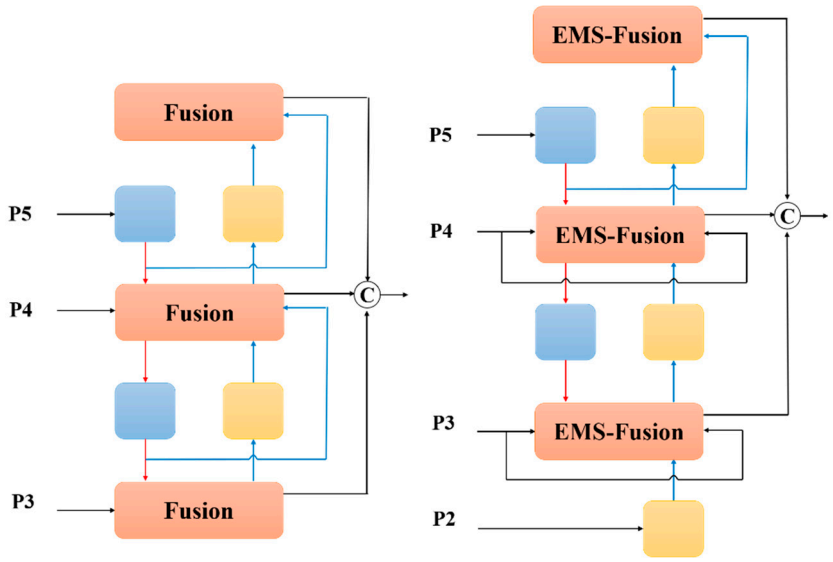


Figure 3. Structural diagram of different feature pyramid networks.

Furthermore, we propose an improved fusion module called EMS-Fusion. This module introduces a Multi-Scale Convolution Block (MSCB)[30] to replace the RepBlock in the original structure, enhancing semantic representation capability (see Figure 4). Moreover, EMS-Fusion integrates the Cross Stage Partial (CSP) design principle to improve feature extraction efficiency and fusion performance, thus strengthening both shallow detail representation and the integrity of deep semantic information.

 doi:10.20944/preprints202510.2237.v1

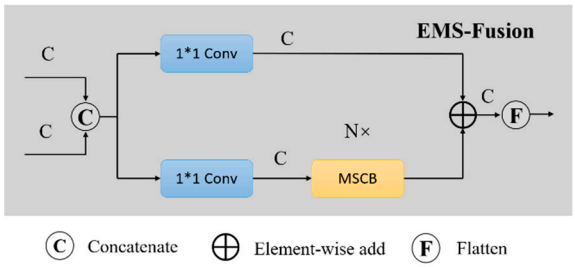


Figure 4. Structure of EMS-Fusion.

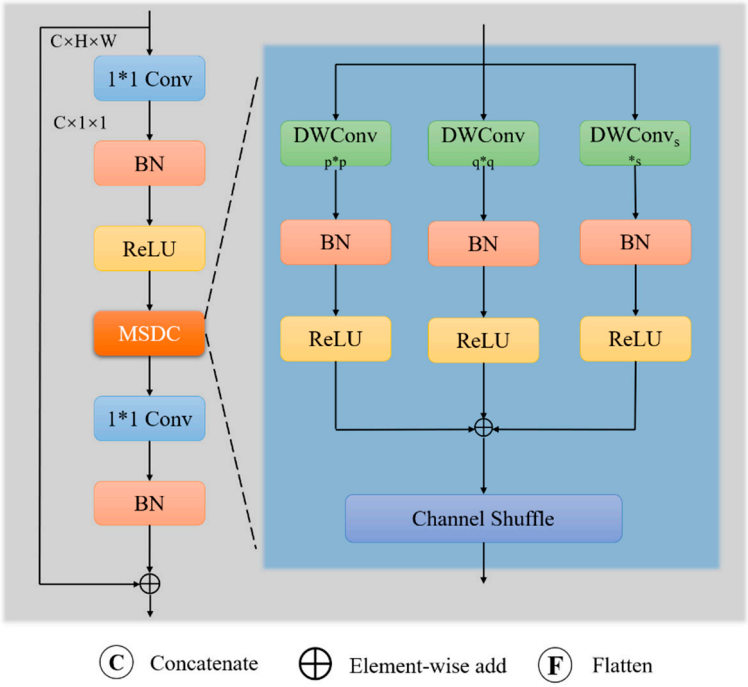
Figure 5 illustrates the detailed structure of the MSCB module. MSCB employs Multi-Scale Depthwise Separable Convolutions (MSDC) to capture contextual information at varying receptive fields and leverages Channel Shuffle to enhance inter-channel communication. Specifically, MSCB first applies a 1×1 convolution to expand channel dimensions, followed by Batch Normalization (BN) and ReLU activation. The features are then split into multiple groups, each processed by depthwise separable convolutions with different kernel sizes (e.g. 3×3, 5×5, 7×7) to extract multi-scale information. Each convolution is followed by BN and ReLU. Subsequently, the outputs undergo channel shuffle to promote cross-channel dependencies, and a 1×1 convolution restores the original channel number. A residual connection adds the original input back to the output to preserve initial information. The overall formulation is:

$$MSCB(X) = X + BN(Conv_{1\times1}(MSDC(ReLU(BN(Conv_{1\times1}(X)))))), \tag{3}$$

where MSDC is defined as:

$$MSDC(X) = ChannelShuffle(\sum_{i} ReLU(BN(DWConv_{i\times i}(X)))) \tag{4}$$

We incorporate a global heterogeneous convolution kernel selection mechanism within the multi-scale convolutions, inspired by TridentNet[31]. Larger receptive fields benefit large object detection, while smaller receptive fields are better suited for small targets. Accordingly, during the FPN stage, we select appropriate convolution kernel sizes for different scale feature maps, effectively adapting to multi-scale objects and progressively building richer spatial context.



 doi:10.20944/preprints202510.2237.v1

Figure 5. Structure of MSCB.

3.4. Local-Aware Gated Module (LAGM)

To enhance the model's feature modeling capabilities in small object and dense scene detection, we propose a lightweight convolutional module with dynamic perception capability, termed the Local-Aware Gated Module (LAGM). This module builds upon and extends the original Gated Convolution (gConv) mechanism.

The gConv module, initially proposed by Song et al.[32], combines Depthwise Separable Convolutions[33] with a gating mechanism. In this structure, the main branch performs feature extraction using pointwise and depthwise convolutions, while the gating branch applies activation functions to dynamically modulate the importance of each information channel. Although gConv offers computational efficiency and selective feature enhancement, its representational power is limited due to the fixed channel dimensionality of the main branch and the absence of intermediate nonlinear transformations, restricting its ability to capture complex semantic information.

To address these limitations, we propose an improved Gated Inverted Convolution (GIConv) module. As illustrated in Figure 6, GIConv introduces an inverted bottleneck architecture: a 1×1 convolution first expands the channel dimension to four times the original size, creating a richer nonlinear transformation space. This is followed by a 3×3 depthwise convolution to capture local spatial features, and finally another 1×1 convolution projects the feature map back to the original channel size. Meanwhile, the gating branch is retained to generate attention maps that model channel-wise feature salience. The entire operation can be formulated as:

$$Y = Conv_{1\times1}^{proj}(\emptyset(DWConv_{3\times3}(\emptyset(Conv_{1\times1}^{expand}(X))))) \cdot G(X), \quad (5)$$

where ϕ denotes the activation function, and G(X) is the gating map generated by the gating branch. This structure not only enhances the diversity of feature transformations but also preserves the dynamic channel selection capability through gating.

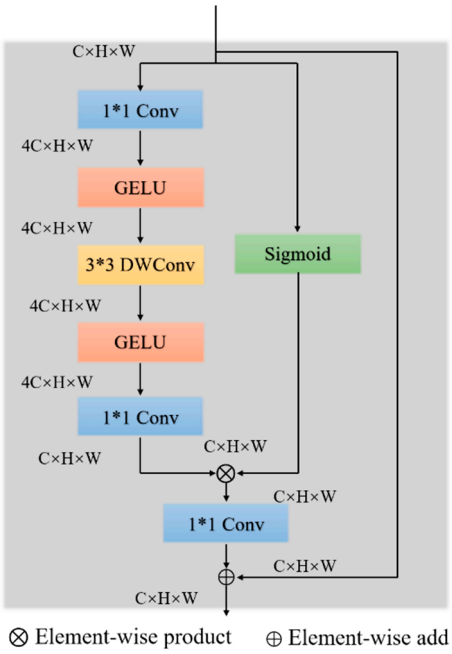


Figure 6. Structure of GIConv.

Building upon the GIConv, we design the LAGM to replace the original Bottleneck structure in the C2f block of the YOLOv8 backbone. While the C2f block is advantageous in terms of speed and parameter efficiency, it suffers from a fixed convolutional path and lacks adaptive modulation mechanisms, making it inadequate for precisely modeling small objects or targets with ambiguous

 doi:10.20944/preprints202510.2237.v1

boundaries. The LAGM module adopts a dual GIConv architecture with double residual connections, forming a global-to-local modeling path that significantly enhances the network's ability to represent multi-scale and multi-semantic features while maintaining a lightweight profile.

