algorithms

MDPI

Article

Spatial–Frequency Response Aware Synergy for Small-Object Detection in UAV Aerial Imagery

Dejie Luan [1,2], Chunlong Yang [3], Chunjie Zhang [3,*] and Peng Li [1]

[1] School of Automation and Intelligence, Beijing Jiaotong University, Beijing 100044, China; 23115007@bjtu.edu.cn (D.L.); lipeng@bjtu.edu.cn (P.L.)
[2] Signal and Communication Research Institute, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China
[3] Institute of Information Science, School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044, China; 25115076@bjtu.edu.cn
* Correspondence: cjzhang@bjtu.edu.cn

Abstract

Low-altitude UAV aerial imagery often has complex backgrounds with densely distributed small objects, posing challenges to accurate small-object detection. To address these problems, we propose a spatial–frequency response aware synergistic network for small-object detection in low-altitude UAV aerial imagery. A Frequency-Response-Aware Enhancement Module (FRAEM) is designed to effectively extract discriminative features. The module employs a deterministic stage-aware filtering strategy: Scharr-based edge-sensitive filtering is used in the shallow stage, whereas Gaussian smoothing is used in deeper stages, enabling complementary enhancement of hierarchical representations. A Detail Feature Fusion module (DFFusion) is then developed to improve the efficiency of multi-scale feature fusion. The existing Content-Aware Reassembly of Features (CARAFE) operator is employed for content-aware upsampling and feature alignment, after which DFFusion uses learnable scalar weighting to integrate high-resolution detail information with low-resolution contextual information. A Lightweight Adaptive Decoupled Head (LADH) is also designed to reduce complexity. LADH asymmetrically allocates computational capacity across the prediction tasks: the confidence branch retains stronger spatial processing, whereas the classification and regression branches use lightweight projections; depthwise separable convolution serves as an efficiency-oriented implementation choice. Experiments on the VisDrone2019 and DOTA-v2.0 datasets demonstrate that the proposed method can achieve balance between detection performance and model complexity over mainstream detection methods. Ablation experiments also prove the effectiveness of the proposed components.

Keywords: small-object detection; UAV vision; spatial–frequency response aware feature fusion; deep learning

Check for updates

Academic Editor: Guanqiu Qi

Received: 18 June 2026
Revised: 5 August 2026
Accepted: 5 August 2026
Published: 21 August 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) license.

1. Introduction

Low-altitude economy has emerged as an important role in new infrastructure development and has promoted more efficient allocation and utilization of low-altitude airspace resources [1,2]. Small unmanned aerial vehicles (UAVs) are characterized by their portability, maneuverability, and sensing capabilities, which are playing more and more important roles in integrated aerial–ground monitoring systems [3,4]. UAVs have been widely applied in intelligent traffic management, multispectral crop monitoring, and thermal-

imaging-based wildfire surveillance [5,6]. Reliable onboard visual perception is vital to these applications, while accurate object detection serves as one of its core capabilities.

Modern object detectors are predominantly based on deep networks [7–27] and have benefited from fast developments in visual representation learning and image recognition [28–30]. Object detection methods can be generally divided into two categories: (1) two-stage detectors (e.g., Faster R-CNN), (2) single-stage detectors (e.g., the YOLO series and SSD). Two-stage detectors often have good detection accuracy with high computational overhead and large model sizes, which can hinder real-time deployment on resource-constrained UAV platforms. However, single-stage detectors jointly conduct localization and classification, thereby offering favorable trade-off between detection accuracy and inference efficiency, which are more suitable for onboard UAV applications with computational and latency constraints.

However, accurate small-object detection in low-altitude UAV imagery is challenging. Owing to long imaging distances and view variations, objects (e.g., pedestrians and vehicles) often occupy only a few pixels, and are often densely distributed with scale variations. In addition, visual cues and spatial details are progressively attenuated during deep feature extraction, making the objects difficult to distinguish from the background. Moreover, complex aerial scenes contain repetitive patterns and target-like textures, which further increase the difficulty of object detection.

Most existing UAV-oriented detection methods address these challenges primarily in the spatial domain by introducing dedicated small-object detection heads, attention mechanisms, or enhanced multi-scale feature fusion. However, small objects has weak contrast and similar visual patterns compared with the background. Only using spatial-domain cues may be insufficient for reliable detection. Frequency response aware analysis provides a complementary perspective. High-frequency-based components contain edge, boundary, and fine-texture information, while low-frequency-based components characterize slowly varying structures and global contextual information [31]. Frequency response aware features can help spatial representations by highlighting subtle structural differences.

Recent studies have explored joint spatial–frequency aware feature modeling. Nevertheless, many of them employ fixed or depth-independent processing strategies. Such design cannot fully model the hierarchical characteristics of convolutional features. Shallow layers generally preserve rich local details and spatial boundaries, which may benefit from appropriate high-frequency-based enhancement. In contrast, deeper layers contain more semantic information and require greater suppression of irrelevant responses. Applying the same operation throughout the network may amplify background noise in deep features and weaken discriminative details in shallow features. This observation motivates a deterministic stage-aware filtering strategy that assigns different operations according to the network-stage index rather than through a dynamically learned selector.

Computational efficiency is another challenge. Many accuracy-oriented networks rely on complex feature-fusion necks or heavily coupled detection heads, resulting in increased parameters with computational costs and memory consumption. This makes it incompatible with onboard deployment on UAV edge devices. Therefore, an effective UAV detector should enhance small-object representation while maintaining a lightweight architecture, rather than improving accuracy at the expense of deployment efficiency.

Motivated by these observations, we propose a spatial–frequency response aware synergistic network for small-object detection in low-altitude UAV imagery. Rather than introducing new basic operators, the key idea is to coordinate complementary feature-processing mechanisms according to their roles at different stages of the detection pipeline. Shallow representations receive edge-sensitive processing to preserve local details, deeper representations receive smoothing to suppress irrelevant rapidly varying responses, cross-

scale fusion balances high-resolution details with low-resolution context, and the detection head allocates different computational structures to distinct prediction tasks. Therefore, the novelty of the proposed framework does not lie in the individual use of Scharr filtering, Gaussian filtering, CARAFE, or depthwise separable convolution, but in their stage-specific and task-specific coordination for UAV small-object detection. The main contributions are summarized as follows.

1. We develop an Efficient Feature Extraction Network (EFENet) with a stage-aware complementary filtering strategy. Instead of applying a uniform enhancement operation across all network depths, Scharr-based edge-sensitive filtering is employed in the shallow stage, while Gaussian smoothing is applied in deeper stages according to the hierarchical characteristics of convolutional features. This deterministic stage-dependent rule provides an explicit inductive bias for detail preservation in shallow representations and suppression of irrelevant rapidly varying responses in deeper representations.
2. We construct a Multi-Scale Detail-Enhanced Feature Fusion Network (MDEFNet) for detail-preserving cross-scale integration. CARAFE is employed as an existing content-aware upsampling operator for feature alignment, whereas the proposed DFFusion mechanism focuses on the subsequent learnable integration of high-resolution detail information and low-resolution contextual information.
3. We design a Lightweight Adaptive Decoupled Head (LADH) whose contribution lies in an asymmetric task-specific structure rather than in a new convolutional primitive. The confidence branch retains deeper spatial processing, while the classification and regression branches use lightweight projections, thereby allocating computational capacity according to the requirements of the three prediction tasks. Depthwise separable convolution is adopted as an efficiency-oriented implementation choice rather than being treated as a novel operator.
4. Extensive experiments on the VisDrone2019 and DOTA v2.0 datasets demonstrate that the proposed network achieves a favorable trade-off of detection accuracy, model complexity, and inference speed compared with mainstream detectors. Ablation studies further verify the effectiveness of the proposed components.

2. Related Work

2.1. Spatial-Domain Small-Object Detection for UAVs

Most existing UAV small-object detection methods used spatial-domain feature enhancement, including increased effective resolution, attention-based feature refinement, and multi-scale feature aggregation. Gao et al. [32] combined parallel dilated convolutions with a dedicated small-object detection head to enlarge the receptive field with fine-grained target information. Sun et al. [33] proposed MIS-YOLOv8 to alleviate feature degradation by integrating multi-level feature extraction with adaptive spatial fusion. Wang et al. [34] introduced PETNet which incorporated a prior-enhanced Transformer, a one-to-many feature fusion mechanism, and additional shallow detection heads. To improve onboard efficiency, Fan et al. [35] proposed LUD-YOLO, a lightweight YOLOv8-based detector that employed multi-scale feature fusion and dynamic sparse attention to balance detection accuracy and computational efficiency for deployment on resource-constrained UAV platforms. Wang et al. [36] proposed RPLFDet which preserved low-level features through RSConv and a residual downsampling module. Although these methods had achieved steady detection improvements, they mainly enhanced feature representations within the spatial domain. When small targets exhibited weak contrast or visual patterns similar to background, spatial cues alone might insufficient for reliable discrimination. However, it

was still hard for cluttered aerial scenes. The incorporation of complementary frequency response aware information was necessary.

2.2. Spatial–Frequency Response Aware Detection

Frequency response aware analysis provided information that complemented spatial appearance cues by separating visual components with different variation rates. High-frequency components generally contained edge, boundary, and fine-texture information, whereas low-frequency components mainly represented slowly varying structures and global context [37,38]. Such frequency response aware representations could help reveal subtle structural differences between small objects and complex backgrounds. Liu et al. [31] developed a spatial–frequency interaction backbone while Hu et al. [39] proposed CSFPR-RTDETR, which integrated feature extraction with enhancement to strengthen the discrimination of small objects from complex backgrounds. These studies demonstrated the effectiveness of frequency-based cues. However, many existing methods used fixed or depth-independent processing strategies. Such designs did not fully explore the hierarchical characteristics of convolutional features. Shallow features retained abundant edges and fine-grained spatial details, and may benefit from frequency response aware enhancement. In contrast, deeper features contained more abstract semantic information and required greater suppression of irrelevant responses. Applying a fixed processing strategy may either amplify background noise in deeper layers or weaken discriminative details in shallow layers.

