

Article

EFSL-YOLO: An Improved Model for Small Object Detection in UAV Vision

Meng Zhou ¹, Shuke He ¹, Chang Wang ^{2,3} and Jing Wang ^{1,*}¹ School of Electrical and Control Engineering, North China University of Technology, Beijing 100144, China; zhoumeng@ncut.edu.cn (M.Z.)² Beijing Aerospace Automatic Control Institute, Beijing 100854, China³ College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China

* Correspondence: jwang@ncut.edu.cn

Highlights

What are the main findings?

- An improved YOLOv13 framework is proposed for small object detection in UAV aerial imagery, integrating a multi-stage feature enhancement mechanism to address dense occlusion, complex backgrounds, and weak target features.
- Experimental validation on the VisDrone dataset shows that the proposed method outperforms the baseline YOLOv13 by 6.2% in mAP and 5.0% in recall.

What are the implications of the main findings?

- The proposed mechanism (EFFA-Net, SwiGLUConv, and LF-MSCF) effectively enhances small-target representation and anti-interference capability, providing a practical solution for detection in complex aerial environments.
- This work enriches the technical approaches to UAV-based small-target detection, offering a valuable reference for algorithm optimization in similar aerial scenarios.

Abstract

To address the challenges in UAV remote sensing imagery, such as small object size, dense occlusion and complex background interference, this paper proposes an enhanced small object detection algorithm based on an improved YOLOv13 model for drone applications in complex weather environments. First, an enhanced feature fusion attention network (EFFA-Net) is designed in the preprocessing stage to reduce image degradation and suppress the interference caused by smoke and haze. Then, in the backbone, a swish-gated convolution (SwiGLUConv) module is designed to adaptively expand the receptive field and enhance multi-scale feature extraction, which strengthens the representation of small targets while maintaining efficient computation. Furthermore, a locally enhanced multi-scale context fusion (LF-MSCF) module is integrated into the feature fusion neck of YOLO, combining multi-head self-attention, channel attention, and spatial attention to suppress background noise and redundant responses, thereby improving detection accuracy. Extensive experiments on the VisDrone-DET2019 dataset, UAVDT dataset, and HazyDet dataset demonstrate that the proposed algorithm outperforms other mainstream methods, showcasing excellent detection accuracy and robustness in complex UAV aerial scenarios.

Keywords: UAV vision; small object detection; SwiGLUConv; LF-MSCF

Academic Editor: Thomas P Karnowski

Received: 4 February 2026

Revised: 24 March 2026

Accepted: 24 March 2026

Published: 27 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

In recent decades, with the rapid advancement of object detection and unmanned aerial vehicle (UAV) technologies, aerial vision has emerged as a rapidly evolving field of research within computer vision [1]. UAVs equipped with onboard vision systems are widely applied in intelligent transportation, forest fire surveillance, power line inspection, emergency rescue, and smart agriculture [2]. By integrating real-time object detection algorithms, UAVs can automatically analyze aerial imagery, perform autonomous decision-making, and enhance situational awareness in complex environments [3,4].

However, UAV aerial images differ significantly from natural-scene photographs. They typically contain a large proportion of small targets; exhibit dense spatial distributions; and involve complex background interference caused by varying illumination, smoke, haze, and occlusions. Due to these characteristics, conventional detection algorithms often suffer from low recall rates and poor robustness when applied to UAV scenarios. The small size of objects also leads to weak semantic features and vanishing gradients in deep layers, which makes it difficult for detectors to maintain discriminative representations [5].

Recent developments in deep learning, especially the YOLO family, have greatly improved detection accuracy and speed. Nevertheless, standard YOLO architectures remain limited in UAV applications due to three main issues: insufficient multi-scale feature representation for small targets, degraded feature extraction under complex weather conditions, and limited receptive field and weak global–local interaction during downsampling [6].

The current mainstream object detection algorithms based on UAVs can be mainly divided into Transformer-based object detection algorithms and CNN-based object detection algorithms. For transformer-based object detection algorithms, these models have gained substantial attention in object detection and driven notable advancements in UAV perception tasks. In [7], to tackle the difficulty of extracting small object features in occluded, overlapping complex backgrounds for AAV-borne traditional real-time detectors, the research team proposed FM-RTDETR, a dedicated real-time detection algorithm optimized for small object detection. The Mamba-in-Transformer DETR (MIT-DETR) framework was proposed in [8]. This framework extracts global information with linear complexity via Mamba, further divides visual sentences into visual words and analyzes local information with the Transformer, and designs a multi-scale dilated attention (MSDA) module to obtain abundant multi-scale information at a low cost. Building on the demand for balanced performance, OWRT-DETR and its lightweight variant [9] offer a well-calibrated trade-off between accuracy, efficiency, and practical adaptability, supporting missions like maritime monitoring and search-and-rescue. Similarly, Drone-DETR [10] is a lightweight, high-precision model optimized for embedded UAV systems, while ESO-DETR [11] integrates theoretical advancements with engineering-oriented lightweight design to reconcile efficiency and detection strength. Complementarily, UAV-DETR [12] is tailored for drone aerial imagery, addressing spatial–contextual alignment challenges under dynamic conditions and balancing accuracy with lightweight architecture, demonstrating strong theoretical innovation and practical value.

For CNN-based object detection algorithms, despite the aforementioned advantages of two-stage and Transformer-based detectors, they often struggle to balance accuracy and computational efficiency. In contrast, single-stage object detectors, such as SSD, RetinaNet and the YOLO series, have become the mainstream choice for real-time visual tasks in UAV applications, owing to their high inference speed and strong resource efficiency [13]. In [14], an enhanced SDWS-YOLO model based on YOLOv8 is proposed, which achieves a balanced improvement in performance and efficiency via SDWS module replacement, feature loss control, Four-FPN fusion, detection head simplification and composite WFS-IoU loss optimization. In [15], a multi-scale feature fusion network is proposed to address

scale variation challenges inherent in UAV small-target imagery. In [16], a lightweight detection model named ESGN-YOLO is proposed, which achieves the performance improvement of multi-scale small object detection by virtue of the Efficient Feature Fusion Module (EFFM), Shift-Wise Convolution (SWC) Bottleneck and Group Normalisation Scale Head (GNSH). Moreover, the method proposed in [17] integrates object-oriented information with wave-function theory, addressing persistent issues such as insufficient discriminative features, weak global representation, and suboptimal multi-scale fusion in UAV small-target detection.

Despite significant advancements in deep learning-based detection, UAV small-object recognition remains challenging due to inherent difficulties like occlusion and limited pixel information, leading to frequent missed or false detections. Moreover, accuracy-oriented improvements often come at high computational cost, which is prohibitive for resource-limited UAV platforms. To address this, in this paper, an enhanced YOLOv13 framework integrating SwiGLUConv and the LF-MSCF module is proposed to reduce missed and false detections while maintaining lightweight, real-time performance.

Compared with existing feature-enhancement modules (such as CBAM, SE-Net), the SwiGLUConv module combines the SwiGLU activation function with convolution operations, which not only enhances the feature extraction capability but also reduces the computational cost. Compared with existing multi-scale fusion methods (such as FPN, PANet), the LF-MSCF module adopts a hybrid attention mechanism combining channel attention, adaptive multi-scale selection, spatial attention, and multi-head self-attention with a feed-forward fusion network, which can simultaneously model inter-channel dependencies, multi-scale receptive fields, local spatial saliency, and long-range global context. By integrating local convolution enhancement and global context modeling in a unified residual structure, LF-MSCF achieves stronger feature refinement and detail preservation for small targets, effectively alleviating feature dilution and occlusion loss in complex UAV detection scenarios.

To overcome these challenges, this paper proposes an enhanced YOLOv13-derived small object detection method specifically designed for UAV vision under complex atmospheric environments, targeting the core challenges of dense occlusion, complex background interference, and weak small-target features in aerial scenes. The main contributions of this paper are summarized as follows:

1. An enhanced feature fusion attention network (EFFA-Net) is applied in preprocessing to reduce the effects of haze and atmospheric scattering, providing cleaner inputs for subsequent detection stages.
2. A Swish-Gated Convolution (SwiGLUConv) module is designed for the backbone network to enhance multiscale feature extraction and adaptively expand the receptive field, improving the expressiveness of small-object representations.
3. A locally enhanced multi-scale context fusion (LF-MSCF) module is incorporated into the neck, combining multi-head self-attention, channel attention, and spatial attention mechanisms to suppress redundant responses and background noise.
4. Extensive ablation experiments and comparative experiments are conducted on three public UAV detection datasets, verifying the effectiveness and robustness of each proposed module in handling complex UAV detection scenarios.

2. Materials and Methods

In this section, the overall architecture of the improved YOLOv13 model is presented, with a focus on the design of its key modules. The integration of EFFA-Net, SwiGLUConv, and LF-MSCF is detailed, highlighting their contributions to improved detection performance.

2.1. Overall Framework

The overall structure of the proposed improved YOLOv13 model for UAV small object detection is depicted in Figure 1. This framework comprises three core components:

1. To address the challenges of haze and low contrast in UAV imagery, a novel EFA-Net is designed in image preprocessing.
2. Backbone feature extraction with the proposed SwiGLUConv module to adaptively capture multi-scale contextual information.
3. Detection head enhancement via the LF-MSCF module to integrate global–local semantics for small-object discrimination.

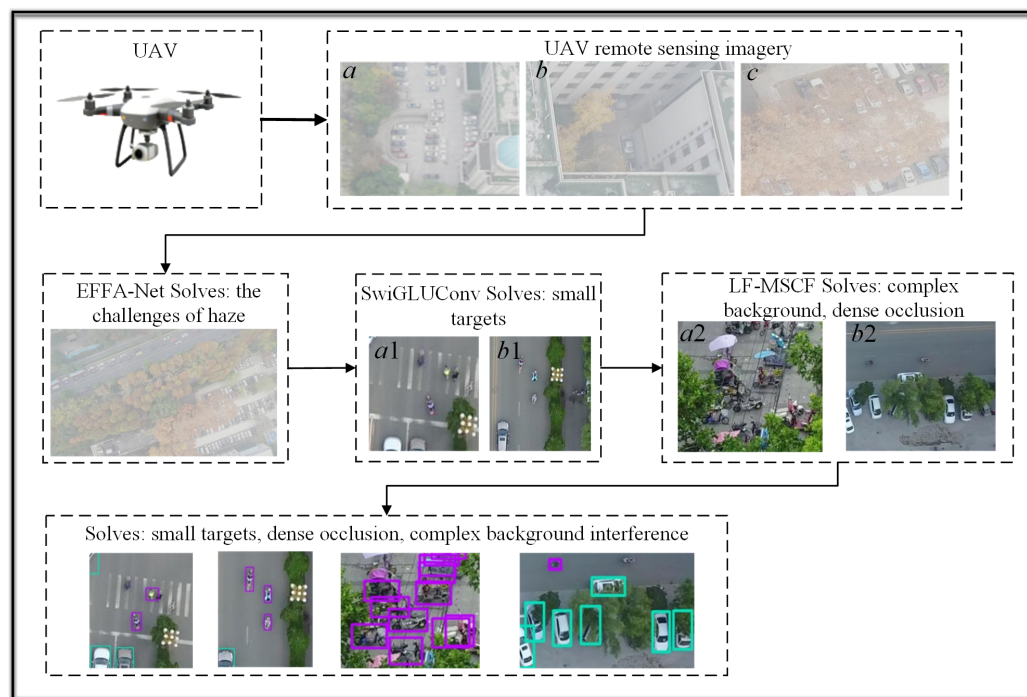


Figure 1. Structure of the proposed system, where (a) denotes the small-target scene under foggy conditions, (b) denotes the complex background scene under foggy conditions, and (c) denotes the heavily occluded scene under foggy conditions. (a1,b1) represent the small-target scene on the street, (a2) represents the complex background scene, and (b2) represents the occluded scene. The purple and turquoise boxes in the figure represent motors and cars, respectively, the two categories detected by the model.