We integrate the LAGM module into the backbone to construct a novel PG-Backbone, which inherits the structural style of YOLOv8 while substantially improving overall performance, particularly in small object detection scenarios.

4. Experiments

4.1. Datasets and Experimental Setup

We conduct our experiments on the VisDrone2019 dataset, a large-scale benchmark specifically designed for object detection in aerial views captured by drones[34]. This dataset is released by the Vision and Learning Laboratory at Tianjin University and has been widely adopted in small object detection research.

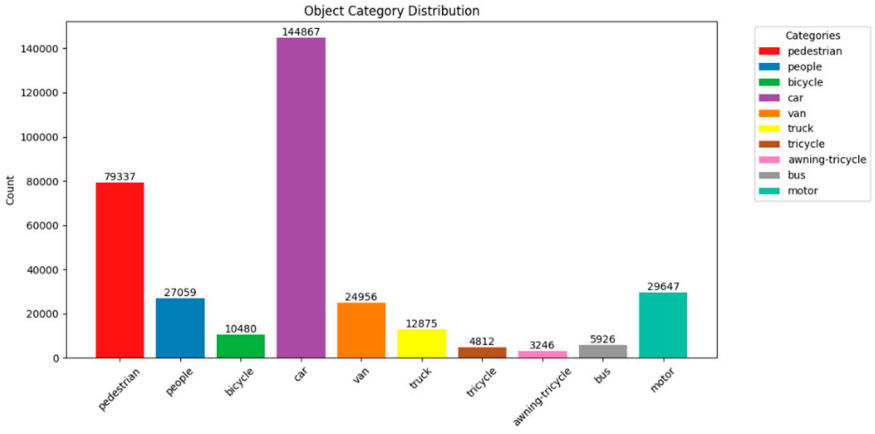
The VisDrone2019 dataset covers diverse urban and suburban scenes from multiple cities across China, including Tianjin, Hong Kong, and Daqing. All images are captured using high-resolution cameras mounted on drones, with a typical resolution of 1920×1080. The dataset is characterized by a wide variety of viewpoints, complex lighting conditions, and densely distributed targets—making it particularly challenging for robust object detection models.

We utilize the detection subset of VisDrone2019, which consists of 10,209 images, split into 6,471 for training, 548 for validation, and 319 for testing. Each image is annotated with bounding boxes, object categories, occlusion levels, truncation ratios, and visible ratios. The dataset defines 10 object categories: pedestrian, person, car, van, bus, truck, motor, bicycle, awning-tricycle, and tricycle.

We conducted a detailed visualization and statistical analysis of the dataset characteristics. As shown in Figure 7a, the most frequent category is car, with a total of 144,867 instances, while the least frequent is awning-tricycle, with only 3,246 instances, indicating a significant class imbalance.

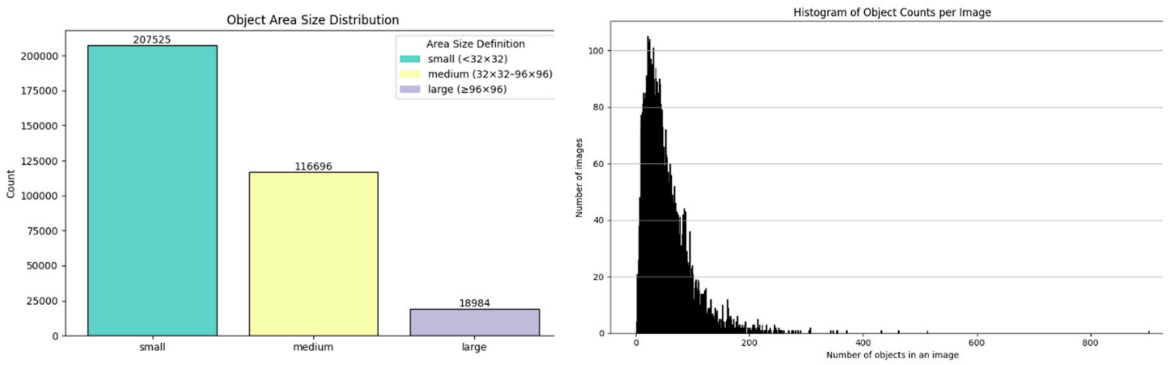
Furthermore, as illustrated in Figure 7b, small objects account for more than half of the total targets, whereas large objects constitute only about 5%. This highlights the dominance of small object instances in the dataset.