2.3. Efficient and Lightweight Detection Components

Deploying object detectors on UAV edge devices imposed strict constraints on model size, computational cost, memory consumption, and inference latency. Substantial models had focused on improving efficiency. For feature extraction, lightweight architectures such as MobileNetV4 [40], StarNet [41], and FasterNet [42] reduced computational overhead through efficient convolutional designs. Depthwise separable convolution [43] further decomposed a standard convolution into depthwise and pointwise operations with substantially reduced computational cost and parameter count. For multi-scale feature fusion, feature pyramid architectures such as FPN [44], BiFPN [45], GFPN [46], AFPN [47], and MAFPN [48] enhanced cross-scale feature interaction through different aggregation structures. CARAFE [49] performed content-aware feature reassembly to improve the quality of upsampled feature maps. Building on these established operators, our work focuses on coordinating complementary mechanisms across hierarchical feature extraction, cross-scale fusion, and task-specific prediction rather than introducing the operators individually.

3. Methods

The overall architecture of the proposed spatial–frequency response aware synergistic network for small-object detection in UAV imagery is given in Figure 1. The single-stage network consists of an Efficient Feature Extraction Network (EFENet), a Multi-Scale Detail-Enhanced Feature Fusion Network (MDEFNet), and a Lightweight Adaptive Decoupled Head (LADH). EFENet adopts a progressively downsampled hierarchical architecture and performs collaborative spatial–frequency response aware feature extraction to generate multi-scale representations from the input image. MDEFNet then uses DFFusion to integrate features at different resolutions through a learned scalar balance between aligned high- and low-resolution streams. Finally, the fused features are fed into LADH, which employs task-specific branches to predict bounding-box locations, object categories, and confidence scores.

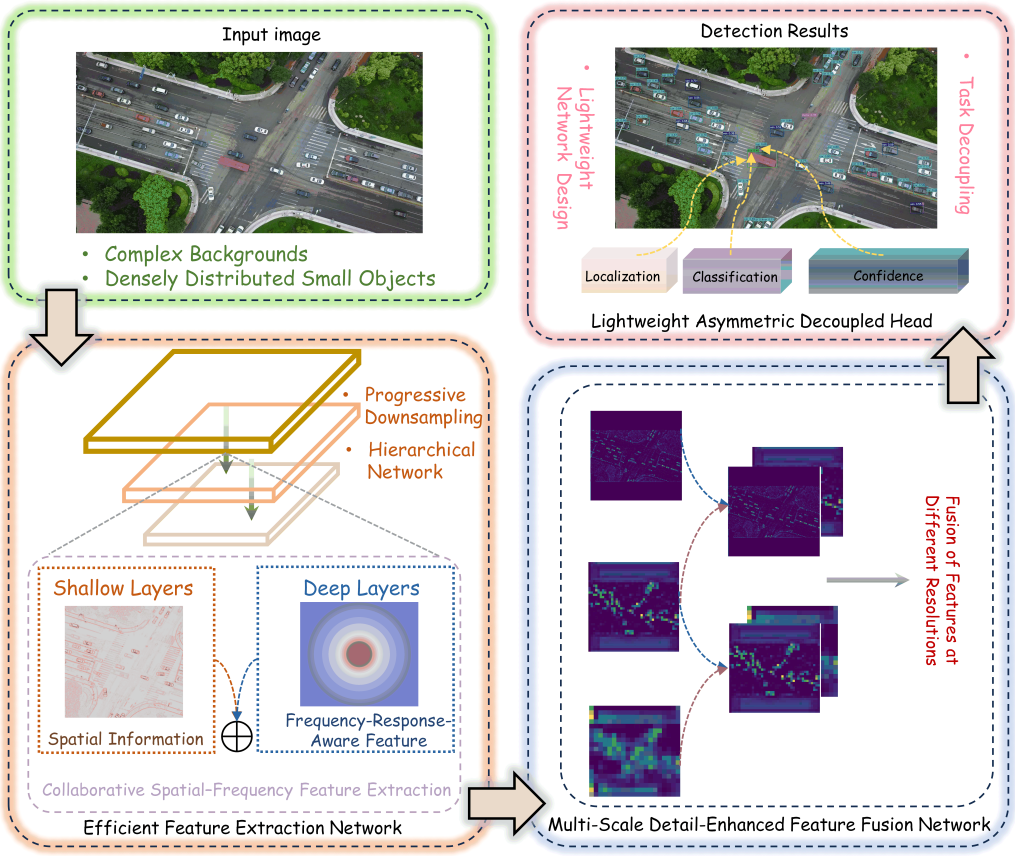


Figure 1. Schematic of the proposed network architecture.

3.1. Efficient Feature Extraction Network

By jointly modeling with spatial–frequency response aware feature extraction, EFENet enhances the representation of small objects in complex aerial scenes. The architecture is illustrated in Figure 2a. EFENet adopts a four-stage hierarchical architecture with progressive downsampling. Each stage consists of a convolutional downsampling module followed by a FRAEM.

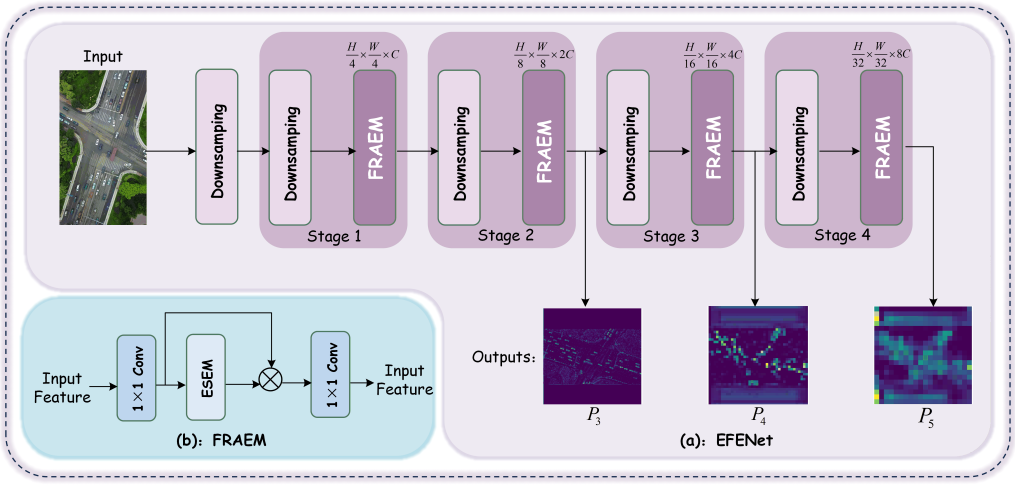


Figure 2. Schematic of the EFENet.

Specifically, the input image is processed by two convolutional downsampling modules, reducing its spatial resolution to 1/4 of the original size. The resulting features are then refined by FRAEM, which emphasizes edge and fine-grained spatial informa-

tion. In each subsequent stage, a convolutional downsampling module first reduces the spatial resolution of the input features, after which FRAEM performs stage-dependent Gaussian smoothing. This progressive design enables EFENet to construct hierarchical spatial–frequency response aware representations across multiple feature levels. Finally, the network outputs feature maps P_3, P_4, and P_5 at resolutions of 1/8, 1/16, and 1/32 of the input image, respectively. These feature maps provide complementary spatial details and semantic information at different scales for subsequent feature fusion.

The structure of FRAEM is illustrated in Figure 2b. It applies a 1×1 convolution on both the input and output to adjust the channel dimensions. An internal residual connection is introduced to facilitate gradient propagation and stabilize the optimization of deep networks. The intermediate features are processed by an Edge–Smoothing Enhancement Module (ESEM), as shown in Figure 3.

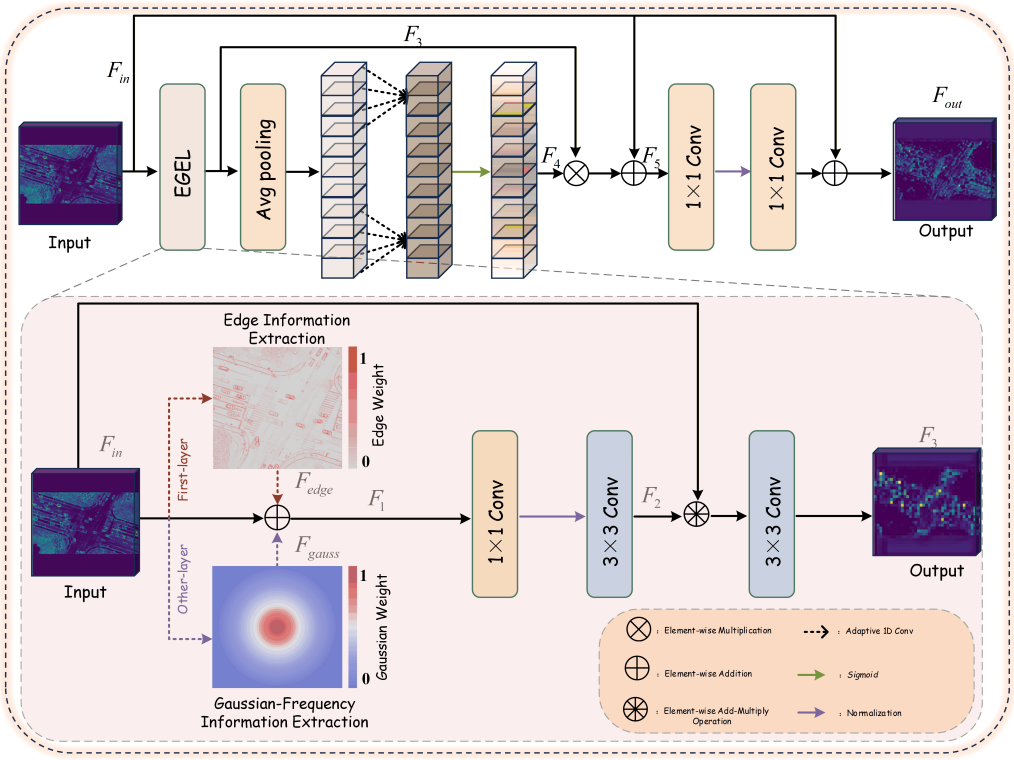


Figure 3. Schematic of the ESEM.