2.2. Datasets

Experiments are carried out on three benchmark datasets: VisDrone [18], UAVDT [19], and HazyDet [20].

Among them, VisDrone serves as a standardized benchmark for UAV-based object detection in real-world environments, with a particular emphasis on small object detection; over 60% of the objects are smaller than 32×32 pixels. It was partitioned into 6471 training, 548 validation, and 3190 testing images, covering ten categories. The class distribution is as follows: cars (42.2%), pedestrians (23.1%), people (7.9%), motors (8.6%), vans (7.2%), trucks (3.8%), bicycles (3.1%), tricycles (1.4%), buses (0.7%), and awning-tricycles (0.9%). The dataset encompasses a diverse range of scenarios and provides richly annotated samples sufficient for model training and evaluation. The dataset is publicly available at <https://github.com/VisDrone/VisDrone-Dataset> (accessed on 1 January 2026).

UAVDT is another notable dataset designed for complex UAV scenarios, released by the University of Chinese Academy of Sciences. This study utilizes its official split of 23,258 training and 15,069 testing images. As no validation set is provided in the

benchmark, hyperparameter tuning was performed via cross-validation on the training subset. As a large-scale vehicle detection and tracking dataset, UAVDT focuses on three predefined vehicle categories: car, truck, and bus. Cars dominate the dataset, comprising approximately 95% of all annotated objects, followed by trucks (4.0%) and buses (1.0%). In terms of object size, small objects account for 61.9%, medium objects account for 36.3%, and large objects account for 1.8%. The dataset is publicly available at <https://github.com/dataset-ninja/uavdt> (accessed on 10 February 2026).

The HazyDet dataset is purpose-built for object detection under hazy conditions, featuring three categories: car, truck, and bus, with over 75% of the objects being small ($\leq 32 \times 32$ pixels). The dataset follows its prescribed split of 8000 training, 1000 validation, and 2000 testing images. HazyDet is designed to evaluate two key capabilities of detection models: accuracy and robustness under degraded visibility, making it particularly suitable for validating UAV-based detection algorithms in adverse weather conditions. The dataset is publicly available at <https://github.com/GrokCV/HazyDet> (accessed on 15 February 2026). Regarding data augmentation, no additional geometric or photometric transformations (e.g., random flipping, cropping, or color jittering) were applied during the training phase. This protocol was intentionally maintained to assess the model's inherent robustness and to ensure a fair comparison with baseline methods under raw input conditions. The class distributions for all datasets correspond to the official annotations, covering ten categories for VisDrone and three vehicle-focused categories for both UAVDT and HazyDet. Detailed statistics regarding total instances and category counts are summarized in Table 1, while fine-grained per-class distributions follow the specifications in the respective dataset publications.

Table 1. Details of the datasets used in UAV detection experiments.

Dataset Name	Total Images	Data Split (Train/Val/Test)	Object Categories	Instances
VisDrone	10,209	6471/548/3190	10	540,000
UAVDT	38,327	23,258/-/15,069	3	840,000
HazyDet	11,000	8000/1000/2000	3	383,000

2.3. EFSL-YOLO

2.3.1. Architecture

UAV-based object detection in adverse weather conditions demands real-time performance while maintaining high accuracy for small, obscured targets. The YOLO series models have long been recognized for their superior balance between detection speed and end-to-end trainability compared to other detection architectures. Furthermore, YOLOv13 has demonstrated strong capabilities in object localization and recognition across diverse scenarios, along with computational and deployment advantages. Based on these considerations, the improved network model proposed in this study is an enhancement of YOLOv13. For UAV-based small object detection under hazy conditions, YOLOv13 is better equipped for feature learning specific to small targets and can efficiently handle interference issues such as degraded visibility, scale variations, and complex backgrounds encountered in aerial scenarios.

Despite its advantages, YOLOv13 exhibits certain limitations. For instance, small objects in UAV imagery—such as vehicles, pedestrians, and cyclists—often occupy only a few pixels in the image. When compounded by haze-induced feature degradation, inter-object occlusion, and background clutter, the model struggles to maintain the same efficiency as it does when detecting large targets, frequently resulting in missed detections and false positives. Moreover, the multi-scale nature of objects in aerial imagery—ranging from

tiny pedestrians to large vehicles—further challenges the model’s feature representation capacity. Enhancing YOLOv13’s ability to handle these interference issues inherent to UAV scenarios, particularly small target detection under degraded visibility, will necessarily involve architectural innovations that improve feature extraction and fusion while carefully balancing computational efficiency.

To address the aforementioned challenges—specifically small object representation, haze-induced feature degradation, multi-scale variation, and background interference in UAV imagery—this paper introduces an enhanced model, EFSL-YOLO (Enhanced Feature Selection and Learning YOLO), the architecture of which is depicted in Figure 2. First, an enhanced feature fusion attention network (EFFA-Net) is integrated as a preprocessing stage to restore features degraded by haze and uneven illumination. By introducing cross-layer feature fusion pathways and an enhanced attention module combining channel and spatial mechanisms with a lightweight gating unit, EFFA-Net significantly improves feature representation for small targets while maintaining network lightweight characteristics. This effectively addresses the challenge of degraded visibility in hazy conditions, providing the subsequent backbone with clearer, high-SNR features.

Next, the SwiGLUConv module is proposed to replace standard convolutional layers in the backbone network. The construction of this module includes two core designs: first, the integration of parallel depthwise convolutions with kernel sizes of 3×3 and 5×5 to adaptively expand the receptive field for multi-scale feature extraction; and second, the incorporation of a Swish-Gated Linear Unit (SwiGLU) mechanism that modulates information flow based on feature importance. The advantages of this design are twofold: the parallel multi-scale convolutions enhance the model’s capability to capture objects of varying scales under degraded visibility, while the gating mechanism improves the network’s adaptability to sparse and weak features typical in UAV imagery, thereby comprehensively strengthening the robust detection of small objects in hazy aerial scenes.

Finally, the Local Feature Multi-Scale Contextual Fusion (LF-MSCF) module is designed to improve the feature fusion stage between the neck and head of YOLOv13. This module integrates four complementary mechanisms: channel attention for inter-channel dependency recalibration, adaptive multi-scale selection with dilated depthwise convolutions (dilation rates 3, 5, 7) for receptive-field expansion, spatial attention for emphasizing salient target regions, and multi-head self-attention with a feed-forward network for global context modeling. This multi-component design enables LF-MSCF to capture both fine-grained local details and global contextual cues concurrently, significantly enhancing detection robustness for small, dense, or occluded targets in cluttered UAV scenarios while effectively suppressing background interference. Figure 2A illustrates the overall architecture of the EFSL-YOLO model, while Figure 2B–D detail the structures of EFFA-Net, SwiGLUConv, and LF-MSCF modules, respectively.

2.3.2. EFFA-Net

Severe weather, such as smoke, can significantly degrade the quality of drone imagery. Therefore, image dehazing preprocessing has become a critical step for achieving accurate surveillance. In [21], a Feature Fusion Attention Network (FFA-Net) is proposed, and it is recognized as an efficient and specialized solution for image dehazing. By effectively integrating feature fusion with a dual-attention mechanism, it significantly enhances both the visual quality and detail preservation of dehazed images. However, the original FFA-Net struggles with weak texture response and inadequate scale adaptability in group-level feature interaction, hindering its performance on drone-captured small objects. To overcome this issue, this paper proposes an enhanced feature fusion attention network (EFFA-Net), which makes improvements in both cross-layer architecture and attention

enhancement mechanisms, making it better suited to the sparse feature distributions and scale variations typical of small objects in drone imagery.