Finally, as shown in Figure 7c, most images contain more than 200 annotated targets, indicating a highly dense object distribution, which introduces considerable occlusion and increases the difficulty of detection. These characteristics collectively pose substantial challenges to the robustness and generalization capability of object detection models.



(**a**)

 doi:10.20944/preprints202510.2237.v1



(b) (c)

Figure 7. Overview of the VisDrone2019 dataset characteristics: (a) Class distribution indicates significant imbalance, with the car class dominating the dataset; (b) Object size distribution reveals a predominance of small objects, challenging for detectors; (c) Most images contain more than 200 annotated instances, highlighting the density and potential occlusion in UAV-captured scenes.

Since the original model is intended for deployment on embedded UAV devices with limited computational resources and memory, we focus exclusively on optimizing a lightweight variant of RT-DETR. Both training and inference are performed on images resized to 640×640.

Table 1. Hardware and software environment used in the experiments.

Environment	Specification
CPU	Intel(R) Xeon(R) Platinum 8352
GPU	vGPU-32GB
VRAM	32GB
RAM	90GB
Operating System	Ubuntu 22.04
Language	Python 3.12
Framework	PyTorch 2.5.1
CUDA Version	12.4

Table 2. Hyperparameter settings used in the experiments.

Environment	Specification
Input Size	640×640
Batch Size	16
Training Epochs	300
Optimizer	AdamW
Initial Learning Rate	0.0001
Learning Rate Factor	0.01
Momentum	0.9
Warmup Steps	2000

To ensure fair comparisons, all experiments are conducted on a workstation equipped with a vGPU featuring 32GB of memory, an Intel(R) Xeon(R) Platinum 8352 CPU running at 2.10GHz, and 90GB of RAM. The operating system is Ubuntu 22.04. All models are trained and evaluated in an environment based on CUDA 12.4, Python 3.12, and PyTorch 2.5.1. The detailed hardware configuration is summarized in Table 1. The hyperparameter settings used in training are listed in Table 2. Input images are resized to 640×640 pixels. We adopt a batch size of 16 and train all models for 300 epochs using the AdamW optimizer. The initial learning rate is set to 0.0001 with a learning

 doi:10.20944/preprints202510.2237.v1

rate factor of 0.01 and a momentum of 0.9. Additionally, a warmup phase of 2000 steps is applied to stabilize training.

4.2. Ablation Experiment

Based on the original RT-DETR model, this paper proposes an Adaptive Attention Fusion module. Additionally, an efficient Multi-scale Convolutional Feature Pyramid Network is introduced. Building upon the proposed Gated Inverted Bottleneck Convolution, we further develop a locally-aware gated bottleneck structure and reconstruct the backbone network of YOLOv8 accordingly, resulting in a novel backbone named PG-Backbone. This new backbone maintains a low parameter count and computational complexity.

RT-DETR-R18 and RT-DETR-R50 serve as baseline models to systematically evaluate the impact of each proposed module. Ablation studies utilize Precision (P), Recall (R), mean Average Precision at 50% IoU (mAP50), parameter count, and computational cost (FLOPs) as evaluation metrics. These metrics are widely recognized in object detection research for their ability to comprehensively reflect the influence of individual components on model performance.

Table 3. Ablation study results for LEA-DETR-S on the VisDrone2019 dataset.

Data	PG-Backbone	AAF	EMS-FPN	P	R	mAP@50	Params	FLOPs
Val	–	–	–	59.1	43.8	45.0	20.1M	58.7G
	√	–	–	57.9	42.9	44.1	10.0M	30.7G
	√	√	–	59.2	43.1	45.1	10.8M	31.6G
	√	√	√	**61.0**	**45.2**	**46.3**	**9.2M**	**27.3G**
Test	–	–	–	52.0	36.3	34.5	20.1M	58.7G
	√	–	–	52.0	35.4	33.7	10.0M	30.7G
	√	√	–	53.5	36.7	35.0	10.8M	31.6G
	√	√	√	**54.1**	**37.3**	**36.5**	**9.2M**	**27.3G**

* Bold values indicate the best results.

The ablation results of the proposed LEA-DETR are presented in Tables 3 and 4, where the best results are highlighted in bold. Tables 3 and 4 respectively show the performance of two lightweight models of different scales after sequentially applying PG-Backbone, the AAF module, and the EMS-FPN module.

Table 4. Ablation study results for LEA-DETR-M on the VisDrone2019 dataset.