The core component of ESEM is the Edge–Gaussian Enhancement Layer (EGEL). EGEL uses a deterministic stage-aware rule, rather than a learned gating mechanism, to select the corresponding operation from the network-stage index. Specifically, edge enhancement is applied when $s = 1$ to preserve fine-grained spatial details, whereas Gaussian smoothing is applied when $s > 1$ to suppress irrelevant rapidly varying responses. This process is formulated as follows:

$$F_1 = \begin{cases} F_{\text{edge}} + F_{\text{in}}, s = 1, \\ F_{\text{gauss}} + F_{\text{in}}, s > 1. \end{cases} \quad (1)$$

where F_{in} denotes the input feature, F_{edge} represents the feature obtained through edge enhancement, and F_{gauss} represents the feature obtained through Gaussian smoothing. The variable s is the current network-stage index and directly determines the branch in Equation (1); it is not a dynamically learned selection variable.

For a spatial convolution $y = k * x$, the convolution theorem gives the following Fourier representation:

$$Y(\omega) = K(\omega)X(\omega). \tag{2}$$

where $K(\omega)$ specifies how the kernel modulates each frequency response aware component of the input. The Scharr operator approximates first-order spatial derivatives and therefore responds strongly to rapidly varying structures such as edges and object boundaries. In contrast, a Gaussian kernel has a low-pass characteristic and suppresses rapidly varying responses. This complementarity matches the CNN feature hierarchy: shallow feature maps retain higher spatial resolution and abundant boundaries and local structures, which are particularly important for tiny objects, whereas deeper feature maps have lower spatial resolution and stronger semantic abstraction, where excessive rapidly varying responses may reflect irrelevant local fluctuations or background interference. Accordingly, the proposed stage-aware strategy does not treat Scharr and Gaussian filtering as independent enhancements; instead, they serve complementary roles at different network depths, forming a hierarchical detail-preservation–noise-suppression mechanism matched to the representation characteristics of the network.

For edge-feature extraction, we adopt the Scharr operator [50], which provides improved rotational symmetry and gradient estimation accuracy compared with conventional first-order edge operators. The horizontal and vertical Scharr kernels are defined as

$$S_x = \begin{bmatrix} -3 & 0 & 3 \\ -10 & 0 & 10 \\ -3 & 0 & 3 \end{bmatrix}, \qquad S_y = \begin{bmatrix} -3 & -10 & -3 \\ 0 & 0 & 0 \\ 3 & 10 & 3 \end{bmatrix}. \tag{3}$$

where both kernels are fixed during network training. The horizontal and vertical gradient responses are obtained by convolving the input feature F_{in} with S_x and S_y, respectively. The resulting edge feature is computed as

$$F_{\text{edge}} = \sqrt{\left[\text{Conv}\,2D^{S_x}(Fin)\right]^2 + \left[\text{Conv}\,2D^{S_y}(Fin)\right]^2}. \tag{4}$$

The spatial resolution of feature maps decreases with network depth, while fine-grained edge details become less prominent. We employ Gaussian filtering to suppress irrelevant responses while retaining the dominant structural information. A fixed 5×5 depthwise Gaussian convolution is adopted:

$$F_{\text{gauss}} = \text{DConv}^{G}_{5\times5}(F_{\text{in}}), \tag{5}$$

where G denotes the fixed Gaussian kernel. The selected feature F_1 is subsequently processed by a 1×1 convolution, a normalization layer, and a 3×3 convolution:

$$F_2 = \text{Conv}^{3\times3}_{2D}\Big(\text{Norm}\Big(\text{Conv}^{1\times1}_{2D}(F_1)\Big)\Big). \tag{6}$$

The resulting feature is used to modulate the input feature through element-wise multiplication. A residual connection [51] is then introduced:

$$F_3 = \text{Conv}^{3\times3}_{2D}(F_{\text{in}} + F_{\text{in}} \odot F_2), \tag{7}$$

where $\odot$ denotes element-wise multiplication. Global average pooling and an adaptive one-dimensional convolution are used to generate channel weights:

$$F_4 = \text{Sigmoid}(\text{Conv}^{z}_{1D}(\text{GAP}(F_3))). \tag{8}$$

The channel-recalibrated feature is computed as:

$$F_5 = F_{\text{in}} + F_4 \odot F_3. \tag{9}$$

Finally, the channel dimension is expanded from C to $2C$ and then projected back to C:

$$F_{\text{out}} = F_{\text{in}} + \text{Conv}_{2D}^{1\times1}\Big(\text{AN}\Big(\text{Conv}_{2D}^{1\times1}(F_5)\Big)\Big). \tag{10}$$

3.2. Multi-Scale Detail-Enhanced Feature-Fusion Network

We construct a MDEFNet, whose architecture is illustrated in Figure 4. MDEFNet incorporates top-down, bottom-up, and cross-scale fusion pathways to aggregate the multi-level features generated by EFENet. Convolutional modules are used for downsampling, while the Content-Aware Reassembly of Features (CARAFE) operator is employed for content-aware upsampling. The proposed Detail Feature Fusion module (DFFusion) then performs the subsequent learnable integration of aligned features at different resolutions. The fused features are refined by FRAEM to further enhance the discriminative power of visual representations.

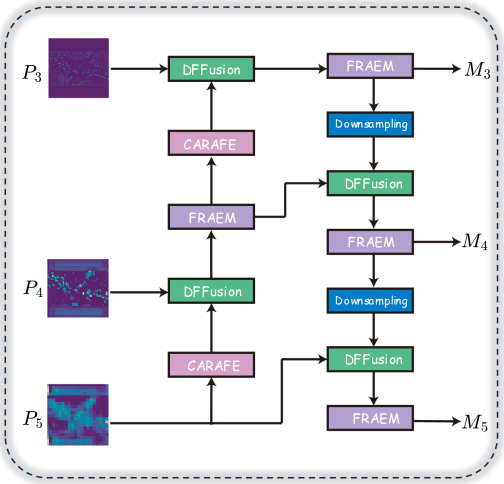


Figure 4. Schematic of the MDEFNet.

When high-resolution features are fused with low-resolution features containing richer contextual information, fine-grained details associated with small objects may be weakened. We propose the DFFusion module, as illustrated in Figure 5. Let $F_l \in \mathbb{R}^{H_l\times W_l\times C_l}$ and $F_h \in \mathbb{R}^{H_h\times W_h\times C_h}$ denote the low-resolution and high-resolution feature maps, respectively. F_l is upsampled to match the spatial resolution of F_h. We employ ARAFE [49], which predicts content-aware reassembly kernels from local feature information and reconstructs the low-resolution feature map accordingly. This operation improves the quality of the upsampled features and facilitates subsequent multi-scale feature fusion. The structure of CARAFE is illustrated in Figure 6. In MDEFNet, CARAFE is employed to perform content-aware upsampling and align the low-resolution features with the high-resolution branch. The aligned features are subsequently processed by DFFusion, which integrates low-resolution contextual information with high-resolution fine-grained details to improve cross-scale feature aggregation. The effects of the upsampling strategy and the fusion mechanism are further analyzed separately in the comparative experiments.

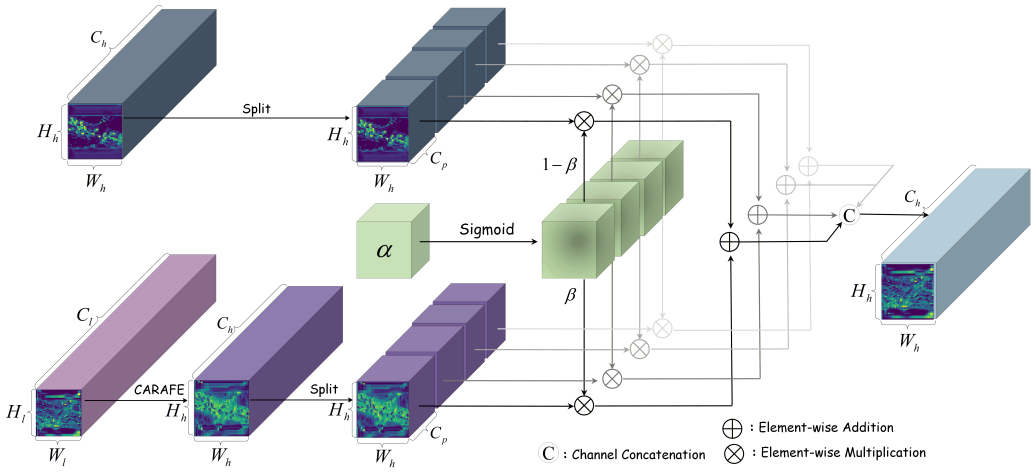


Figure 5. Schematic of the DFFusion module.

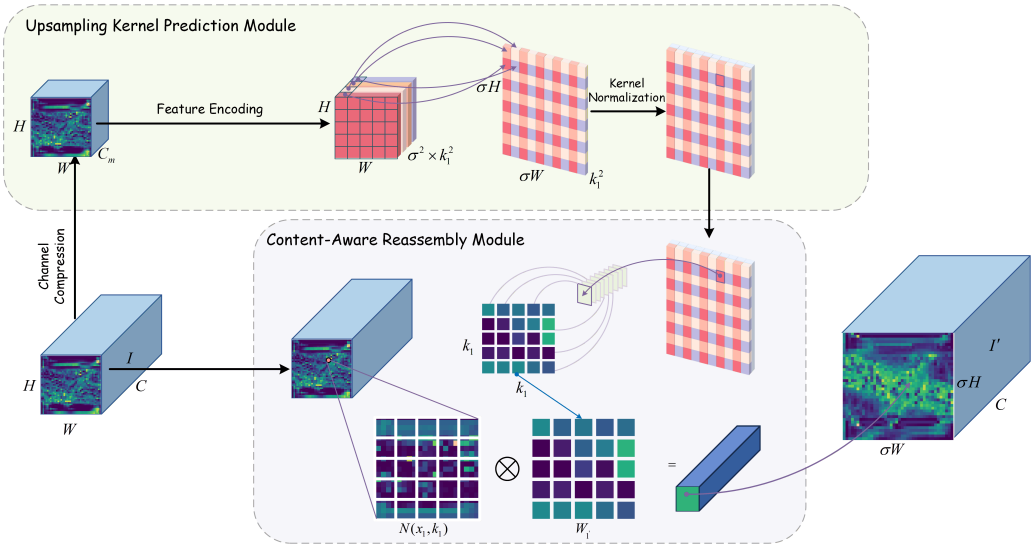


Figure 6. Schematic of the CARAFE module.