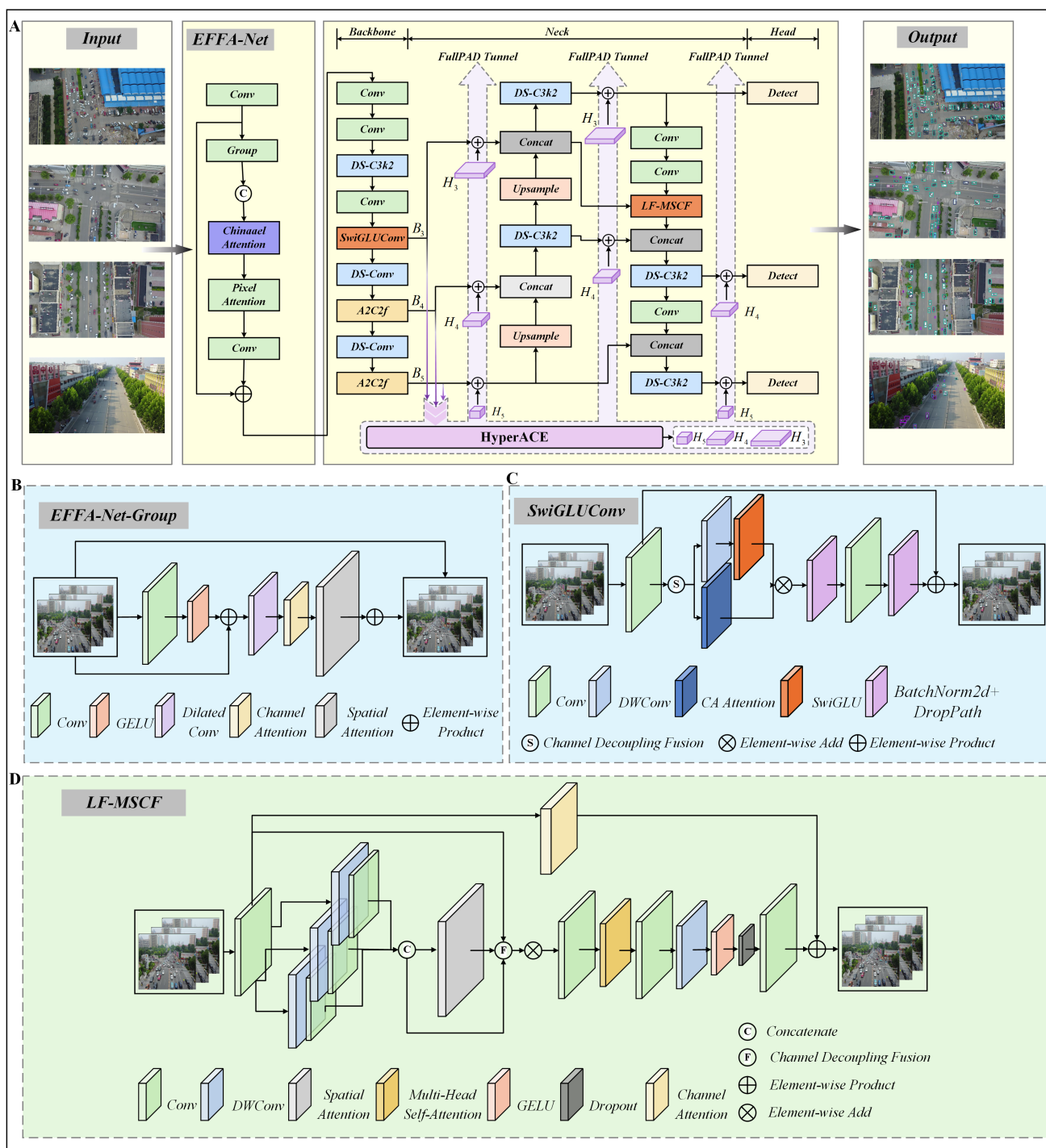


Figure 2. Architecture of the proposed EFSL-YOLO. (A) Overall architecture. (B) Architecture of EFAA-Net-Group. (C) Architecture of SwiGLUConv. (D) Architecture of LF-MSCF.

To ensure effective propagation from shallow texture details to deep semantic representations, the proposed EFAA-Net preserves FFA-Net's core architecture, feature grouping and inter-group recursive fusion. First, to further enhance multi-scale representation of small targets, the model introduces cross-layer feature fusion pathways that systematically integrate shallow textures, critical mid-level features, and deep semantics. Then, based on this structure, an enhanced attention module is developed by combining channel and spatial attention mechanisms to effectively amplify sparse and weak features typical in

UAV imagery, which addresses FFA-Net's weak texture response. Finally, the EFFA-Net incorporates a dual-branch multi-scale aggregation framework that adaptively combines high-resolution detail with rich low-resolution semantics, producing more discriminative and stable representations for small object detection. Through these enhancements, the proposed EFFA-Net achieves significant improvements in feature representation for small target detection in drones while maintaining network lightness, demonstrating higher detection robustness in scenarios with sparse features and significant scale variations.

EFFA-Net begins by feeding raw drone-captured images into a convolutional module to extract initial feature representations. These features are then divided into multiple sub-feature groups to capture local information at different scales, which are sequentially fed into a group structure composed of stacked Blocks. Each Block first extracts local features through standard convolutional layers, followed by nonlinear transformations via ReLU activation functions. The convolved features undergo dilated convolution to expand the receptive field, enabling broader contextual information capture without increasing parameter count. Features from multiple groups are concatenated along the channel dimension, with global average pooling and max pooling extracting global information. A shared multi-layer perceptron generates adaptive weights to achieve cross-level feature fusion. The concatenated features undergo weighted processing through channel attention and spatial attention modules, then refined via convolutional layers and residual connections to form enhanced features. This cross-level feature fusion path integrates shallow, key mid-level, and deep features, ultimately generating fog-free images through adaptive fusion. This approach achieves high-fidelity restoration of smog-affected areas while maintaining fine-grained local features and global context, demonstrating stronger feature representation capability and robustness in complex scene image dehazing and small object detection tasks.

In summary, the differences between EFFA-Net and FFA-Net lie in the integration of cross-layer fusion pathways that bypass FFA-Net's sequential bottleneck to ensure small object details are propagated to the reconstruction stage, the adoption of dilated convolution that expands the receptive field without additional parameters to balance contextual capture for smoke and high-resolution details for small targets, and the introduction of dual-branch aggregation that decouples detail preservation and semantic enrichment to outperform FFA-Net's single-stream processing in scale adaptability, with all these modifications collectively transforming FFA-Net into a specialized lightweight architecture for drone surveillance.

In this paper, the proposed EFFA-Net module is integrated as a preprocessing stage to restore features degraded by haze and uneven illumination in UAV images. Enhancing the input quality at the source can provide the subsequent backbone network with clearer, high-SNR features, ensuring stable detection performance even under complex atmospheric and lighting conditions.

2.3.3. SwiGLUConv

In UAV aerial photography, small targets occupy a minimal proportion in the image. After multiple downsampling processes, their features are often compressed to just a few pixels or even completely vanish, resulting in a significant drop in detection accuracy. Meanwhile, traditional convolutional layers tend to lose discriminative information during propagation. To address these challenges, we integrate the SwiGLUConv module into the backbone network. Building upon the gated convolution (GConv) design [22], which combines depthwise separable convolution with a dynamic gating mechanism, SwiGLUConv further incorporates parallel multi-scale convolutions and a Swish-Gated Linear Unit (SwiGLU) to enhance nonlinear expressiveness and adaptability for small-object detection.

Specifically, standard convolutional layers are replaced by a SwiGLU mechanism coupled with parallel depthwise separable convolutions. The architectural parameters were meticulously optimized through ablation studies: We employ parallel depthwise convolutions with kernel sizes of 3×3 and 5×5 . Comparative experiments revealed that a single 3×3 kernel yields an insufficient receptive field for capturing long-range feature correlations, while a single 5×5 kernel incurs high computational costs without proportional performance gains. The proposed dual-branch configuration optimally balances fine-grained local detail extraction with broader contextual awareness. The intermediate channel dimension is expanded by a factor of 2 to facilitate the SwiGLU gating mechanism ($x \cdot \text{SiLU}(v)$). This ratio was determined to be the minimal requirement for effective gating; a ratio of 1 restricts the gating branch's representational capacity, whereas ratios ≥ 3 introduce redundant computations with negligible accuracy improvements.

The mathematical formulation of the SwiGLU operation is given as

$$\begin{aligned} \text{SwiGLU}(x, W, V) &= \text{Swish}_{\beta}(xW) \otimes xV, \\ \text{Swish}_{\beta}(x) &= x\sigma(\beta x). \end{aligned} \quad (1)$$

where x is the input feature tensor, W and V are learnable weight matrices, and $\text{Swish}_{\beta}(\cdot)$ denotes the Swish function with parameter β . Nonlinear transformation is achieved through the sigmoid function $\sigma(\cdot)$; $\otimes$ represents element-wise multiplication.

Through this nonlinear transformation, the module enhances channel-wise activation dynamics, allowing it to dynamically filter information flow based on feature importance. Following feature modulation, BatchNorm2d layers and residual connections are employed to stabilize gradient propagation and preserve essential structural information during deep network training.

Distinct from the original GConv version, the improved SwiGLUConv module uniquely incorporates these parallel depthwise convolutions alongside the Swish-Gated Linear Unit to enhance interactive feature expression. Furthermore, the explicit addition of BatchNorm2d stabilizes the training process, enabling the model to capture fine-grained local details more accurately. By leveraging this optimized structure, the module not only expands the receptive field but also effectively mitigates information loss during downsampling, thereby strengthening the model's multi-scale perception and response capability for small targets in the YOLOv13 backbone.

This integration improves contextual aggregation and strengthens the discrimination of dense or overlapping small targets in UAV imagery. By leveraging the gating mechanism and the optimized multi-scale parallel convolution structure, SwiGLUConv not only expands the receptive field but also effectively mitigates information loss during downsampling, enabling precise and efficient detection without significant computational overhead.

2.3.4. LF-MSCF

Detailed features of small targets are prone to dilution during transmission, especially in dense occlusion scenarios where features of occluded targets tend to be missing or fragmented, leading to a significant decline in detection accuracy. Traditional fusion mechanisms often lack sufficient capability for detail enhancement and occlusion restoration in complex UAV scenarios. To address missed and false detections, we propose the LF-MSCF module, integrated between the neck and head of YOLOv13. Building on the multi-scale contextual fusion (MSCF) framework [22], LF-MSCF refines the internal mechanisms by incorporating a feedforward network and optimizing parameter configurations for small-target sensitivity.

Specifically, LF-MSCF consists of four synergistic components: channel attention for inter-channel dependency recalibration, adaptive multi-scale selection for receptive-field

expansion, spatial attention for emphasizing salient regions, and multi-head self-attention (MHSA) with a feed-forward network for global context modeling. A residual shortcut is employed throughout to stabilize gradient propagation. Key parameter selections are justified as follows: The adaptive multi-scale selection block utilizes three parallel 3×3 depthwise convolutions with dilation rates set to $d = [3, 5, 7]$. This specific sequence was chosen to ensure dense, non-overlapping coverage of the receptive field, enabling the efficient aggregation of contextual information from varying distances without spatial downsampling. To mitigate the computational cost associated with large dilation rates, a bottleneck structure with a channel reduction ratio of 4 is implemented. Ablation studies demonstrated that this ratio significantly reduces FLOPs while preserving critical semantic information, outperforming both narrower (ratio = 2) and wider (ratio = 8) configurations.

In the processing flow, channel attention first learns global inter-channel relationships via global average pooling and a lightweight bottleneck transformation. Next, the multi-scale block performs feature fusion using learned scale weights, enabling dynamic focus on optimal spatial ranges. Subsequently, spatial attention leverages average and max pooling to generate a 2D attention map, highlighting salient target regions while suppressing background noise. Following local optimization, an MHSA mechanism captures long-distance positional dependencies, and a feed-forward fusion network employing GELU activation is added to further refine discriminative representations.

The improved LF-MSCF model differs from the original one in that a feed-forward fusion network is added after the multi-head attention mechanism, and the internal mechanism is further optimized. Pre-convolution is introduced to avoid information loss caused by dimension mismatch. DWConv is used to perform local spatial refinement on the global features output by MHSA and strengthen the spatial position information of small targets. As a smooth nonlinear activation function, GELU can retain the weak feature signals of small targets and enhance the nonlinear representation ability of features. Dropout is adopted to stabilize training. These improvements greatly boost the discriminability of small targets and reduce missed detections and false detections.

This comprehensive design allows LF-MSCF to concurrently acquire fine-grained details and global contextual cues. By combining local convolution enhancement with optimized multi-scale fusion, the module effectively suppresses background redundancy and strengthens the network's ability to capture the saliency of small targets. These improvements significantly boost detection robustness in cluttered UAV scenarios, reducing both missed and false detections.

2.4. Experimental Settings

2.4.1. Experimental Environment and Training Settings

The model training and experiments for the proposed EFSL-YOLO were conducted on a high-performance computing server. The hardware configuration included an Intel(R) Xeon(R) Silver 4210R CPU operating at 2.40 GHz, 256 GB of RAM, and an NVIDIA GeForce RTX 3090 GPU as the primary computational unit. The system ran on the Ubuntu 20.04 LTS (64-bit) operating system with 4 TB of storage capacity. All software environments were built upon Python 3.11.13, with PyTorch 2.3.1 with CUDA 12.1. as the deep learning framework and CUDA 11.7, along with CuDNN 8.9.2, to fully leverage the parallel computing capabilities of the GPU. The detailed hardware and software configurations are summarized in Table 2.