Data	PG-Backbone	AAF	EMS-FPN	P	R	mAP@50	Params	FLOPs
Val	–	–	–	60.9	46.5	47.9	42.8M	134.5G
	√	–	–	60.2	46.0	47.2	**13.7M**	**50.0G**
	√	√	–	60.9	46.8	48.1	14.6M	50.9G
	√	√	√	**61.7**	**47.9**	**49.2**	14.7M	57.0G
Test	–	–	–	54.2	38.3	37.5	42.8M	134.5G
	√	–	–	53.7	37.8	36.7	**13.7M**	**50.0G**
	√	√	–	55.4	38.8	38.1	14.6M	50.9G
	√	√	√	**56.8**	**40.4**	**39.2**	14.7M	57.0G

* Bold values indicate the best results.

Initially, integrating PG-Backbone significantly reduces both the parameter count and computational cost, albeit with a slight decrease in accuracy. Subsequently, as the AAF and EMS-FPN modules are progressively incorporated, both models demonstrate steady improvements in accuracy on the validation and test sets. This indicates that the proposed modules effectively complement each other, jointly enhancing detection performance while preserving inference speed.

 doi:10.20944/preprints202510.2237.v1

4.3. Comparisons with Other Object Detection Networks

This work focuses on object detection methods for UAV aerial imagery that are better suited to real-world engineering scenarios, characterized by low hardware requirements and strong generalization capability. Therefore, the comparison models selected in this paper are limited to one-stage and end-to-end real-time object detectors, which are easy to deploy and have been widely adopted in various industrial and practical applications.

Based on differences in detection mechanisms, we compare our method with state-of-the-art approaches from two mainstream families: YOLO-based and DETR-based detectors. Specifically, we include several representative YOLO models and their advanced variants, including YOLOv8-M, YOLOv8-L[12], YOLOv11-S, YOLOv11-M[15], Drone-YOLO-S[26], HIC-YOLO[35] and DFAS-YOLO[36]. For DETR-based models, we include DETR[16], Deformable DETR[17], Sparse DETR[37], and RT-DETR[19]. These models represent a broad coverage of the current landscape and have been thoroughly validated in the literature for their robustness and effectiveness across diverse scenarios.

To comprehensively evaluate the performance of the proposed LEA-DETR, we adopt standard metrics including mAP@0.5 (mAP50), mAP@0.5:0.95 (mAP50-95), model parameters, and FLOPs. The results are summarized in Table 5.

Table 5. Comparison of Detection Accuracy, Model Size, and Computation

Model	Input Size	mAP@50	mAP@50:95	Params	FLOPs
YOLOv8-M	640×640	40.7	24.6	25.9M	78.9G
YOLOv8-L	640×640	42.7	26.1	43.7M [1]	165.2G
YOLOv11-S	640×640	38.7	23.0	9.4M	**21.3G**
YOLOv11-M	640×640	43.1	25.9	20.0M	67.7G
Drone-YOLO-S	640×640	44.3	27.0	10.9M	–
HIC-YOLOv5	640×640	44.3	26.0	9.4M	31.2G
DFAS-YOLO	640×640	44.8	27.3	7.5M	–
Deformable DETR	1333×800	43.1	27.1	40.0M	173.0G
RT-DETR-R18	640×640	45.0	27.2	20.1M	58.7G
RT-DETR-R50	640×640	47.9	29.3	42.8M	134.5G
UAV-DETR-R18	640×640	48.8	29.8	20.0M	77.0G
LEA-DETR-S(Ours)	640×640	46.3	28.5	**9.2M**	27.3G
LEA-DETR-M(Ours)	640×640	**49.2**	**30.5**	14.7M	57.0G

* Bold values indicate the best results.

In comparison with the YOLO series models, LEA-DETR-S demonstrates significant advantages. Compared to YOLOv8-M, LEA-DETR-S improves mAP50 and mAP50-95 by 5.6 and 3.9 percentage points respectively, while reducing the parameter count by approximately 54% and lowering FLOPs by 65%. Even when compared with the larger YOLOv8-L, LEA-DETR-S still achieves superior accuracy, with only 27% of its parameter count and 17% of its computational cost, reflecting a better trade-off between efficiency and performance. Compared with YOLOv11-S, which has a similar model size, LEA-DETR-S performs noticeably better in detection accuracy, validating the representational power and modeling advantages of our architecture under lightweight constraints. Even against the mid-sized model YOLOv11-M, LEA-DETR-S achieves a better balance between accuracy and resource consumption. In comparison with the lightweight optimized model HIC-YOLOv5, LEA-DETR-S achieves a 3.0 point improvement in mAP50 and a 2.5 point improvement in mAP50-95 with a comparable number of parameters, while also requiring fewer FLOPs. This demonstrates the higher computational efficiency and stronger feature representation capability of our architectural design.