CARAFE consists of two components: an upsampling kernel prediction module and a content-aware feature reassembly module. In the kernel prediction module, a 1×1 convolution first compresses the input channels from C to C_m to reduce computational cost. Given an upsampling ratio σ and a reassembly kernel size of $k_1 \times k_1$, the feature encoding operation maps the compressed features to $\sigma^2 k_1^2$ channels. The encoded features are then spatially rearranged and normalized to generate content-aware upsampling kernels with a size of $\sigma H \times \sigma W \times k_1^2$. In the content-aware feature reassembly module, a local neighborhood $N(x_1, k_1)$ of size $k_1 \times k_1$ is extracted around each corresponding sampling location. The features within this neighborhood are weighted by the predicted upsampling kernel and aggregated through a dot product. This process produces an upsampled feature map with a spatial size of $\sigma H \times \sigma W$ and C output channels.

In DFFusion, F_l is matched to F_h via CARAFE. The aligned maps are then split evenly along the channel dimension into four groups ($c_p = c/4$), denoted $\{h_i\}_{i=1}^4 \in \mathbb{R}^{H \times W \times c_p}$ and $\{I_i\}_{i=1}^4 \in \mathbb{R}^{H \times W \times c_p}$ for the high- and low-resolution fragments, respectively. A learnable scalar $\alpha' \in \mathbb{R}^{1 \times 1 \times 1}$ is broadcast across the spatial and channel dimensions to match the fragment shape, giving $\alpha \in \mathbb{R}^{H \times W \times c_p}$. Because this operation only replicates one scalar,

the entries of α are shared rather than independently predicted for different locations or channels. A weight $\beta \in [0, 1]$ is obtained via a sigmoid:

$$\beta = \text{sigmoid}(\alpha), \tag{11}$$

where the broadcast coefficient β controls the contribution of the low-resolution features and $1 - \beta$ that of the high-resolution features. The cross-resolution integration is then realized as a weighted sum:

$$F_i' = \beta \otimes I_i + (1 - \beta) \otimes h_i, \tag{12}$$

where $F_i' \in \mathbb{R}^{H \times W \times c_p}$ is the fused map of the i-th fragment. This learned global mixing coefficient balances the detail of h_i (when $\beta \to 0$) against the context of I_i (when $\beta \to 1$); it should not be interpreted as a separately predicted per-location weight. The fused fragments are concatenated along the channel dimension to produce the final output:

$$F_{out}' = [\, F_1', F_2', F_3', F_4' \,]. \tag{13}$$

Within MDEFNet, FRAEM further refines and enhances the fused maps, ultimately outputting three feature maps M_3, M_4, and M_5 for decoding by the head.

3.3. Lightweight Adaptive Decoupled Head

As illustrated in Figure 7, LADH employs three independent branches for confidence estimation, classification, and bounding-box regression. The confidence branch contains three consecutive 3×3 depthwise separable convolutions. Stacking these convolutional layers enlarges the effective receptive field and strengthens spatial feature modeling, while progressive channel reduction limits the associated computational overhead. This design improves the ability of the confidence branch to distinguish foreground objects from complex backgrounds. In contrast, the classification and regression branches each employ a lightweight 1×1 convolution to produce category scores and bounding-box predictions, respectively. This task-specific design reduces redundant computation while preserving the distinct representation requirements of the three detection tasks.

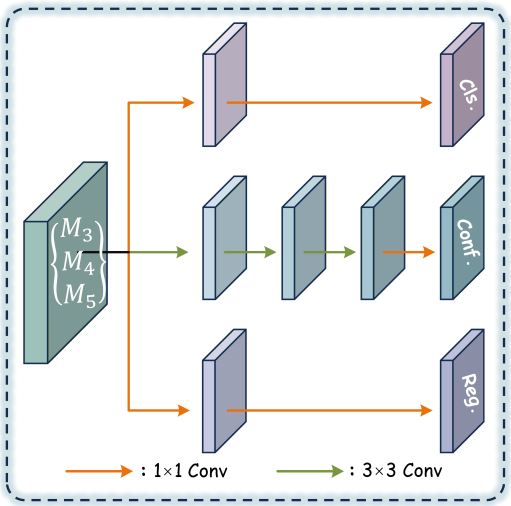


Figure 7. Schematic of the LADH network.

LADH adopts an asymmetric task-specific design to accommodate the different representation requirements of confidence estimation, classification, and localization. Specifically, the confidence branch retains deeper spatial processing to strengthen foreground–

background discrimination, whereas the classification and regression branches employ lightweight projections to reduce redundant computation. This task-dependent allocation of computational capacity enables the detection head to balance representation capability and efficiency.

Depthwise separable convolution is adopted in the confidence branch to improve computational efficiency. It decomposes a standard convolution into a depthwise convolution for channel-wise spatial filtering and a pointwise convolution for cross-channel information interaction [43,52]. Compared with standard convolution, this decomposition substantially reduces the number of parameters and floating-point operations, particularly when the number of input and output channels is large. As a result, LADH improves spatial modeling in the confidence branch without introducing excessive computational complexity.

4. Experimental Results and Analysis

4.1. Datasets and Evaluation Metrics

4.1.1. Datasets

The first dataset used in this study is VisDrone2019 [53], a challenging UAV-view object detection benchmark containing ten categories. We use 6471 images for training, 548 for validation, and 1610 for testing. The dataset covers diverse scenes and imaging conditions, with substantial variations in object scale, density, illumination, and background complexity. Objects with bounding boxes smaller than 50×50 pixels are regarded as small objects. As shown in Figure 8, small objects account for a large proportion of the annotated instances and frequently appear in dense distributions, making VisDrone2019 suitable for evaluating small-object detection in complex UAV imagery. It should be noted that VisDrone distinguishes the categories 'pedestrian' and 'people' according to human posture. A human target in a standing or walking pose is annotated as 'pedestrian', whereas a human in other postures is assigned to the 'people' category.

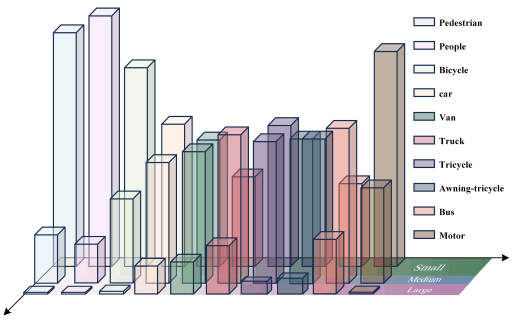


Figure 8. Visualization of object sizes in VisDrone2019.

4.1.2. Evaluation Metrics

We adopt the standard MS COCO evaluation metrics to comprehensively assess detection accuracy across different object scales. The primary metrics are AP, averaged over IoU thresholds from 0.50 to 0.95 with a step size of 0.05, and AP_{50}, evaluated at an IoU threshold of 0.50. We further report AP_S, AP_M, and AP_L to evaluate detection performance on small, medium, and large objects, respectively. Model complexity is measured using the number of parameters (Params) and giga floating-point operations (GFLOPs), while inference efficiency is evaluated in frames per second (FPS).

4.2. Implementation Details

All experiments were conducted on a Linux server equipped with four NVIDIA GeForce RTX 3090 GPUs (24 GB each). The software environment used Python 3.10 and PyTorch 2.0.0 for training and inference. The main hyperparameters are listed in Table 1.

Table 1. Model training configuration.

Parameter	Setting
Image size	640 × 640
Epochs	300
Learning rate	0.01
Optimizer	SGD
Momentum	0.937
Data augmentation	Mosaic

4.3. Comparative Experiments

4.3.1. Comparison of Feature Extraction Backbones

To evaluate the effectiveness of EFENet, we compare it with several representative feature extraction backbones. These include the CNN-based CSPDarkNet [54], MobileNetV4 [40], StarNet [41], and FasterNet [42], as well as the Transformer-based EfficientViT [55]. All backbones are evaluated under the same training and testing settings, and the results are reported in Table 2.

Table 2. Comparison of feature-extraction backbones. In each experiment, only the backbone of the proposed network is replaced while all other components remain unchanged. Per-category and mean AP (%); best results are shown in bold.

Backbone	Params	GFLOPs	AP (%)										mAP (%)
			Ped.	People	Bicycle	Car	Van	Truck	Tri.	A-Tri.	Bus	Motor	
CspDarkNet	2.3 M	7.1	35.6	**28.2**	7.4	76.1	39.1	27.0	21.2	12.4	45.9	37.3	33.0
MobileNetV4	5.5 M	26.4	32.5	25.6	7.5	74.4	35.7	24.6	17.7	9.7	43.8	**43.3**	30.6
StarNet	1.5 M	5.5	31.1	24.0	7.2	73.9	34.1	24.8	17.6	10.5	44.7	34.7	30.3
FasterNet	3.4 M	10.1	**35.8**	27.4	**8.0**	**77.2**	40.5	30.2	21.3	12.3	49.9	37.9	34.4
EfficientViT	3.3 M	8.6	30.8	24.8	6.5	72.4	35.6	26.8	17.9	9.6	42.4	33.1	30.2
EFENet	2.1 M	8.5	**35.8**	27.0	**8.0**	77.1	**41.2**	**31.1**	**21.6**	**13.8**	**54.1**	38.6	**34.8**

EFENet achieves the highest overall mAP of 34.8%, outperforming FasterNet by 0.4 percentage points. Meanwhile, EFENet uses only 2.1 M parameters and 8.5 GFLOPs, representing reductions of 1.3 M parameters and 1.6 GFLOPs compared with FasterNet. These results indicate that EFENet achieves a more favorable balance between detection accuracy and computational complexity. In terms of category-specific performance, EFENet obtains the best or tied-best AP in seven of the ten categories, including pedestrian, bicycle, van, truck, tricycle, awning-tricycle, and bus. In particular, it achieves AP values of 41.2%, 31.1%, and 54.1% for van, truck, and bus, respectively. These improvements demonstrate that the proposed feature extraction strategy effectively preserves discriminative information for objects with different scales and appearance characteristics. Although StarNet has fewer parameters and lower computational cost, its detection accuracy is notably lower. EfficientViT also exhibits lower category-level accuracy while requiring a similar amount of computation. Overall, EFENet provides a favorable trade-off between feature representation capability and computational efficiency.