For consistent model input, all images from the three benchmark datasets were uniformly resized to a resolution of 640×640 pixels. The training configurations were adapted to accommodate the distinct characteristics of each dataset. For the VisDrone dataset, which contains a diverse set of ten categories with significant class imbalance, the model

was trained for 300 epochs. For the UAVDT and HazyDet datasets, which focus on vehicle detection in complex scenarios and hazy conditions respectively, the training was set to 270 epochs. All datasets were trained with a batch size of 16 to balance memory consumption and training stability.

The optimization process employed the AdamW optimizer with an initial learning rate of 0.01, momentum of 0.937, and weight decay of 0.005 to enhance the model's generalization ability and mitigate overfitting. For the VisDrone dataset, a step-wise learning rate decay strategy was applied at epochs 200 and 270, while for UAVDT and HazyDet, decay was applied at epochs 180 and 240. The IoU loss function was adopted to guide the bounding box regression optimization process. The main hyperparameter configurations used for training across all datasets are detailed in Table 3.

Table 2. Computational implementation.

Device	Parameter Configuration
Central processor unit (CPU)	Intel(R) Xeon(R) Silver 4210RCPU @2.40 GHz
Graphics processor unit (GPU)	NVIDIA GeForce RTX 3090
Running memory (RAM)	256 GB
Operating system	Ubuntu 20.04 LTS (64-bit)
Computer language	Python 3.11.13
Acceleration library	CuDNN 8.9.2

Table 3. Algorithm hyperparameters.

Parameters	Setup
Epochs	300/270
Batch size	16
Learning rate	0.001
Image size	640 × 640
Weight Decay	0.005
Momentum	0.937
Optimizer	AdamW

2.4.2. Evaluation Metrics

Precision and recall are key performance metrics for detection models. Precision reflects the model's ability to avoid false alarms, defined as the fraction of correctly identified positives out of all predicted positives. Recall characterizes the model's sensitivity in detecting true targets, defined as the fraction of true positives successfully retrieved from all actual positives. Their formal definitions are given as

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

where TP denotes the detector correctly identifying annotated objects, FP denotes the detector mistakenly classifying background regions as annotated objects, and FN denotes the detector mistakenly classifying annotated objects as background regions.

Average precision (AP): refers to the average accuracy of various categories in the dataset under different confidence thresholds, which is expressed as

$$AP = \int_0^1 P(R)dt. \quad (4)$$

where P denotes precision, and R denotes recall.

Mean average precision (mAP): this refers to the mean average precision of all categories in the dataset. It has the form of

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_n \quad (5)$$

where N represents the number of categories, and AP_n represents the AP of class n .

3. Experimental Results and Analysis

3.1. Ablation Experiment

To verify the effectiveness of the proposed improvements, the impact of different modules on the performance of the improved YOLOv13 is evaluated through ablation experiments. The results of the ablation experiments are shown in Table 4. The following three components are examined:

1. UAV aerial images are often affected by haze, illumination imbalance, and atmospheric degradation, which reduce the clarity of input features and hinder small object detection. The lightweight EFFA-Net is introduced at the input stage to restore low-quality images and improve the signal-to-noise ratio of the extracted features. Table 4 shows that on the HazyDet dataset, adding EFFA-Net increases $mAP@0.5:0.95$ from 52.8% to 53.7% and $mAP@0.5$ from 74.1% to 74.5%. These results confirm that improving image quality at the source effectively enhances the stability and robustness of downstream detection under complex atmospheric conditions.
2. Small UAV targets usually contain very limited feature cues, and conventional convolution layers tend to lose discriminative information during propagation. The SwiGLUConv module enhances multi-scale feature interaction through gated activation and lightweight dynamic convolution, enabling more selective enhancement of small object features. When SwiGLUConv is incorporated into the backbone, the $mAP@0.5:0.95$ (%) and $mAP@0.5$ (%) metrics on the Hazydet dataset improved by 2.5% and 0.8% respectively. On the VisDrone dataset, precision increased by 1.9%, recall increased by 2.0%, and $mAP@0.5$ (%) increased by 2.3%, proving that the gated convolution structure significantly improves the model's responsiveness to small targets while maintaining lightweight efficiency.
3. Dense UAV scenes often suffer from feature aliasing, scale inconsistency, and occlusion interference, which weaken the effectiveness of high-level semantic fusion. The LF-MSCF module combines multi-head self-attention with a feedforward network to jointly capture long-range dependencies and local spatial relationships. Benefiting from its multi-scale feature fusion strategy and adaptive channel weighting, the $mAP@0.5:0.95$ (%) and $mAP@0.5$ (%) metrics on the HazyDet dataset improved by 3.0% and 1.8% respectively. On the VisDrone dataset, precision increased by 3.8%, recall increased by 1.7%, and $mAP@0.5$ (%) increased by 2.9%. On the UAVDT dataset, precision increased by 3.0%, recall increased by 8.2%, $mAP@0.5:0.95$ (%) increased by 2.4%, and $mAP@0.5$ (%) increased by 3.5%. This demonstrates that LF-MSCF effectively mitigates feature aliasing in dense backgrounds and contributes substantially to accuracy improvement.

It can be seen from the ablation experiments that adding the LF-MSCF module alone brings a greater accuracy improvement than adding the SwiGLUConv module alone, and the LF-MSCF module contributes more to mAP improvement.

The results of the ablation experiments are presented in the table, and the visualizations of the ablation experiments are shown in Figure 3. Among the visualization results,

yellow bounding boxes indicate missed detections, while red bounding boxes denote false detections. It can be observed from the provided visualizations that the false detection and missed detection rates of the improved model are significantly reduced, thereby verifying the effectiveness of the proposed improvements.

Table 4. Ablation experiment results on different datasets.

Dataset	YOLOv13s	EFFA-Net	SwiGLU Conv	LF- MSCF	P (%)	R (%)	mAP @0.5:0.95 (%)	mAP @0.5 (%)
HazyDet	✓				84.1	69.5	55.2	76.8
	✓	✓			84.4	70.8	55.3	77.4
	✓	✓	✓		85.9	71.5	57.8	78.2
	✓	✓	✓	✓	86.8	71.6	58.2	78.6
Visdrone	✓	-			47.1	35.0	20.9	35.5
	✓	-	✓		49.0	37.0	22.4	37.8
	✓	-		✓	48.6	37.6	22.7	38.0
	✓	-	✓	✓	52.8	38.7	24.3	40.7
UAVDT	✓	-			38.9	31.5	19.1	30.5
	✓	-	✓		38.0	32.9	17.9	31.1
	✓	-		✓	38.9	35.2	19.1	31.5
	✓	-	✓	✓	41.9	39.7	21.5	34.0

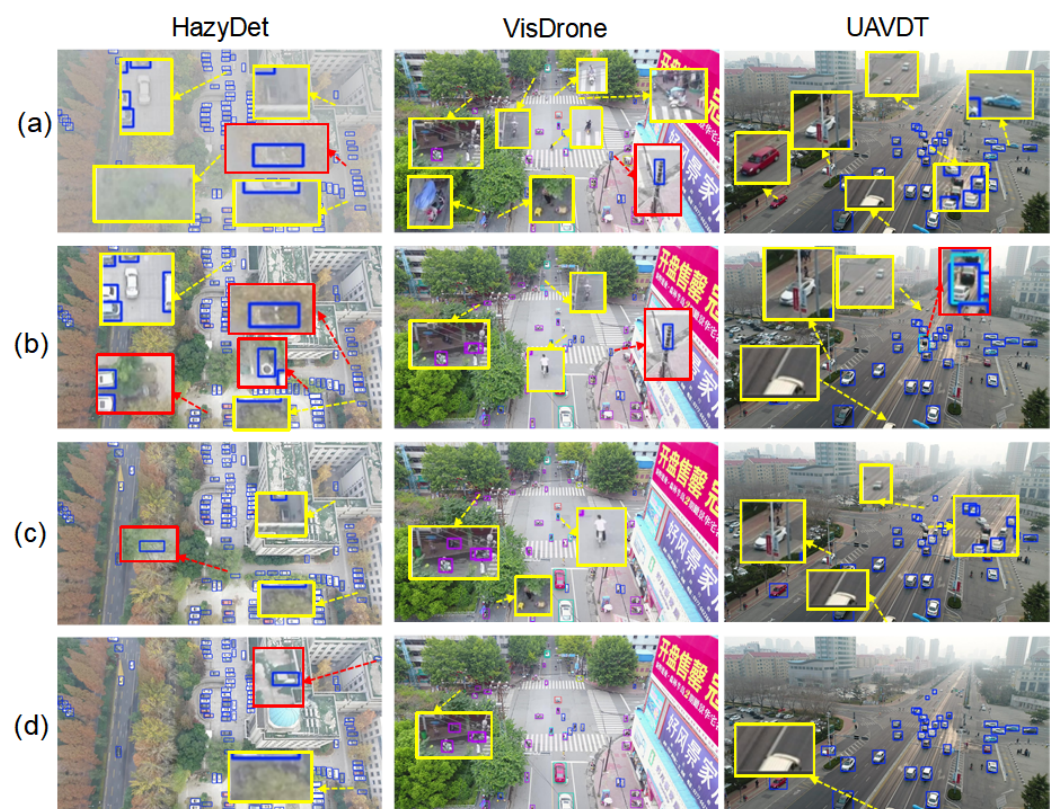


Figure 3. Visualization of ablation experimental results on three datasets. (a) Baseline model YOLOv13 on HazyDet, VisDrone, and UAVDT datasets; (b) YOLOv13 integrated with EFFA-net/SwiGLUConv; (c) YOLOv13 integrated with EFFA-net/SwiGLUConv and LF-MSCF; (d) The final proposed EFSL-YOLO model. Red boxes denote false detections, and yellow boxes denote missed detections.

3.2. Comparison Experiments

3.2.1. Comparison of Backbone Network

Comparative experiments were conducted with the YOLOv13 backbone network and four lightweight backbone networks, including EfficientViT [23], FasterNet [24],

RepViT [25], and ConvNeXtV2 [26]. These comparative models represent typical design paradigms of current lightweight vision models: EfficientViT balances speed and accuracy on resource-constrained edge devices (e.g., drones) by optimizing the Transformer architecture and adopting efficient computational modules; FasterNet employs PConv and MLP hierarchical blocks to enhance hardware parallelism and reduce computational overhead, meeting the real-time surveillance requirements of drones; RepViT improves feature representation capability without increasing latency by separating channel mixers via structural re-parameterization, which is suitable for deployment on mobile devices and drone terminals; ConvNeXtV2 integrates GRN normalization and depthwise separable convolutions to strengthen inter-channel interactions while simplifying convolution operations, a classic pure convolutional optimization scheme. All experiments were performed under unified conditions (e.g., identical input resolution, training epochs, and loss function), and the quantitative results are summarized in Table 5.