Furthermore, we also compare LEA-DETR with several Transformer-based detectors. Compared to the baseline RT-DETR-R18, LEA-DETR-S achieves improvements in both mAP50 and mAP50-95, while reducing model size by approximately 54% and FLOPs by more than 50%, showing

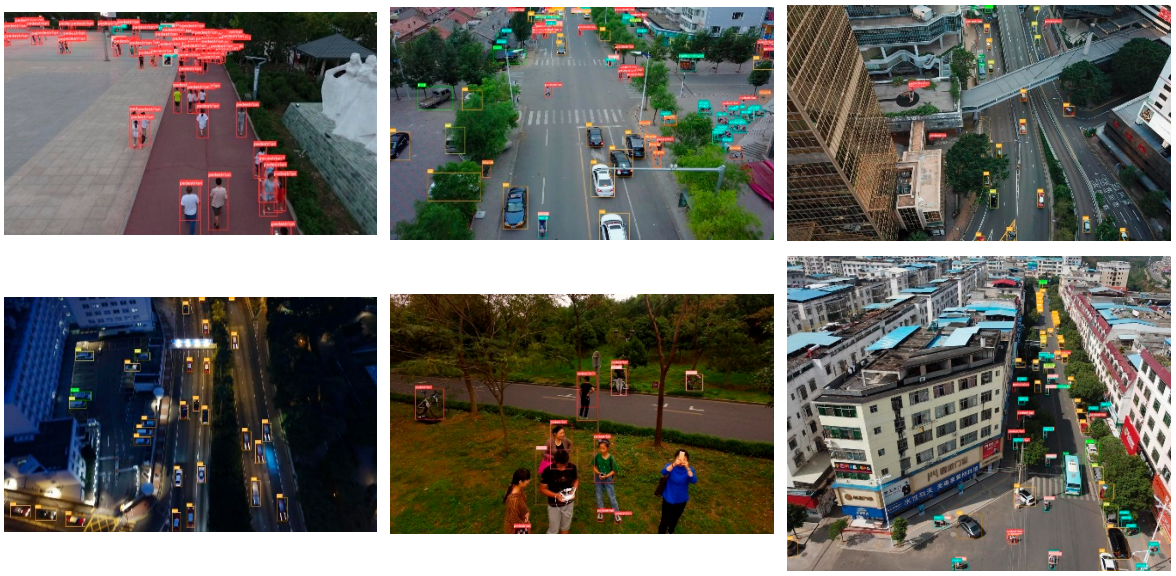
 doi:10.20944/preprints202510.2237.v1

a strong advantage in terms of lightweight design. When compared with more computationally expensive models such as Deformable DETR, RT-DETR-R50, and UAV-DETR-R18, LEA-DETR-M still achieves superior or comparable detection accuracy while significantly reducing the number of parameters and computational burden. This further confirms the effectiveness of our method in achieving an excellent balance between accuracy and resource efficiency.

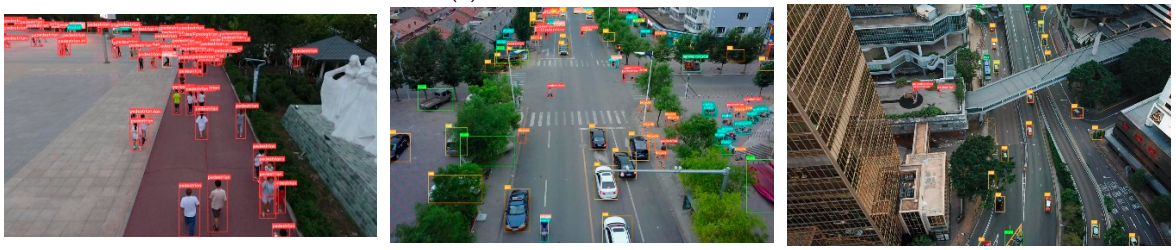
In summary, the proposed LEA-DETR achieves superior detection accuracy with significantly reduced resource consumption compared to a wide range of detectors. This makes it particularly suitable for deployment on platforms with limited computational resources, such as embedded systems, UAVs, and mobile devices, where real-time performance and energy efficiency are critical. Compared with existing lightweight detectors, our method exhibits a more rational architectural design and stronger feature representation capabilities. It achieves an effective balance between accuracy, parameter count, and computational complexity, making it a highly practical and scalable solution for real-world scenarios.