4.3.2. Ablation Study of the Stage-Aware Filtering Strategy

To further investigate the effects of the filtering operations in FRAEM, we conduct controlled ablation experiments using Scharr-only and Gaussian-only variants and compare them with the complete FRAEM. In all experiments, the remaining network architecture, training configuration, and evaluation settings are kept unchanged. The results are reported in Table 3.

Table 3. Ablation study of different filtering configurations in FRAEM. All other network components and training settings are kept unchanged. The best results are shown in bold.

Method	AP	AP_{50}	AP_S	AP_M	AP_L	GFLOPs	Params (M)
Scharr-only	18.5	31.7	9.9	28.2	37.2	8.5	2.1
Gaussian-only	18.7	32.0	9.7	28.8	36.6	8.5	2.1
FRAEM	**20.5**	**34.8**	**10.9**	**32.5**	**39.6**	8.5	2.1

As shown in Table 3, the Scharr-only and Gaussian-only variants achieve AP values of 18.5% and 18.7%, respectively, whereas the complete FRAEM increases AP to 20.5%. Similar improvements are observed for AP_{50}, which increases from 31.7% with Scharr-only filtering and 32.0% with Gaussian-only filtering to 34.8% with FRAEM. Compared with the Scharr-only variant, FRAEM improves AP, AP_{50}, AP_S, AP_M, and AP_L by 2.0, 3.1, 1.0, 4.3, and 2.4 percentage points, respectively. Compared with the Gaussian-only variant, the corresponding improvements are 1.8, 2.8, 1.2, 3.7, and 3.0 percentage points.

The results indicate that relying on either filtering operation alone is less effective than using them according to the hierarchical characteristics of the network. Scharr filtering emphasizes local boundaries and fine-grained structures, which are important for shallow feature representations, while Gaussian smoothing suppresses rapidly varying responses and provides smoother representations in deeper stages. Their stage-aware use in FRAEM therefore exploits the complementary properties of the two filtering operations and improves feature representation across different object scales. Notably, the three configurations have the same parameter count and comparable computational complexity, indicating that the performance improvement mainly results from the filtering strategy rather than an increase in model size.

4.3.3. Comparison of Upsampling Modules

To evaluate the effect of different upsampling strategies in the feature fusion network, we compare CARAFE with DySample [56], the upsampling module used in YOLOv6 [57], and EUCB [58]. The results are reported in Table 4. CARAFE achieves the highest overall AP and the best AP_S among the compared methods. This result indicates that content-aware feature reassembly is effective in retaining fine-grained information associated with small objects during upsampling. Although CARAFE introduces slightly higher computational complexity, the improvement in detection accuracy remains notable. CARAFE provides a favorable trade-off between small-object detection performance and computational cost.

Table 4. Comparison of different upsampling modules. In all experiments, DFFusion and the other network components are kept unchanged, and only the upsampling operator is replaced. Best results are shown in bold.

Method	AP	AP_{50}	AP_S	AP_M	AP_L	GFLOPs	Params (M)
DySample	20.0	33.8	10.5	31.2	38.8	8.3	2.0
Upsample	19.8	33.6	10.4	30.8	39.1	8.2	2.0
EUCB	20.1	33.7	10.4	31.2	39.1	8.7	2.1
CARAFE	**20.5**	**34.8**	**10.9**	**32.5**	**39.6**	8.5	2.1

4.3.4. Comparison of Fusion Methods

DFFusion is designed for efficient fusion of multi-resolution feature maps. We compared it with Concat [59], PSFusion [60], and MFM [61]; results are shown in Table 5. DFFusion achieves the best AP, AP_{50}, AP_S, and AP_M, demonstrating more effective multi-scale fusion enabled by its adaptive mechanism, which fully exploits informative cues at different resolutions. Although MFM is slightly higher on AP_L, DFFusion shows stronger

overall multi-scale representation and fusion. Its complexity and parameter count are not appreciably higher than those of the alternatives. DFFusion improves fusion performance while maintaining model complexity. These results demonstrate that the performance gain does not arise solely from the use of CARAFE. CARAFE provides content-aware feature alignment, whereas DFFusion further improves cross-scale integration of high-resolution detail and low-resolution contextual information.

Table 5. Comparison of different feature fusion methods. CARAFE is used as the upsampling operator in all experiments, while only the feature fusion module is replaced and all other components remain unchanged. Best results are shown in bold.

Method	AP	AP_{50}	AP_S	AP_M	AP_L	GFLOPs	Params (M)
Concat	20.1	33.9	10.4	31.3	38.7	8.4	2.0
PSFusion	20.0	34.0	10.5	31.4	38.5	8.6	2.2
MFM	20.2	34.0	10.7	31.8	**40.2**	8.4	2.0
DFFusion	**20.5**	**34.8**	**10.9**	**32.5**	39.6	8.5	2.1

4.3.5. Comparison of Feature-Fusion Networks

To evaluate the effectiveness of MDEFNet in aggregating and enhancing multi-scale features, we compare it with several representative feature fusion networks, including FPN [44], BiFPN [45], GFPN [46], AFPN [47], and MAFPN [48]. The results are reported in Table 6.

Table 6. Comparison of feature-fusion networks. In each experiment, only the feature-fusion network of the proposed model is replaced while all other components remain unchanged; Per-category and mean AP (%); best mAP in bold.

Network	Params	GFLOPs	AP (%)										mAP (%)
			Ped.	People	Bicycle	Car	Van	Truck	Tri.	A-Tri.	Bus	Motor	
FPN	2.2 M	8.7	35.4	27.4	7.1	76.4	40.3	29.4	21.2	11.3	43.7	36.1	32.8
BiFPN	1.4 M	7.7	35.0	26.4	6.4	76.1	40.7	28.5	21.1	11.4	51.1	37.0	33.6
GFPN	2.9 M	9.4	34.7	27.0	7.5	76.5	40.4	29.7	20.2	13.6	46.6	37.4	33.3
AFPN	2.2 M	11.1	34.0	28.6	6.9	76.0	39.7	28.2	21.0	12.2	47.6	36.2	33.0
MAFPN	2.3 M	9.4	36.7	27.7	7.9	77.0	41.0	30.3	21.5	12.9	50.0	37.2	34.3
MDEFNet	2.1 M	8.5	35.8	27.0	8.0	77.1	41.2	31.1	21.6	13.8	54.1	38.6	**34.8**

MDEFNet achieves mAP of 34.8%, outperforming all compared feature fusion networks. It obtains the best category-level AP in eight of the ten categories, including bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor. These results indicate that MDEFNet effectively integrates complementary information across different feature scales and improves the representation of objects with diverse appearances and scales. MAFPN achieves the highest AP for pedestrian, while AFPN performs best for people. However, their overall mAP values remain lower than that of MDEFNet. MDEFNet requires 2.1 M parameters and 8.5 GFLOPs. Compared with MAFPN, it improves mAP by 0.5 percentage points while reducing the parameter count by 0.2 M and the computational cost by 0.9 GFLOPs. Although BiFPN has lower model complexity, its mAP is 1.2 percentage points lower than that of MDEFNet. These results demonstrate that MDEFNet provides a favorable trade-off between detection accuracy and computational complexity for small-object detection in UAV imagery.

4.3.6. Comparison of Detection Heads

To verify the advantage of LADH, we compared it with the coupled head Coupled-Head [62] and several decoupled heads: Efficient-head [63], LSCSBD [64], LSCD-LQE [65], and LSDECD [66]; results are shown in Table 7.

Table 7. Comparison of detection heads. In each experiment, only the detection head of the proposed network is replaced while all other components remain unchanged; Per-category and mean AP (%); best mAP in bold.

Head	Params	GFLOPs	AP (%)										mAP (%)
			Ped.	People	Bicycle	Car	Van	Truck	Tri.	A-Tri.	Bus	Motor	
Coupled-Head	2.2 M	9.1	31.8	19.9	6.0	73.2	36.2	26.2	16.7	10.4	45.1	29.3	29.7
Efficient-head	2.1 M	8.5	35.5	26.8	8.4	76.6	40.0	29.4	20.7	12.9	49.5	37.4	33.7
LSCSBD	2.2 M	9.6	36.4	26.9	8.2	76.4	40.9	29.6	19.3	11.3	47.6	35.0	33.2
LSCD-LQE	2.3 M	8.9	37.1	28.2	9.0	77.0	40.5	29.9	21.1	12.2	50.7	38.2	34.5
LSDECD	2.1 M	9.4	37.3	28.4	7.7	77.6	41.0	31.0	20.5	12.7	51.0	37.9	34.6
LADH	2.1 M	8.5	35.8	27.0	8.0	77.1	41.2	31.1	21.6	13.8	54.1	38.6	**34.8**

Decoupled detection heads consistently outperform the coupled head across the evaluated categories, indicating that separating classification and regression can reduce interference between task-specific feature representations. Among the decoupled designs, LADH achieves the highest AP in six categories. LADH and Efficient-Head have the lowest parameter counts and computational costs among the compared methods. Nevertheless, LADH surpasses Efficient-Head by 1.1 percentage points in mAP. These results demonstrate that LADH achieves a favorable balance between detection accuracy and computational complexity.