As shown in Table 5, compared with the baseline YOLOv13, the proposed model achieves significant improvements in mAP@0.5 and mAP@0.5:0.95 metrics (specific values can be supplemented according to experimental results), with a slight increase in recall rate, indicating that the model exhibits stronger perception capability for sparse small targets in drone scenarios. Compared with other mainstream backbone networks, FasterNet delivers relatively lower detection accuracy due to its excessive pursuit of computational reduction, leading to insufficient feature extraction for small objects. EfficientViT and RepViT achieve moderate performance with balanced efficiency and precision, yet still lag behind our method in terms of overall detection quality. ConvNeXtV2 demonstrates competitive feature representation ability but suffers from higher computational complexity compared to the proposed framework. In contrast, our EFSL-YOLO model achieves the optimal trade-off between detection accuracy and model complexity, outperforming all comparative lightweight backbones in most key metrics while maintaining practical parameters and computational cost for drone-based object detection.

Table 5. Results for comparison of different backbone networks on VisDrone Datasets.

Methods	P (%)	R (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Params (M)	GFLOPs (G)
YOLOv13	47.1	35.0	35.5	21.1	9.0	20.7
EfficientViT	46.2	32.1	32.8	18.5	7.2	15.3
FasterNet	45.8	31.5	32.1	17.9	6.8	12.6
RepViT	46.5	32.9	33.5	19.2	7.9	18.1
ConvNeXtV2	46.0	32.4	33.0	18.8	8.5	22.4
EFSL-YOLO	51.9	39.7	40.7	24.3	8.9	28.9

3.2.2. Comparison of Different Detection Models

To evaluate the proposed model's advantages and generalization performance, comparative experiments are conducted with YOLO series models and real-time end-to-end target detection models based on Transformer and RCNN across the three datasets. For a comprehensive comparison, this paper selects a set of leading models: the end-to-end methods RT-DETR-R18 [27], TinyDef-DETR [28], DETR-R18 [29], DCM-DETR [29], Freq-DETR [30], BAP-DETR-S [31] and one-stage Drone-YOLOs [32], Drone-YOLOm [32], YOLO-RCS-FasterNet [33], OSD-YOLOv10 [34], YOLOv11(O+E) [35], and YOLO-GML [36]. As shown in Table 6, the experimental results demonstrate that the proposed method achieves optimal target detection performance while maintaining a lightweight architecture, demonstrating significant competitive advantages over these models.

To verify the superiority of the proposed EFSL-YOLO model, this study examines mainstream object detection methods and conducts a performance comparison with EFSL-

YOLO on three benchmark datasets: HazyDet, VisDrone, and UAVDT. Five core indicators are mainly evaluated, including precision, recall, mAP@0.5:0.95, mAP@0.5, and model parameters, with the comparison results shown in Table 6.

Table 6. Comparison results of different networks on HazyDet, VisDrone, and UAVDT datasets.

Dataset	Model	mAP@0.5:0.95 (%)	mAP@0.5 (%)	Params (M)
HazyDet	YOLO-GML [36]	53.3	73.6	11.66
	UFDT-YOLO [36]	/	75.92	11.80
	DETR-R18 [29]	53.1	71.4	19.8
	DCM-DETR [29]	54.6	74.3	22.1
	EFSL-YOLO	58.2	78.6	8.9
VisDrone	Drone-YOLOs [32]	20.4	35.6	10.9
	Drone-YOLOm [32]	22.5	38.9	33.9
	YOLO-RCS-FasterNet [33]	21.9	37.6	11.0
	RT-DETR-R18 [27]	18.5	27.8	20.0
	TinyDef-DETR [28]	21.7	37.2	20.5
	EFSL-YOLO	24.3	40.7	8.93
UAVDT	ST-YOLO [34]	/	33.4	9.0
	YOLOv11(O+E) [35]	17.7	26.5	/
	FBRT-YOLO-L [37]	18.4	31.1	14.6
	RT-DETR [27]	18.6	32.4	42.0
	Freq-DETR [30]	18.6	32.4	/
	BAP-DETR-S [31]	19.9	31.4	25.0
	EFSL-YOLO	21.5	34.0	8.9

On the HazyDet foggy dataset, EFSL-YOLO exhibits the optimal comprehensive performance. Its precision and recall are higher than those of the second-best DCM-DETR, and mAP@0.5 is improved compared with DCM-DETR. Meanwhile, the number of parameters is much lower than that of DCM-DETR and DETR-R18, even lower than the lightweight YOLOv8s, achieving the dual advantages of accuracy and light weight. Compared with YOLO-GML and YOLOv8s, EFSL-YOLO achieves significant improvements in all accuracy indicators, further proving its adaptability in complex foggy scenarios. On the VisDrone UAV aerial dataset, the mAP@0.5:0.95 of EFSL-YOLO is higher than all comparative models, including higher than Drone-YOLOm and YOLO-RCS-FasterNet. The mAP@0.5 also ranks first, and the number of parameters is much lower than that of Drone-YOLOm, RT-DETR-R18, and TinyDef-DETR, solving the deployment problem of limited memory on UAV embedded platforms. On the UAVDT small target dataset, the mAP@0.5:0.95 of EFSL-YOLO is significantly higher than that of comparative models such as BAP-DETR-S and RT-DETR. The mAP@0.5 is also better than all comparative algorithms, and its parameter count is lower than that of FBRT-YOLO-L and BAP-DETR-S, showing strong robustness and lightweight advantages in small target detection. In summary, EFSL-YOLO achieves the best performance in accuracy indicators on three datasets of different scenarios and effectively controls the number of model parameters. Compared with existing mainstream object detection methods, it has better comprehensive performance and deployment potential.

To verify the object detection performance of the proposed EFSL-YOLO method in complex scenarios such as foggy weather and UAV aerial photography, comprehensive comparative experiments are conducted between the EFSL-YOLO model and mainstream object detection algorithms (RT-DETR, YOLOv8, YOLOv10, YOLO11, YOLOv12) on three benchmark datasets: HazyDet, VisDrone, and UAVDT. The evaluation covers multiple dimensions, including precision (P), recall (R), mAP metrics, model parameters (Params),

computational complexity (GFLOPs), and inference speed (FPS), with results presented in Table 7.

Table 7. Comparison of detection performance of different detection models.

Dataset	Methods	P (%)	R (%)	mAP@0.5:0.95 (%)	mAP@0.5 (%)	Params (M)	GFLOPs (G)	FPS
HazyDet	RT-DETR	84.2	67.5	54.1	75.3	20.1	58.3	37.62
	YOLOv8	86.0	69.3	55.7	76.6	11.14	28.7	123.46
	YOLOv10	85.4	68.1	55.2	75.9	15.3	24.8	92.31
	YOLO11	84.8	69.7	55.9	76.6	9.4	21.3	96.15
	YOLOv12	86.0	68.9	55.3	76.2	15.3	19.3	70.42
	EFSL-YOLO	86.8	71.6	58.2	78.6	8.9	28.9	43.29
VisDrone	RT-DETR	50.1	36.2	21.8	38.8	20.1	58.3	34.00
	YOLOv8	45.1	35.0	20.4	34.8	11.14	28.7	116.28
	YOLOv10	46.2	34.3	20.6	35.1	15.3	24.8	68.26
	YOLO11	47.0	33.7	20.3	34.7	9.4	21.3	94.34
	YOLOv12	46.8	34.1	20.5	35.0	15.3	19.3	70.42
	EFSL-YOLO	52.8	38.7	24.3	40.7	8.9	28.9	43.86
UAVDT	RT-DETR	48.6	35.4	20.1	33.6	20.1	58.3	36.10
	YOLOv8	37.0	32.0	17.7	30.0	11.14	28.7	81.97
	YOLOv10	37.0	27.3	17.6	28.7	15.3	24.8	62.35
	YOLO11	41.1	28.6	18.8	31.1	9.4	21.3	68.49
	YOLOv12	33.8	38.1	18.1	30.6	15.3	19.3	51.02
	EFSL-YOLO	41.9	39.7	21.5	34.0	8.9	28.9	46.08

On the HazyDet foggy dataset, the EFSL-YOLO model significantly outperforms all comparative algorithms in both precision and recall, showing strong adaptability to foggy environments and robust target recognition capabilities. Precision reaches 86.8%, and recall reaches 71.6%. For the comprehensive evaluation metric mAP@0.5:0.95, EFSL-YOLO achieves 58.2%, an improvement of more than 6 percentage points over the second-best YOLO11. The mAP@0.5 value reaches 78.6%, markedly surpassing the YOLO series models. Meanwhile, the parameters and GFLOPs of EFSL-YOLO are controlled within a reasonable range, at 8.9 M and 28.9 G respectively, and are superior to the RT-DETR model with 20.1M parameters. The inference speed is 43.29 FPS, maintaining high inference efficiency while ensuring high accuracy. On the VisDrone UAV aerial dataset with challenges of multi-scale objects and complex backgrounds, the EFSL-YOLO model still leads with a mAP@0.5:0.95 score of 24.3%, outperforming YOLOv10, YOLO11 and YOLOv8. The parameter count is only 8.9M, much lower than the 20.1M of RT-DETR, making it more suitable for deployment on embedded UAV platforms with limited memory. Although dense objects and complex backgrounds in UAV scenes reduce the inference speed to 43.86 FPS, this speed still meets the requirements of most real-time detection tasks and is more efficient than RT-DETR, verifying the model's deployment potential on resource-constrained devices. On the UAVDT small object dataset focusing on tiny and distant targets, the EFSL-YOLO model achieves 21.5% in mAP@0.5:0.95, significantly higher than RT-DETR and all YOLO series models, reflecting stronger perception and localization ability for small targets. The inference speed is 46.08 FPS, which not only outperforms the computation-intensive RT-DETR but also strikes a good balance between detection accuracy and speed within the speed range of YOLOv10 and YOLO11, demonstrating remarkable comprehensive advantages in all three complex scenarios.

The native inference speed of the YOLOv13 baseline model used in this paper is around 47 FPS across the three datasets, which is relatively low among the comparison methods. The inference speed of EFSL-YOLO on each dataset is only slightly lower than that of the baseline model, while detection accuracy is greatly improved. This result further proves that

the proposed improvement achieves an efficient trade-off between accuracy and efficiency, without excessive sacrifice of inference speed caused by performance enhancement.

Figure 4 presents the comparative results of confusion matrices for the YOLOv8, YOLOv9, YOLOv10, YOLO11, YOLOv12, and EFSL-YOLO models. The proposed EFSL-YOLO achieves the highest recall on challenging categories such as truck (0.50) and bus (0.79) among all compared methods. This demonstrates that our designed modules effectively enhance the feature discriminability for medium- and small-scale targets in complex UAV scenes while maintaining superior performance on large-scale targets (car) and background separation.