4.4. Visualization

In Figure 8, we showcase the performance of LEA-DETR-S on the VisDrone dataset across a variety of challenging scenarios, including densely populated scenes, cluttered backgrounds, close-range views, low-light nighttime environments, and scenes with extremely small objects. These scenarios encompass crowded pedestrian and vehicle areas, tall buildings and bridges, and complex intersections.



(a) RT-DETR-R18



 doi:10.20944/preprints202510.2237.v1

(b) LEA-DETE-S

Figure 8. Visualization results of RT-DETR-R18 and LEA-DETR-S on VisDrone2019-test.

In scenes with dense targets and complex backgrounds, objects are often occluded by other objects or the background itself, which can easily lead to missed detections. Compared to the baseline model, LEA-DETR-S demonstrates significantly fewer missed detections, indicating improved performance under these challenging conditions.

In close-range scenes, the model faces a generalization challenge, as the training data primarily contains small-scale pedestrian instances. As illustrated in the figures, the baseline model tends to miss close-up pedestrians more frequently than LEA-DETR-S.

Under low-light conditions, targets are more likely to blend into dark backgrounds, making detection more difficult. As shown in the visualizations, LEA-DETR-S maintains robust detection performance even in dimly lit environments.

5. Conclusions

In this paper, we propose LEA-DETR, a lightweight and efficient object detection model tailored for UAV-based applications, built upon the RT-DETR architecture. Through three key innovations, LEA-DETR achieves significant reductions in both parameter count and computational cost, while maintaining or even improving detection accuracy. This makes it particularly suitable for scenarios where both efficiency and precision are critical.

First, we introduce the AAF (Adaptive Attention Fusion) module, which employs an early fusion strategy to adaptively integrate sparse and dense attention mechanisms. By incorporating channel attention, AAF enhances the model's perception of critical spatial regions and semantic dimensions, thereby improving feature selection accuracy.

Second, the EMS-FPN module redefines the feature fusion strategy by incorporating efficient multi-scale convolutions. Additionally, it integrates the P2 feature layer into the fusion hierarchy, significantly enhancing the model's ability to capture fine-grained semantic information crucial for small object detection.

Third, we design the GIConv module, a convolutional block based on inverted bottleneck and depthwise separable convolutions. By embedding a gating mechanism, GIConv achieves highly efficient local feature modeling. Building upon GIConv, we further propose the LAGM module as a drop-in replacement for the C2f Bottleneck in the YOLOv8 backbone. This leads to the construction of a new backbone architecture, termed PG-Backbone, which drastically reduces the parameter count and computational load compared to the baseline, while preserving accuracy.

Overall, LEA-DETR demonstrates superior performance in UAV object detection tasks, achieving higher accuracy with significantly reduced resource consumption on public benchmarks. In future work, we plan to explore the integration of lightweight design principles with cross-scale feature fusion to further improve model robustness and generalization in complex environments. We also aim to enhance its deployment in real-time and edge-based detection scenarios.

Author Contributions: Conceptualization, Haohao Ma.; methodology, Haohao Ma; software, Haohao Ma; validation, Haohao Ma; formal analysis, Haohao Ma; investigation, Haohao Ma; resources, Haohao Ma; data

 doi:10.20944/preprints202510.2237.v1

curation, Haohao Ma writing—original draft preparation, Haohao Ma; writing—review and editing, Mingliang Zuo; visualization, Haohao Ma; supervision, Qiqi Ge; project administration, Qiqi Ge; funding acquisition, Qiqi Ge. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data that support the findings of this study are available in http: //www.aiskyeye.com/ at ICCV 2019 Open Access Repository, reference number [34].

Acknowledgments: Thanks are due to Qiqi Ge for supporting experiment device.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DETR	Detection Transformer
AAF	Adaptive Attention Fusion
FPN	Feature Pyramid Networks
LAGM	Local-Aware Gated Module