4.3.7. Comparison with State-of-the-Art Detectors

To evaluate the performance and generalization capability of the proposed network for low-altitude UAV small-object detection, we evaluated it on the public VisDrone2019 and DOTA v2.0 datasets against mainstream detectors: the two-stage Faster R-CNN [67]; the Transformer-based RT-DETR [68] and DEIM [69]; and the CNN-based YOLO family [70–75]. Results are shown in Table 8.

Table 8. Comparison with state-of-the-art detectors on VisDrone2019 [53] and DOTA v2.0 [76]. Best results are shown in bold.

Network	Params	GFLOPs	VisDrone2019					DOTA v2.0				
			AP_{50}	AP_S	AP_M	AP_L	FPS	AP_{50}	AP_S	AP_M	AP_L	FPS
Faster R-CNN	41.4 M	208	32.9	9.5	30.9	**42.9**	23.1	31.0	13.1	35.8	38.4	25.6
RT-DETR	19.9 M	57	33.3	**13.9**	27.5	42.3	68.1	44.1	18.4	45.2	51.1	87.9
DEIM	3.7 M	7.2	32.2	9.0	26.2	37.6	177.7	39.8	15.2	38.0	43.1	232.6
YOLOX-Tiny	2.2 M	8.7	27.8	7.6	22.1	27.8	164.4	44.6	17.1	41.2	50.1	199.0
YOLOv8n	3.0 M	8.1	25.9	5.9	22.5	33.9	171.4	48.7	16.3	45.9	56.1	225.9
YOLOv10n	2.3 M	6.5	26.1	6.3	22.4	29.2	210.6	45.2	18.1	40.9	49.7	301.3
YOLOv11n	2.6 M	6.3	25.8	5.8	22.5	31.6	199.8	49.4	**21.2**	**48.8**	54.0	298.6
YOLOv12n	2.6 M	6.3	25.9	5.7	22.4	34.6	197.6	47.5	17.7	41.2	48.6	**303.5**
YOLOv13n	2.5 M	6.2	24.4	5.5	21.0	31.7	175.4	42.1	18.0	38.5	43.7	211.7
Ours	**2.1 M**	8.5	**34.8**	10.9	**32.5**	39.6	187.6	**50.3**	19.6	46.2	**57.1**	237.4

In terms of detection accuracy, the proposed network achieves the highest AP_{50} among the evaluated methods on both datasets, reaching 34.8% on VisDrone2019 and 50.3% on DOTA v2.0. It also obtains competitive results for AP_S, AP_M, and AP_L, indicating stable detection performance across different object scales. On DOTA v2.0, the improvement over YOLOv11n in AP_{50} is 0.9 percentage points, while the proposed model requires 8.5 GFLOPs compared with 6.3 GFLOPs for YOLOv11n.

In terms of model complexity, the proposed network contains only 2.1 M parameters, which is the smallest among the evaluated methods. However, parameter count alone does not directly determine practical inference speed. The proposed model requires 8.5 GFLOPs, and the CARAFE-based feature reassembly and multi-scale feature-fusion operations introduce additional computation and memory-access overhead. Practical FPS is also affected by operator parallelism and hardware/software optimization. Consequently, a smaller parameter count does not necessarily correspond to a higher inference speed.

Regarding inference efficiency, the proposed network reaches 187.6 FPS on VisDrone2019 and 237.4 FPS on DOTA v2.0, remaining competitive with the evaluated lightweight detectors. Overall, the results demonstrate a favorable trade-off among detection accuracy, parameter count, computational complexity, and inference efficiency.

4.3.8. Qualitative Comparison

To provide comprehensive qualitative evaluation, we compare the proposed method with RT-DETR, DEIM, YOLOv11n, YOLOv12n, and YOLOv13n. These methods cover different detector architectures, including Transformer-based end-to-end detectors and lightweight CNN-based detectors. Four representative images from the VisDrone2019 validation set are selected, containing densely distributed tiny objects, distant targets, complex backgrounds, and substantial scale variations.

As shown in Figure 9, correct detections, false positives, and false negatives are marked in green, blue, and red, respectively. Enlarged regions are additionally provided to facilitate the comparison of challenging small targets. The compared detectors exhibit different degrees of missed detections and false positives, particularly in densely populated regions and areas with complex background interference. In contrast, the proposed method produces fewer visible detection errors in the highlighted regions and retains more complete detection results for small and densely distributed objects.

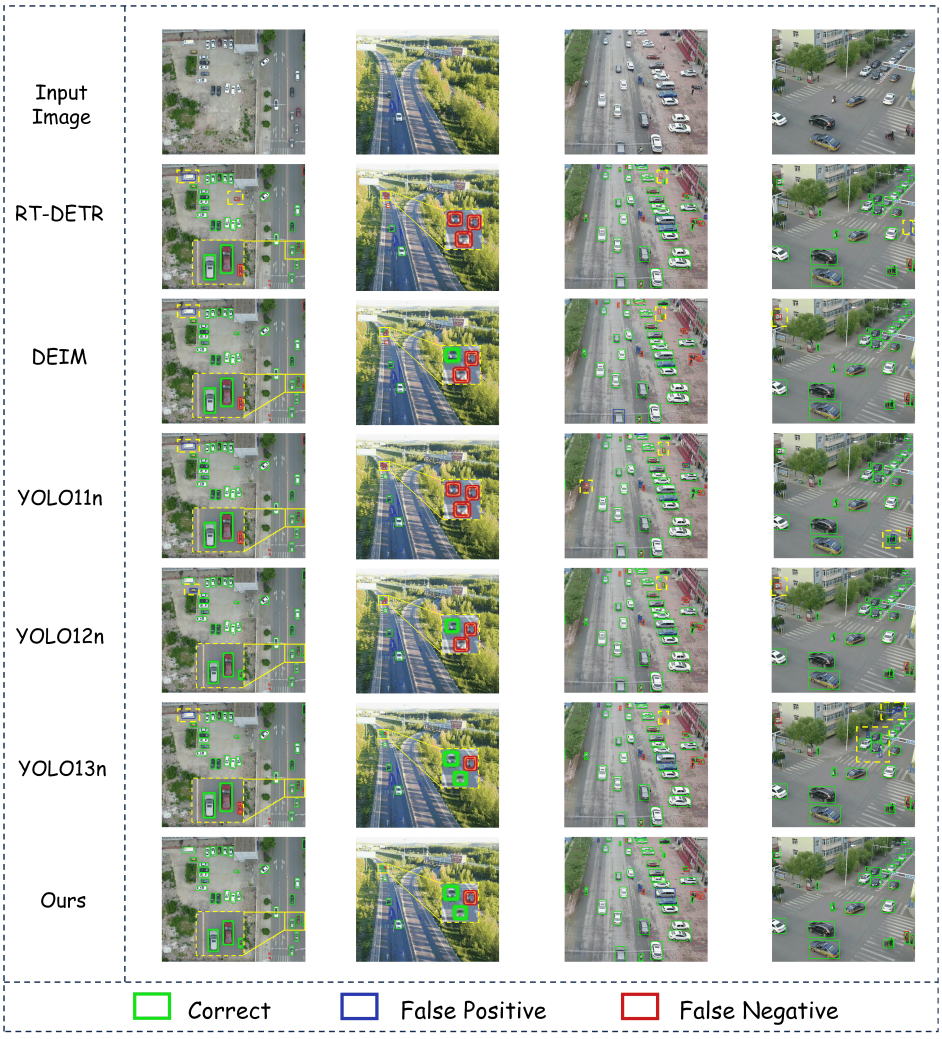


Figure 9. Visualization of detection results on VisDrone2019.

4.3.9. Visualization of Hierarchical Feature Responses

To further examine the representation of small objects during hierarchical feature extraction, we visualize the intermediate feature responses at the P3, P4, and P5 levels for different backbone networks. The same input images and visualization procedure are used for all compared methods, as shown in Figure 10.

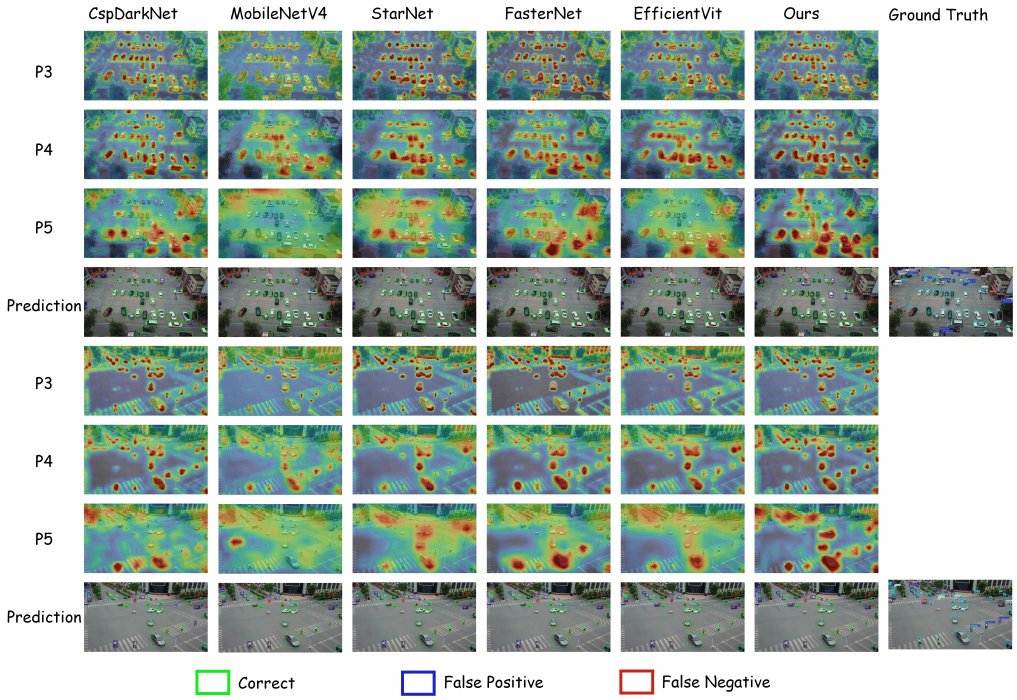


Figure 10. Visualization of intermediate feature responses at different hierarchical levels for representative backbone networks. P3, P4, and P5 denote feature maps with progressively deeper semantic representations.