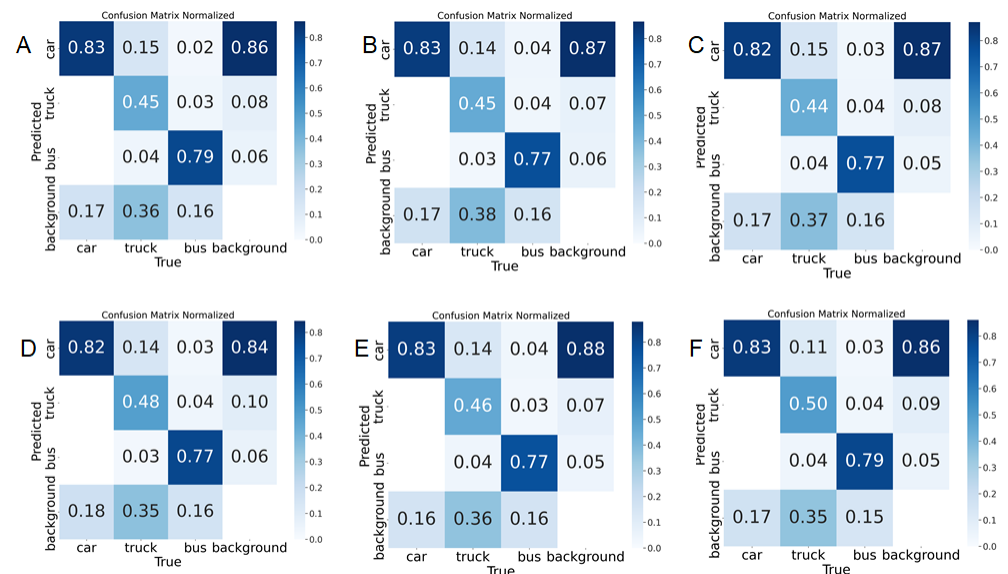


Figure 4. Comparison of confusion matrices of different models evaluated on HazyDet dataset. (A) RT-DETR. (B) YOLOv8. (C) YOLOv10. (D) YOLO11. (E) YOLOv12. (F) EFSL-YOLO.

The proposed EFSL-YOLO, as shown in Figure 5F, achieves the best performance among all compared methods. While maintaining accurate background recognition, this method enables balanced and efficient detection across all target categories: on one hand, the recall rates of core large-scale targets such as car and bus reach 0.68 and 0.47 respectively, both the highest values among all compared methods; on the other hand, it significantly improves the detection performance of medium-scale targets (e.g., van, truck, motor) and small-scale targets (e.g., pedestrian, tricycle, awning-tricycle), with the recall rates of all target categories ranking at the forefront, and no obvious detection shortcomings. Meanwhile, the misclassification ratio between different categories is more balanced without extreme misjudgment, which fully demonstrates the comprehensive detection advantages of the proposed method for targets of different scales and proportions in complex UAV scenarios.

Figure 6 illustrates the normalized confusion matrix comparison of various detection models on the UAVDT dataset. The proposed EFSL-YOLO model yields more stable and accurate detection performance than all comparative methods. Specifically, our model attains the optimal recall for most object categories while effectively alleviating cross-class misclassification. It notably improves the recognition capability for small and occluded targets in UAVDT scenarios. The overall confusion between different categories is the lowest among all models, which validates the strong feature discriminability and robust detection performance of the proposed EFSL-YOLO in complex UAV aerial image scenes.

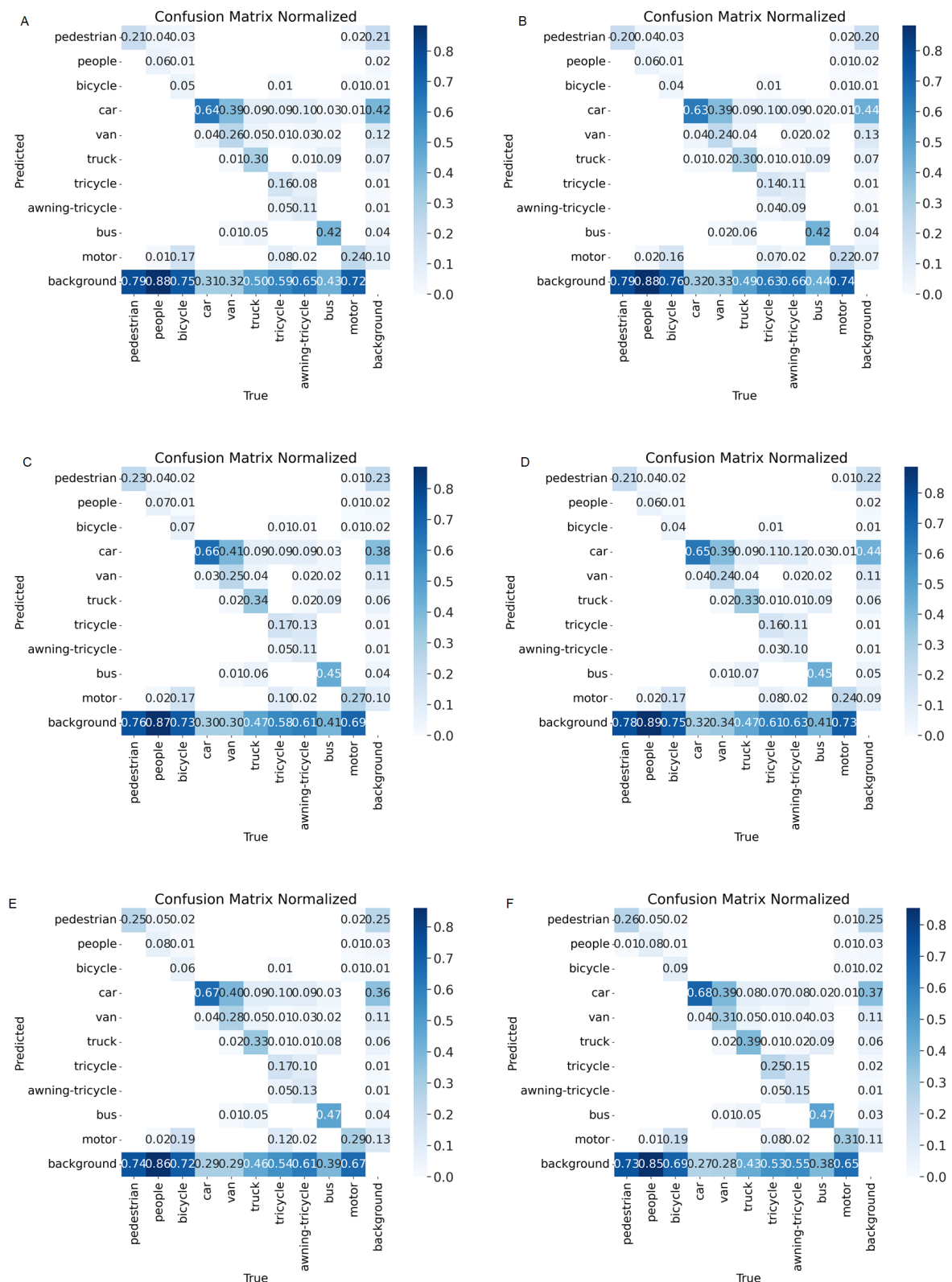


Figure 5. Comparison of confusion matrices of different models evaluated on VisDrone dataset. (A) RT-DETR. (B) YOLOv8. (C) YOLOv10. (D) YOLO11. (E) YOLOv12. (F) EFSL-YOLO.

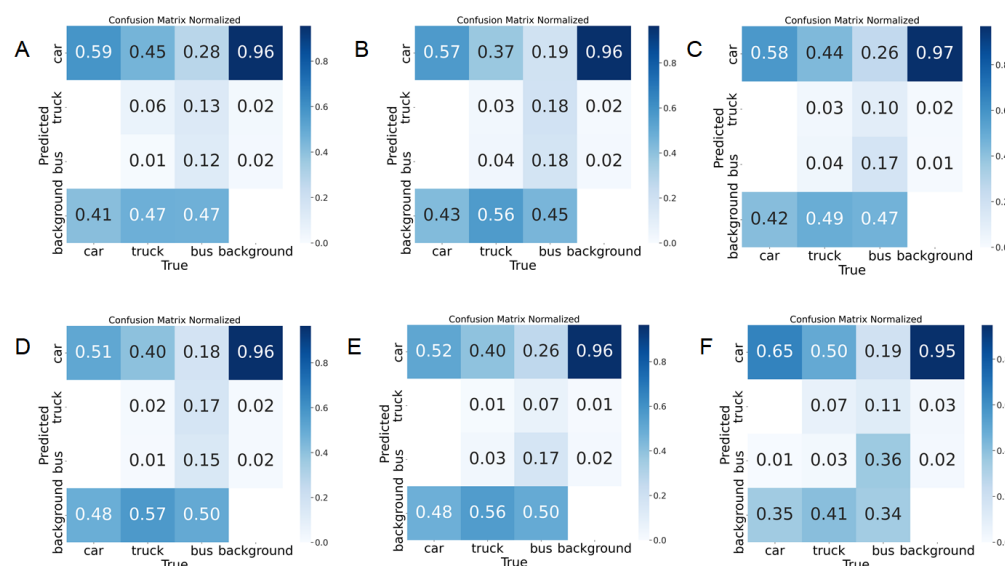


Figure 6. Comparison of confusion matrices of different models evaluated on UAVDT dataset. (A) RT-DETR. (B) YOLOv8. (C) YOLOv10. (D) YOLO11. (E) YOLOv12. (F) EFSL-YOLO.

3.3. Visualization

To verify the detection capability of the proposed model, several visual comparison experiments are conducted. In this study, EigenCAM, a feature map-based visualization technique, is adopted to generate heatmaps by performing principal component analysis and factorization on feature maps. This method can highlight object regions more accurately and demonstrate excellent object localization ability. The corresponding heatmaps are presented in Figure 7.

To quantitatively and qualitatively evaluate the target detection performance of different methods in complex scenarios, this paper conducts a visual comparative analysis of the generated heatmap results. Evaluated from three core dimensions—target recall, localization precision, and noise suppression—the proposed method in this paper demonstrates the optimal comprehensive performance. In two typical complex scenarios, urban streets and foggy roads, the proposed method can effectively cover more small-scale targets (e.g., pedestrians, non-motor vehicles, and distant vehicles) to achieve a higher detection recall rate while maintaining clear and sharp target edge contours, thus significantly avoiding target adhesion and false detection in dense areas. In contrast, other comparative methods exhibit obvious shortcomings: although RT-DETR and YOLOv10 perform well in local target clarity, their recall rates are significantly lower, resulting in the omission of a large number of potential targets; while YOLOv8 and YOLO11 improve the target coverage to a certain extent, they introduce varying degrees of noise diffusion and artifact interference, which are prone to false responses in non-target regions, especially in complex background areas. Overall, the proposed method achieves an optimal balance among target clarity, color contrast, and noise suppression, as well as the best trade-off between high recall and high precision, making it more suitable for practical engineering applications such as UAV inspection and traffic monitoring.