References

1. Feng, J.; Wang, J.; Qin, R. Lightweight detection network for arbitrary oriented vehicles in UAV imagery via precise positional information encoding and bidirectional feature fusion. *International Journal of Remote Sensing* **2023**, *44*, 4529–4558.
2. Bhadra, S.; Sagan, V.; Sarkar, S.; Braud, M.; Mockler, T.C.; Eveland, A.L. Prosail-net: A transfer learning-based dual stream neural network to estimate leaf chlorophyll and leaf angle of crops from UAV hyperspectral images. *ISPRS Journal of Photogrammetry and Remote Sensing* **2024**, *210*, 1–24.
3. Alexan, W.; Aly, L.; Korayem, Y.; Gabr, M.; El-Damak, D.; Fathy, A.; Mansour, H.A.A. Secure communication of military reconnaissance images over UAV-assisted relay networks. *IEEE Access* **2024**, *12*, 78589–78610.
4. Cao, Z.; Kooistra, L.; Wang, W.; Guo, L.; Valente, J. Real-time object detection based on UAV remote sensing: A systematic literature review. *Drones* **2023**, *7*, 620. Available online: https://www.mdpi.com/2504-446X/7/10/620.
5. Zhang, Z. Drone-yolo: An efficient neural network method for target detection in drone images. *Drones* **2023**, *7*, 526. Available online: https://www.mdpi.com/2504-446X/7/8/526.
6. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
7. Girshick, R. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2016**, *39*, 1137–1149.
9. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
10. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
11. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
12. Jocher, G.; Chaurasia, A.; Stoken, A.; Borovec, J.; Kwon, Y.; Michael, K.; Fang, J.; Wong, C.; Yifu, Z.; Montes, D.; et al. ultralytics/yolov5: v6.2 - YOLOv5 classification models, Apple M1, reproducibility, ClearML and Deci.ai integrations. *Zenodo* **2022**.
13. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.

14. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. 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*, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
15. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLO (Version 8.0.0) [Computer Software]. **2023**.
16. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision*; Springer: Glasgow, UK, 23–28 August 2020; pp. 213–229.
17. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
18. Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L.M.; Zhang, L. DN-DETR: Accelerate DETR training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 18–24 June 2022; pp. 13619–13627.
19. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974.
20. Zhai, X.; Huang, Z.; Li, T.; Liu, H.; Wang, S. YOLO-Drone: An optimized YOLOv8 network for tiny UAV object detection. *Electronics* **2023**, *12*, 3664.
21. Sun, W.; Dai, L.; Zhang, X.; Chang, P.; He, X. RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring. *Applied Intelligence* **2022**, *53*, 1–16.
22. Wang, G.; Chen, Y.; An, P.; Hong, H.; Hu, J.; Huang, T. UAV-YOLOv8: A small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios. *Sensors* **2023**, *23*, 7190.
23. Zeng, S.; Yang, W.; Jiao, Y.; Geng, L.; Chen, X. SCA-YOLO: A new small object detection model for UAV images. *The Visual Computer* **2024**, *40*, 1787–1803.
24. Zhao, H.; Zhang, H.; Zhao, Y. YOLOv7-Sea: Object detection of maritime UAV images based on improved YOLOv7. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, HI, USA, 2–7 January 2023; pp. 233–238.
25. Zhang, H.; Liu, K.; Gan, Z.; Zhu, G.-N. UAV-DETR: Efficient end-to-end object detection for unmanned aerial vehicle imagery. *arXiv* **2025**, arXiv:2501.01855.
26. Kong, Y.; Shang, X.; Jia, S. Drone-DETR: Efficient small object detection for remote sensing image using enhanced RT-DETR model. *Sensors* **2024**, *24*, 5496.
27. Liu, Y.; He, M.; Hui, B. ESO-DETR: An improved real-time detection transformer model for enhanced small object detection in UAV imagery. *Drones* **2025**, *9*, 143.
28. Zhou, S.; Chen, D.; Pan, J.; Shi, J.; Yang, J. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 16–22 June 2024; pp. 2952–2963.
29. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
30. Rahman, M.M.; Munir, M.; Marculescu, R. EMCAD: Efficient multi-scale convolutional attention decoding for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 16–22 June 2024; pp. 11769–11779.
31. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z. Scale-aware trident networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6054–6063.
32. Song, Y.; Zhou, Y.; Qian, H.; Du, X. Rethinking performance gains in image dehazing networks. *arXiv* **2022**, arXiv:2209.11448.
33. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
34. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF*

 doi:10.20944/preprints202510.2237.v1

International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019; pp. 0–0.

35. Tang, S.; Zhang, S.; Fang, Y. HIC-YOLOv5: Improved YOLOv5 for small object detection. In *Proceedings of the 2024 IEEE International Conference on Robotics and Automation*, Yokohama, Japan, 13–17 May 2024; pp. 6614–6619.
36. Liu, X.; Zhou, S.; Ma, J.; Sun, Y.; Zhang, J.; Zuo, H. DFAS-YOLO: Dual Feature-Aware Sampling for Small-Object Detection in Remote Sensing Images. *Remote Sens.* **2025**, *17*, 3476. https://doi.org/10.3390/rs17203476
37. Roh, B.; Shin, J.; Shin, W.; Kim, S. Sparse DETR: Efficient end-to-end object detection with learnable sparsity. *arXiv* **2021**, arXiv:2111.14330. https://doi.org/10.48550/arXiv.2111.14330