At the relatively shallow P3 level, most backbone networks preserve visible responses around densely distributed objects. As the feature hierarchy becomes deeper from P3 to P5, however, the responses of several lightweight backbones become increasingly diffuse or less distinguishable around small-object regions. In comparison, EFENet retains relatively concentrated target-related responses at the deeper P4 and P5 levels in the illustrated scenes. This observation is consistent with the stage-aware design of EFENet, in which shallow edge-sensitive filtering emphasizes fine-grained structures while deeper Gaussian smoothing suppresses unnecessary rapidly varying responses. The visualization therefore provides qualitative evidence that EFENet facilitates the preservation of small-object-related representations during hierarchical feature extraction.

5. Conclusions

We proposed a spatial–frequency response aware synergistic network for small-object detection in low-altitude UAV imagery. Within EFENet, FRAEM uses a deterministic stage-aware filtering strategy to enhance edge information in the shallow stage and perform Gaussian smoothing in deeper stages. Within MDEFNet, the existing CARAFE operator provides content-aware upsampling and feature alignment, while the proposed DFFusion mechanism performs the subsequent integration of high-resolution spatial details and low-resolution contextual information. LADH uses an asymmetric task-specific design in which the confidence branch retains stronger spatial processing and the classification and regression branches use lightweight projections; depthwise separable convolution is employed as an efficiency-oriented operator. Experiments on VisDrone2019 and DOTA v2.0

demonstrate that the proposed method achieves a favorable trade-off among detection accuracy, model complexity, and inference speed. The contribution of the framework lies primarily in the coordinated use of stage-aware feature processing, detail-preserving cross-scale integration, and asymmetric task-specific prediction, rather than in the individual basic operators. Future work will focus on further optimizing the network architecture and improving its robustness under adverse imaging conditions, including low illumination, rain, and fog.

Author Contributions: Conceptualization, D.L. and P.L.; Methodology, D.L. and C.Y.; Software, D.L.; Validation, D.L. and C.Y.; Formal analysis, D.L. and C.Y.; Investigation, C.Z. and P.L.; Resources, C.Z. and P.L.; Data curation, C.Y.; Writing—original draft, D.L. and C.Y.; writing—review and editing, D.L., C.Y. and C.Z.; Visualization, C.Z.; Supervision, C.Z. and P.L.; Project administration, C.Z. and P.L.; Funding acquisition, C.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Funds for the Central Universities (grant number 2025JBZX062).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository. The original data presented in the study are openly available in VisDrone-Dataset at https://github.com/VisDrone/VisDrone-Dataset (accessed on 4 August 2026), and in DOTA v2.0 at https://captain-whu.github.io/DOTA/evaluation.html (accessed on 4 August 2026).

Acknowledgments: The authors would like to thank the editors and the anonymous reviewers for their constructive comments, which helped to improve the quality of this manuscript.

Conflicts of Interest: Dejie Luan is employed by Signal and Communication Research Institute, China Academy of Railway Sciences Corporation Limited. He's now pursing his Ph.D degree in School of Automation and Intelligence, Beijing Jiaotong University. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Hua, W.; Chen, Q. A survey of small object detection based on deep learning in aerial images. *Artif. Intell. Rev.* **2025**, *58*, 162. [CrossRef]
2. Feng, Q.H.; Wang, Z.X.; Sun, C.C.; Shao, Z.W. Small object detection in drone images via foreground refinement and multidimensional inductive bias self-attention. *J. Image Graph.* **2025**, *30*, 3547–3563. [CrossRef]
3. Zhang, C.J.; Liu, C.K.; Duan, S.H.; Zheng, X.L.; Yu, T.Y.; Zhang, J.T. Embodied cognitive intelligence guided moon sample collection. *Innovation* **2025**, *6*, 100939. [CrossRef] [PubMed]
4. Zeng, N.Y.; Li, X.Y.; Wu, P.S.; Li, H.; Luo, X. A novel tensor decomposition-based efficient detector for low-altitude aerial objects with knowledge distillation scheme. *IEEE/CAA J. Autom. Sin.* **2024**, *11*, 487–501. [CrossRef]
5. Cao, P.; Zhang, C.J.; Huang, J.; Wang, Y.T.; Fan, L.L.; Zhao, Y.; Zheng, X.L. Cross-light contrastive learning of pavement crack segmentation in low-light conditions for autonomous road inspection vehicles. *IEEE Trans. Intell. Veh.* **2024**. [CrossRef]
6. Duan, S.H.; Zhang, C.J.; Zheng, X.L.; Wang, Y.T.; Zhang, H.; Wang, F.Y. 24-h lane line detection via parallel scene information collaboration. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 15279–15292. [CrossRef]
7. Duan, S.; Zhang, C.; Zheng, X.; Zhao, Y. EvoAR-SFNet: Evolutionary autoregressive spatial-frequency network for remote sensing pansharpening. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, *19*, 18491–18505. [CrossRef]
8. Yu, D.; Tang, Y.; Zhang, C.; Wang, W.; Yang, G.; Zheng, X.; Zhao, Y. IA2GNN: Imbalance-aware adaptive graph construction for multi-modal image fusion. *IEEE Trans. Multimed.* **2026**, *28*, 4747–4758. [CrossRef]
9. Cao, P.; Zhang, C.; Zheng, X.; Zhao, Y. Semantic-spatial guided reasoning for human-object interaction detection. *IEEE Signal Process. Lett.* **2026**, *33*, 551–555. [CrossRef]
10. Cao, P.; Lin, S.; Yang, Y.; Zhang, C. Dual-branch interactive guided network based on gradient prior for image super-resolution. *J. Vis. Commun. Image Represent.* **2025**, *113*, 104607. [CrossRef]

11. Cao, P.; Tang, Y.; Zhang, C.; Zheng, X.; Liang, C.; Wei, Y.; Zhao, Y. Visual relation diffusion for human-object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Honolulu, HI, USA, 2025; pp. 23551–23560.
12. Wang, R.; Zhang, C.; Zheng, X.; Lv, Y.; Zhao, Y. Joint defocus deblurring and superresolution learning network for autonomous driving. *IEEE Intell. Transp. Syst. Mag.* **2024**, *16*, 104–115. [CrossRef]
13. Zhang, C.; Zheng, X. Decomposing visual and semantic correlations for both fully supervised and few-shot image classification. *IEEE Trans. Artif. Intell.* **2024**, *5*, 1658–1668. [CrossRef]
14. Zhang, C.; Cao, P.; Bai, H.; Zheng, X. Semantic coherence guided multiview similarity for image classification with varied supervision. *IEEE Trans. Artif. Intell.* **2024**, *5*, 3072–3082. [CrossRef]
15. Zhang, G.; Tang, Y.; Zhang, C.; Zheng, X.; Zhao, Y. Entity dependency learning network with relation prediction for video visual relation detection. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 12425–12436. [CrossRef]
16. Xu, C.; Zhang, C.; Yang, Y.; Yang, H.; Bo, Y.; Li, D.; Zhang, R. Accelerate adversarial training with loss guided propagation for robust image classification. *Inf. Process. Manag.* **2023**, *60*, 103143. [CrossRef]
17. Zhang, C.; Bai, H.; Zhao, Y. Fine-grained image classification by class and image-specific decomposition with multiple views. *IEEE Trans. Multimed.* **2023**, *25*, 6756–6766. [CrossRef]
18. Li, W.; Hou, S.; Zhang, C.; Cao, C.; Liu, X.; Huang, Y.; Zhao, Y. An in-depth exploration of person re-identification and gait recognition in cloth-changing conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Vancouver, BC, Canada, 2023; pp. 13824–13833.
19. Zhang, C.; Liang, C.; Zhao, Y. Exemplar-based, semantic guided zero-shot visual recognition. *IEEE Trans. Image Process.* **2022**, *31*, 3056–3065. [CrossRef] [PubMed]
20. Zhang, C.; Wang, D.-H.; Li, H. Discriminative semantic region selection for fine-grained recognition. *J. Vis. Commun. Image Represent.* **2021**, *77*, 103084. [CrossRef]
21. Zhang, C.; Wang, D. Exploring the prediction consistency of multiple views for transductive visual recognition. *IEEE Signal Process. Lett.* **2021**, *28*, 668–672. [CrossRef]
22. Zhang, C.; Li, C.; Cheng, J. Few-shot visual classification using image pairs with binary transformation. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 2867–2871. [CrossRef]
23. Zhang, C.; Cheng, J.; Tian, Q. Multi-view image classification with visual, semantic and view consistency. *IEEE Trans. Image Process.* **2020**, *29*, 617–627. [CrossRef] [PubMed]
24. Zhang, C.; Cheng, J.; Tian, Q. Multiview, few-labeled object categorization by predicting labels with view consistency. *IEEE Trans. Cybern.* **2019**, *49*, 3834–3843. [CrossRef] [PubMed]
25. Zhang, C.; Cheng, J.; Tian, Q. Unsupervised and semi-supervised image classification with weak semantic consistency. *IEEE Trans. Multimed.* **2019**, *21*, 2482–2491. [CrossRef]
26. Yan, C.; Li, L.; Zhang, C.; Liu, B.; Zhang, Y.; Dai, Q. Cross-modality bridging and knowledge transferring for image understanding. *IEEE Trans. Multimed.* **2019**, *21*, 2675–2685. [CrossRef]
27. Zhang, C.; Cheng, J.; Tian, Q. Semantically modeling of object and context for categorization. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 1013–1024. [CrossRef] [PubMed]
28. Feng, W.D.; Yu, D.; Zhang, C.J.; Zheng, X.L. Perturbation response-based adaptive ensemble black-box adversarial attack algorithm. *Acta Autom. Sin.* **2025**, *51*, 1788–1799.
29. Fan, K.; Qian, K.; Xu, J.; Liu, B.; Gao, Y.; Zeng, N.; Fang, Z. Few-shot mask-guided controllable anomaly image generation via inpainting diffusion for surface defect detection. *J. Intell. Manuf.* **2026**. [CrossRef]
30. Yang, C.L.; Lv, D.H.; Tian, X.; Wang, C.Z.; Yang, P.H.; Zhang, Y. LEIS-Net: Lightweight and efficient industrial surface defect detection network. *J. Intell. Manuf.* **2026**, *37*, 2485–2502. [CrossRef]
31. Liu, J.; Wang, Y.; Cao, Y.Y.; Guo, C.P.; Shi, P.J.; Li, P. Unified spatial-frequency modeling and alignment for multi-scale small object detection. *Symmetry* **2025**, *17*, 242. [CrossRef]
32. Gao, W.F.; Yi, Y.X.; Huang, L.L.; Li, L.Y.; Li, H.; Xie, J. An efficient algorithm for small object detection in unmanned aerial vehicle images. *Control Decis.* **2025**, *40*, 2525–2533.
33. Sun, T.; Yang, S.Q.; Liu, H.Y.; Gu, J.; Deng, L.X.; Liu, L.D. MIS-YOLOv8: An improved algorithm for detecting small objects in UAV aerial photography based on YOLOv8. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 5020212. [CrossRef]
34. Wang, T.Y.; Ma, Z.J.; Yang, T.; Zou, S.L. PETNet: A YOLO-based prior enhanced transformer network for aerial image detection. *Neurocomputing* **2023**, *547*, 126384. [CrossRef]
35. Fan, Q.; Li, Y.; Deveci, M.; Zhong, K.; Kadry, S. LUD-YOLO: A novel lightweight object detection network for unmanned aerial vehicle. *Inf. Sci.* **2025**, *686*, 121366. [CrossRef]
36. Wang, R.P.; Lin, C.; Li, Y.J. RPLFDet: A lightweight small object detection network for UAV aerial images with rational preservation of low-level features. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 5013514. [CrossRef]
37. She, C.Y.; Han, F.J.; Fang, C.Y.; Duan, S.K.; Wang, L.D. Exploring Fourier prior and event collaboration for low-light image enhancement. *arXiv* **2025**, arXiv:2508.00308.