Comparison results across four typical scenarios are presented in Figure 8. Here, red markers indicate false positives, and yellow markers show missed detections (including occlusions). The first four rows display results from RT-DETR, YOLOv8, YOLOv10, and YOLO11, respectively, while the final row shows the proposed method. The images presented in the experiments are collected from real-world complex scenarios, in which objects generally exhibit the characteristics of small scale and dense distribution. Meanwhile, multiple interference factors such as tree occlusion, uneven illumination, and complex

background textures are widely existing, which easily lead to typical problems, including missed detection and false detection for detection models. It can be clearly observed from the visual comparison results that the proposed method still maintains stable and accurate detection performance under the above complex conditions. It has strong discrimination ability for occluded objects, small-scale objects and confusing background regions, showing an overall lower false detection rate and missed detection rate.

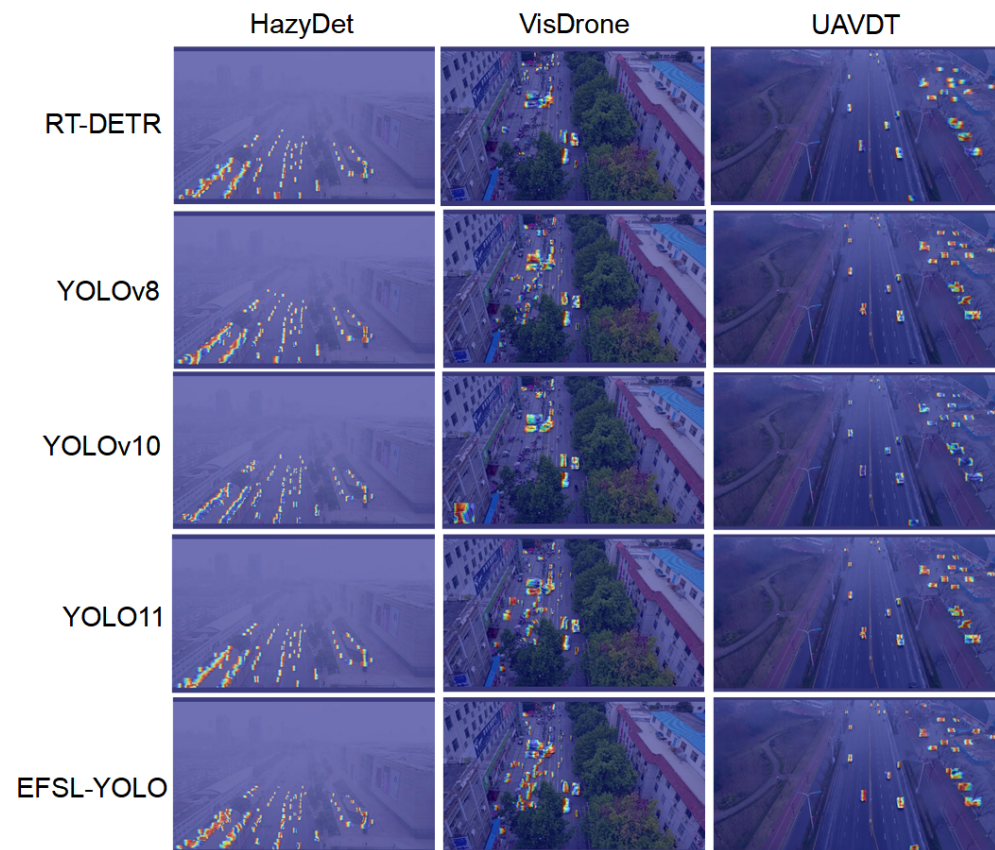


Figure 7. Comparison of heatmaps from different models. The colors in the heatmap represent different intensity levels, with warmer colors indicating higher values and cooler colors representing lower values.

In contrast, the other comparison methods expose obvious limitations on different datasets: In the HazyDet dataset, some models tend to misclassify building contours as vehicles and fail to effectively capture objects occluded by branches and leaves. Moreover, in the complex texture background formed by the interlacing of trees and ground, they cannot accurately distinguish object regions from interfering backgrounds, resulting in numerous false responses and missed detections. In the VisDrone aerial dataset, affected by the aerial altitude and long-distance imaging, the object scales are further reduced. Other methods are difficult to effectively perceive extremely small objects, such as pedestrians, and also present low recall for small vehicles. Vans occluded by trees suffer from obvious missed detection and category misjudgment. In the UAVDT dataset, due to the high similarity between vehicles, zebra crossings and road surfaces in color, the contrast between objects and background is low. Most methods suffer from significant missed detections, especially in edge regions such as the bottom-left corner of images, where multiple objects are not effectively identified.

Based on the visual results of multiple scenarios and datasets, it can be concluded that the proposed method achieves superior performance in complex background suppression, occluded object perception, and small object enhancement. It can effectively improve the

practicality and robustness of the model in real-world scenarios such as UAV inspection and traffic monitoring, which fully demonstrates the superiority of the proposed method.

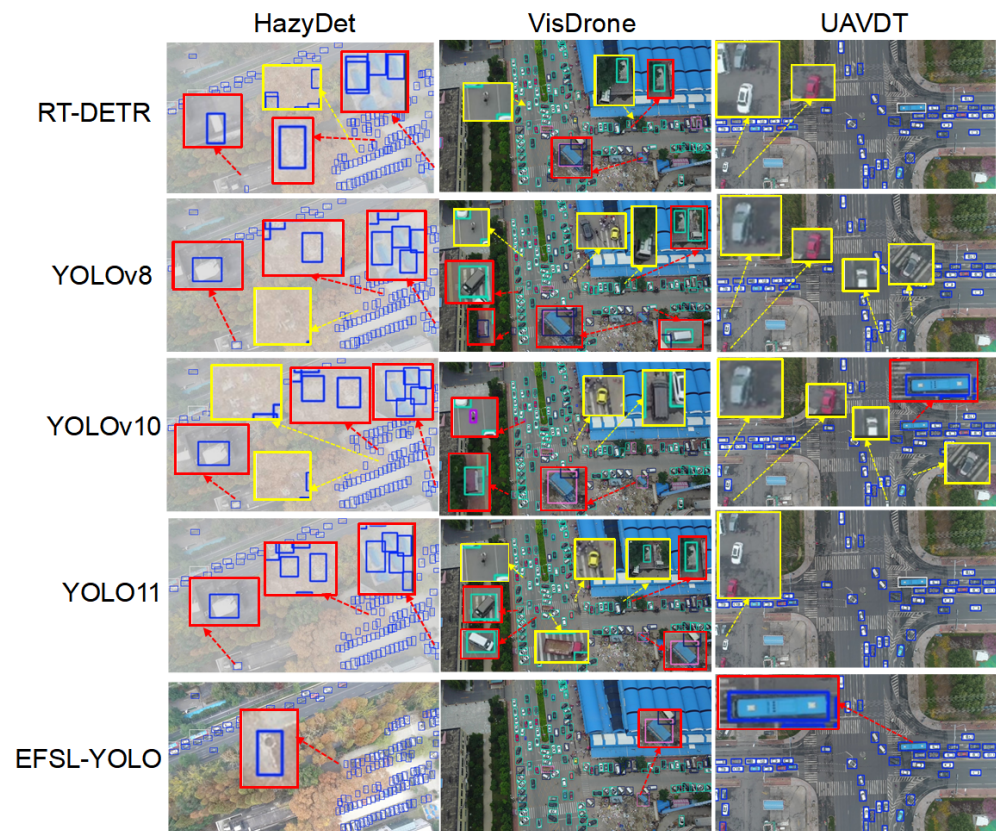


Figure 8. Visualization results of different methods. Red boxes denote false detections, and yellow boxes denote missed detections.

4. Discussion

4.1. Advantages

The detection accuracy obtained in this study is of critical importance to the field of aerial surveillance and intelligent monitoring, particularly for drone-based small object detection in complex environmental conditions. In unstructured aerial scenes, the detection challenge is twofold. First, under adverse weather conditions such as haze, smoke, and uneven illumination, image quality degradation severely compromises feature visibility, where target information becomes blurred and difficult to distinguish from background interference. Second, and perhaps more critical for reliable surveillance, is the detection of small objects in dense occlusion scenarios. Although distinct in structure, small targets in drone imagery often appear in crowded configurations—such as vehicles in traffic, pedestrians in crowds, or animals in herds—causing severe mutual occlusion that often confuses standard detectors into missing a substantial portion of targets. Previous studies have struggled to simultaneously address these conflicting challenges. For instance, Zhang et al. [32] proposed an enhanced YOLO model to address the specific challenges associated with drone image object detection, achieving an mAP50 of 38.9%. Similarly, Shen et al. [28] achieved a recognition accuracy of 37.2% using a DETR-based detector but introduced additional complexity to the training pipeline. In stark contrast, our proposed YOLOv13-based algorithm significantly outperforms these benchmarks with an mAP50 of 40.7%. This performance margin validates that our architecture excels in both scenarios: the EFA-Net preprocessing module effectively restores feature fidelity in degraded imaging conditions through cross-layer feature fusion and enhanced attention mechanisms, while

the LF-MSCF module sharpens the contextual awareness and boundary discrimination for occluded small targets within dense clusters. Beyond technical metrics, the precision of our proposed method holds potential for the UAV industry. Currently, drone-based visual inspection is used in applications such as disaster response and infrastructure monitoring, which are increasingly constrained by low recall rates under challenging conditions, high false alarm rates, and the risk of missing critical targets. Our algorithm serves as a robust “visual intelligence core” for autonomous UAV systems, enabling continuous, reliable operation in complex environments without human intervention. The practical benefits of this advancement are substantial. The achieved 40.7% mAP50 directly minimizes the rate of missed detections and false positives. While traditional methods are subjective and prone to error under fatigue or challenging visual conditions, our model ensures consistent and accurate target localization. This capability not only enhances the effectiveness of drone-based inspections but also facilitates autonomous decision making in time-critical scenarios such as search and rescue operations, thereby translating technical accuracy into tangible operational efficiency.

4.2. Limitations and Future Work

Notwithstanding the superior efficacy of our proposed model across primary benchmarks, including precision, recall, F1 score, mAP50, and mAP50–95, its integration into practical UAV deployment scenarios is still governed by an inherent trade-off between inference latency and the stringent resource constraints of edge computing devices commonly used in drone platforms. If the model is directly deployed onto resource-constrained UAV platforms without optimization, maintaining high detection accuracy may lead to decreased inference speed, making it difficult to meet the real-time processing requirements of drone-based monitoring and autonomous navigation. Conversely, naively simplifying the model to pursue real-time capability would compromise its detection performance in natural aerial scenarios involving small targets and occlusion, thereby negating the practical significance of the architectural improvements.