38. Gao, N.; Jiang, X.Y.; Zhang, X.H.; Deng, Y. Efficient frequency-domain image deraining with contrastive regularization. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Milan, Italy, 2024; pp. 240–257.
39. Hu, L.; Yuan, J.; Cheng, B.; Xu, Q. CSFPR-RTDETR: Real-time small object detection network for UAV images based on cross-spatial-frequency domain and position relation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5638219. [CrossRef]
40. Qin, D.F.; Leichner, C.; Delakis, M.; Fornoni, M.; Luo, S.; Yang, F.; Wang, W.; Banbury, C.; Ye, C.; Akin, B.; et al. MobileNetV4: Universal models for the mobile ecosystem. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Milan, Italy, 2024; pp. 78–96.
41. Ma, X.; Dai, X.Y.; Bai, Y.; Wang, Y.Z.; Fu, Y. Rewrite the stars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Seattle, WA, USA, 2024; pp. 5694–5703.
42. Chen, J.R.; Kao, S.H.; He, H.; Zhuo, W.P.; 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 (CVPR)*; IEEE: Vancouver, BC, Canada, 2023; pp. 12021–12031.
43. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Honolulu, HI, USA, 2017; pp. 1800–1807.
44. Lin, T.Y.; Dollár, P.; Girshick, R.B.; He, K.M.; Hariharan, B.; Belongie, S.J. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Honolulu, HI, USA, 2017; pp. 936–944.
45. Tan, M.X.; Pang, R.M.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Seattle, WA, USA, 2020; pp. 10778–10787.
46. Xu, X.Z.; Jiang, Y.Q.; Chen, W.H.; Huang, Y.L.; Zhang, Y.; Sun, X.Y. DAMO-YOLO: A report on real-time object detection design. *arXiv* **2022**, arXiv:2211.15444.
47. Yang, G.Y.; Lei, J.; Zhu, Z.K.; Cheng, S.Y.; Feng, Z.L.; Liang, R.H. AFPN: Asymptotic feature pyramid network for object detection. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*; IEEE: Honolulu, HI, USA, 2023; pp. 2184–2189.
48. Yang, Z.Q.; Guan, Q.; Zhao, K.E.; Yang, J.M.; Xu, X.L.; Long, H.X.; Tang, Y. Multi-branch auxiliary fusion YOLO with re-parameterization heterogeneous convolution for accurate object detection. In *Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*; Springer: Urumqi, China, 2024; pp. 492–505.
49. Wang, J.Q.; Chen, K.; Xu, R.; Liu, Z.W.; Loy, C.C.; Lin, D.H. CARAFE++: Unified content-aware reassembly of features. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 4674–4687. [CrossRef] [PubMed]
50. Qian, W.J. Edge detection in X-ray images of drill mast welds based on an improved Scharr operator. *Weld. Int.* **2024**, *38*, 574–581. [CrossRef]
51. He, F.X.; Liu, T.L.; Tao, D.C. Why ResNet works? Residuals generalize. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 5349–5362. [CrossRef] [PubMed]
52. Yang, C.L.; Lv, D.H.; Tian, X.; Wang, C.Z.; Yang, P.H.; Zhang, Y. A lightweight convolutional neural network for surface defect detection in strip steel. *Meas. Sci. Technol.* **2025**, *36*, 046001. [CrossRef]
53. Zhu, P.F.; Wen, L.Y.; Du, D.W.; Bian, X.; Fan, H.; Hu, Q.H.; Ling, H.B. Detection and tracking meet drones challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7380–7399. [CrossRef] [PubMed]
54. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: Seattle, WA, USA, 2020; pp. 1571–1580.
55. Liu, X.Y.; Peng, H.W.; Zheng, N.X.; Yang, Y.Q.; Hu, H.; Yuan, Y.X. EfficientViT: Memory efficient vision transformer with cascaded group attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Vancouver, BC, Canada, 2023; pp. 14420–14430.
56. Liu, W.Z.; Lu, H.; Fu, H.T.; Cao, Z.G. Learning to upsample by learning to sample. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Paris, France, 2023; pp. 6004–6014.
57. Li, C.Y.; Li, L.L.; Jiang, H.L.; 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.
58. 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 (CVPR)*; IEEE: Seattle, WA, USA, 2024; pp. 11769–11779.
59. 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 (CVPR)*; IEEE: Vancouver, BC, Canada, 2023; pp. 7464–7475.
60. Tang, L.F.; Zhang, H.; Xu, H.; Ma, J.Y. Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. *Inf. Fusion* **2023**, *99*, 101870. [CrossRef]

61. Zhang, Y.F.; Zhou, S.; Li, H.F. Depth information assisted collaborative mutual promotion network for single image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Seattle, WA, USA, 2024; pp. 2846–2855.
62. Khanam, R.; Hussain, M. What is YOLOv5: A deep look into the internal features of the popular object detector. *arXiv* **2024**, arXiv:2407.20892.
63. Ma, X.L.; Guo, F.M.; Niu, W.; Lin, X.; Tang, J.; Ma, K.S.; Ren, B.; Wang, Y.Z. PCONV: The missing but desirable sparsity in DNN weight pruning for real-time execution on mobile devices. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*; AAAI Press: New York, NY, USA, 2020; pp. 5117–5124.
64. Jin, L.J.; Ding, W.D.; Han, S.J.; Wang, J.Y. A real-time edge inference method for insulator contamination detection with YOLOv11-ssL. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 3534315. [CrossRef]
65. Li, X.; Wang, W.H.; Hu, X.L.; Li, J.; Tang, J.H.; Yang, J. Generalized focal loss V2: Learning reliable localization quality estimation for dense object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Nashville, TN, USA, 2021; pp. 11632–11641.
66. Wang, X.P.; Wan, S.T.; Li, Z.H.; Chen, X.X.; Zhang, B.L.; Wang, Y.L. ECL-Tear: Lightweight detection method for multiple types of belt tears. *Measurement* **2025**, *251*, 117269. [CrossRef]
67. Ren, S.Q.; He, K.M.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
68. Zhao, Y.A.; Lv, W.Y.; Xu, S.L.; Wei, J.M.; Wang, G.Z.; Dang, Q.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 (CVPR)*; IEEE: Seattle, WA, USA, 2024; pp. 16965–16974.
69. Huang, S.H.; Lu, Z.C.; Cun, X.D.; Yu, Y.J.; Zhou, X.; Shen, X. DEIM: DETR with improved matching for fast convergence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Nashville, TN, USA, 2025; pp. 15162–15171.
70. Ge, Z.; Liu, S.T.; Wang, F.; Li, Z.M.; Sun, J. YOLOX: Exceeding YOLO series in 2021. *arXiv* **2021**, arXiv:2107.08430.
71. Varghese, R.; Sambath, M. YOLOv8: A novel object detection algorithm with enhanced performance and robustness. In *Proceedings of the International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*; IEEE: Chennai, India, 2024; pp. 1–6.
72. Wang, A.; Chen, H.; Liu, L.H.; Chen, K.; Lin, Z.J.; Han, J.G.; Ding, G.G. YOLOv10: Real-time end-to-end object detection. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*; Curran Associates: Vancouver, BC, Canada, 2024; pp. 107984–108011.
73. Khanam, R.; Hussain, M. YOLOv11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725.
74. Tian, Y.J.; Ye, Q.X.; Doermann, D.S. YOLOv12: Attention-centric real-time object detectors. *arXiv* **2025**, arXiv:2502.12524.
75. Lei, M.Q.; Li, S.Q.; Wu, Y.H.; Hu, H.; Zhou, Y.; Zheng, X.H.; Ding, G.G.; Du, S.Y.; Wu, Z.Z.; Gao, Y. YOLOv13: Real-time object detection with hypergraph-enhanced adaptive visual perception. *arXiv* **2025**, arXiv:2506.17733.
76. Ding, J.; Xue, N.; Xia, G.S.; Bai, X.; Yang, W.; Yang, M.Y.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7778–7796. [CrossRef] [PubMed]

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.