Subsequent research will continue to explore ways to further optimize the model structure and reduce computational complexity while maintaining high accuracy. This includes investigating more efficient variants of the EFFA-Net preprocessing module with reduced latency, exploring lightweight attention mechanisms for SwiGLUConv, and designing more parameter-efficient multi-scale fusion strategies for LF-MSCF. Additionally, future work will focus on exploring model compression techniques, including knowledge distillation, pruning, and quantization, to enable real-time deployment on edge UAV platforms. The objective is to achieve a better balance between real-time performance and detection accuracy, thereby promoting the practical application and deployment of the algorithm in real-world drone-based surveillance, inspection, and monitoring systems.

Most importantly, future research will focus on enabling dynamic recognition and incremental learning of unknown categories within open-world drone detection frameworks. This includes integrating open-set recognition capabilities to identify and reject unknown objects while maintaining high precision on known classes, exploring incremental learning strategies that allow the model to update its knowledge base with new object categories over time without catastrophic forgetting, and investigating self-supervised and few-shot learning paradigms to reduce dependency on large-scale annotated datasets when introducing new categories. These advancements would significantly enhance the applicability of drone-based object detection in emerging scenarios such as disaster response, where unknown obstacles may appear; environmental monitoring, where novel targets need to be cataloged; and security surveillance, where new threats must be identified promptly.

5. Conclusions

This paper proposes an improved YOLOv13-based drone small object detection algorithm to address challenges in aerial drone photography, including uneven scale, complex backgrounds, and dense occlusion. The algorithm integrates the EFA-Net module into the preprocessing stage, SwiGLUConv into the backbone and LF-MSCF into the neck, enabling effective integration of multi-sensory field information and enhanced contextual representation for small objects. Experiments demonstrate that the proposed method can achieve stable small object recognition in complex drone environments. Future research will focus on enabling dynamic recognition and incremental learning of unknown categories within open-world drone detection frameworks.

Author Contributions: Conceptualization, M.Z. and S.H.; methodology, M.Z.; software, S.H.; validation, M.Z., S.H. and J.W.; formal analysis, M.Z.; investigation, M.Z.; resources, M.Z.; data curation, S.H.; writing—original draft preparation, S.H.; writing—review and editing, M.Z.; visualization, M.Z.; supervision, J.W.; project administration, C.W.; funding acquisition, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China, grant number 2023YFB4704404; the Outstanding Talent Development Program in Yanqing District grant number YQRC2024-18; the R&D Program of Beijing Municipal Education Commission, grant number KM202410009014; and the Yuxiu Innovation Project of NCUT, grant number 2024NCUTYXCX107. The APC was funded by the North China University of Technology.

Data Availability Statement: The data supporting the findings of this study are openly available in the following repositories: VisDrone2019-DET <https://github.com/VisDrone/VisDrone-Dataset> (accessed on 1 January 2026), HazyDet <https://github.com/GrokCV/HazyDet> (accessed on 10 February 2026), and UAVDT <https://github.com/dataset-ninja/uavdt> (accessed on 15 February 2026).

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

References

1. Yan, B.; Wei, Y.; Liu, S.; Huang, W.; Feng, R.; Chen, X. A Review of Current Studies on the Unmanned Aerial Vehicle-Based Moving Target Tracking Methods. *Def. Technol.* **2025**, *51*, 201–219. [\[CrossRef\]](#)
2. Li, Z.; Yan, Y.; Wei, M.; Ge, B.; Su, N. FGOD-YOLOv8: Fine-Grained Object Detection for Crops and Weeds. *IEEE Signal Process. Lett.* **2025**, *32*, 791–795. [\[CrossRef\]](#)
3. Wang, X.; Yao, F.; Li, A.; Xu, Z.; Ding, L.; Yang, X.; Zhong, G.; Wang, S. DroneNet: Rescue Drone-View Object Detection. *Drones* **2023**, *7*, 441. [\[CrossRef\]](#)
4. Szeliski, R. *Computer Vision: Algorithms and Applications*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2022.
5. Uebis, R.; Mami, S.; Chokmani, K. Drones in Precision Agriculture: A Comprehensive Review of Applications, Technologies, and Challenges. *Drones* **2024**, *8*, 686. [\[CrossRef\]](#)
6. Yi, H.; Liu, B.; Zhao, B.; Liu, E. Small Object Detection Algorithm Based on Improved YOLOv8 for Remote Sensing. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2023**, *17*, 1734–1747. [\[CrossRef\]](#)
7. Yang, Y.; Dai, J.; Wang, Y.; Chen, Y. FM-RTDETR: Small Object Detection Algorithm Based on Enhanced Feature Fusion with Mamba. *IEEE Signal Process. Lett.* **2025**, *32*, 1570–1574. [\[CrossRef\]](#)
8. Li, C.; Hei, Y.; Li, W.; Xiao, Z. MIT-DETR: Mamba-in-Transformers for Efficient Small Object Detection in SAR Images. *IEEE Signal Process. Lett.* **2025**, *32*, 3097–3101. [\[CrossRef\]](#)
9. Ma, S.; Zhang, Y.; Peng, L.; Sun, C.; Ding, B.; Zhu, Y. OWRT-DETR: A Novel Real-Time Transformer Network for Small Object Detection in Open Water Search and Rescue from UAV Aerial Imagery. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4205313. [\[CrossRef\]](#)
10. 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. [\[CrossRef\]](#)
11. 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. [\[CrossRef\]](#)

12. Zhou, Y.; Wei, Y. UAV-DETR: An Enhanced RT-DETR Architecture for Efficient Small Object Detection in UAV Imagery. *Sensors* **2025**, *25*, 4582. [\[CrossRef\]](#)
13. Li, Z.; Dong, Y.; Shen, L.; Liu, Y.; Pei, Y.; Yang, H.; Zheng, L.; Ma, J. Development and Challenges of Object Detection: A Survey. *Neurocomputing* **2024**, *598*, 128102. [\[CrossRef\]](#)
14. Bi, J.; Zheng, X.; Li, K.; Zhang, H.; Zhang, G.; Lei, T. SDWS-YOLO: A Lightweight Small Object Detection Model for Aerial Remote Sensing Images. *IEEE Signal Process. Lett.* **2025**, *32*, 4354–4358. [\[CrossRef\]](#)
15. Jiang, L.; Yuan, B.; Du, J.; Chen, B.; Xie, H.; Tian, J.; Yuan, Z. MFFSODNet: Multiscale feature fusion small object detection network for UAV aerial images. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 5015214. [\[CrossRef\]](#)
16. Guo, Z.; Zhong, M.; Duan, S.; Wang, L. ESGN-YOLO: Enhancing Multi-Scale Small Object Detection via Efficient Feature Fusion and Adaptive Spatial Modeling. *IEEE Signal Process. Lett.* **2025**, *33*, 426–430. [\[CrossRef\]](#)
17. Yang, C.; Cao, Y.; Lu, X. Towards Better Small Object Detection in UAV Scenes: Aggregating More Object-Oriented Information. *Pattern Recognit. Lett.* **2024**, *182*, 24–30. [\[CrossRef\]](#)
18. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Ling, 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 International Conference on Computer Vision Workshop (ICCV), Seoul, Republic of Korea, 27–28 October 2019; pp. 213–226.
19. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 370–386.
20. Feng, C.; Chen, Z.; Li, X.; Wang, C.; Yang, J.; Cheng, M.-M.; Dai, Y.; Fu, Q. HazyDet: Open-source benchmark for drone-view object detection with depth-cues in hazy scenes. *arXiv* **2024**, arXiv:2409.19833.
21. Qin, X.; Wang, Z.; Bai, Y.; Xie, X.; Jia, H. FFA-Net: Feature Fusion Attention Network for Single Image Dehazing. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11908–11915.
22. Liu, W.; Shi, L.; An, G. An Efficient Aerial Image Detection with Variable Receptive Fields. *Remote Sens.* **2025**, *17*, 2672. [\[CrossRef\]](#)
23. Liu, X.; Peng, H.; Zheng, N.; Yang, Y.; Hu, H.; Yuan, Y. EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023.
24. Chen, J.; Kao, S.; He, H.; Zhuo, W.; Wen, S.; Lee, C.-H.; Chan, S.-H.G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023.
25. Wang, A.; Chen, H.; Lin, Z.; Han, J.; Ding, G. RepViT: Revisiting Mobile CNN from ViT Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2024; pp. 15909–15920.
26. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.S.; Xie, S. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023.
27. 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*; IEEE: Piscataway, NJ, USA, 2024; pp. 16965–16974.
28. Shen, F.; Cui, J.; Li, W.; Zhou, S. TinyDef-DETR: A Transformer-Based Framework for Defect Detection in Transmission Lines from UAV Imagery. *arXiv* **2025**, arXiv:2509.06035. [\[CrossRef\]](#)
29. Zheng, Z.; Zhao, J.; Fan, J. A DCM-DETR Object Detector for Drone Images in Harsh Environments. *AIP Adv.* **2025**, *15*, 075243. [\[CrossRef\]](#)
30. Chen, J.; Liu, N.; Sun, H.; Wang, Y. Freq-DETR: Frequency-aware Transformer for Real-time Small Object Detection in Unmanned Aerial Vehicle Imagery. *Expert Syst. Appl.* **2025**, *198*, 129710. [\[CrossRef\]](#)
31. Wang, Z.; Yu, P.; Zhang, X.; Sun, H. BAP-DETR: Efficient Drone Object Detection Network Based on Bipartite Attentive Processing and Dual Fusion Encoder. *Comput. Vis. Image Underst.* **2025**, *262*, 104565. [\[CrossRef\]](#)
32. Zhang, Z. Drone-YOLO: An Efficient Neural Network Method for Target Detection in Drone Images. *Drones* **2023**, *7*, 526. [\[CrossRef\]](#)
33. Yang, L.; Wang, Y.; Kong, L.; Bai, Y.; Tao, B. Object Detection of VisDrone Based on Attention Mechanism and FasterNet. In *Proceedings of the 2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*; IEEE: New York, NY, USA, 2024; pp. 257–261.
34. Yan, H.; Kong, X.; Wang, J.; Tomiyama, H. ST-YOLO: An Enhanced Detector of Small Objects in Unmanned Aerial Vehicle Imagery. *Drones* **2025**, *9*, 338. [\[CrossRef\]](#)
35. Liu, X.; Meng, Q.; Zhang, X.; Li, X.; Li, S. Adaptive Clustering-Guided Multi-Scale Integration for Traffic Density Estimation in Remote Sensing Images. *Remote Sens.* **2025**, *17*, 2796. [\[CrossRef\]](#)

36. Zheng, Z.; Zhao, J.; Fan, J. YOLO-GML: An Object Edge Enhancement Detection Model for UAV Aerial Images in Complex Environments. *PLoS ONE* **2025**, *20*, e0328070. [[CrossRef](#)]
37. Xiao, Y.; Xu, T.; Xin, Y.; Li, J. FBRT-YOLO: Faster and Better for Real-Time Aerial Image Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2025; Volume 39, pp. 8673–8681.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.