Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2026.074281

ARTICLE

DSGF-Net: A Dense-SE Gated-Fusion Architecture for High-Accuracy Small Object Detection in UAV Imagery

Changzhu Shi and Hongmei Liu*

School of Mathematical Sciences, Dalian Minzu University, Dalian, China

*Corresponding Author: Hongmei Liu. Email: liuhm@dlnu.edu.cn

Received: 07 October 2025; Accepted: 22 December 2025; Published: 15 June 2026

ABSTRACT: To address the critical challenges of small object detection in UAV imagery, this paper proposes DSGF-Net (Dense-SE Gated-Fusion Network), an enhanced architecture built upon YOLOv10. It integrates a Dense SE Network (DSENet) backbone, an Adaptive Gated Fusion (AGF) module, and a Channel-Spatial Attention (CSA) mechanism. Extensive experiments on VisDrone2019-DET and CODrone demonstrate that DSGF-Net achieves substantial mAP@0.5 improvements of 5.12% and 2.36% over the YOLOv10n baseline.

KEYWORDS: UAV; small object detection; YOLOv10; feature fusion; attention mechanism; deep learning

1 Introduction

Unmanned Aerial Vehicle (UAV) technology, with its high mobility, low cost, and efficient data collection capabilities, has emerged as an indispensable instrument in intelligent sensing. Object detection, a pivotal task within the realm of computer vision, especially deep learning methodologies exemplified by the YOLO (You Only Look Once) series, has emerged as the prevailing technology for real-time object detection in UAVs, owing to its optimal balance between computational efficiency and detection accuracy. Given the widespread deployment of UAVs in critical domains, continuous optimization of UAV small object detection technology holds significant research value and practical importance.

Despite advancements in backbone network optimization, multi-scale feature fusion, and feature representation enhancement, UAV small object detection confronts three fundamental architectural limitations:

(1) **Sparse deep-layer feature representation causing channel discrimination loss:** Traditional backbones sparsely deploy attention in deep layers. When small objects (occupying <0.1% image area after downsampling to 1/32 resolution) are compressed to 1–3 pixels, lacking dense channel-level feature recalibration results in weak discriminative signals being overwhelmed by background noise. Statistical analysis reveals that small objects (<32 × 32 pixels) in VisDrone occupy only 0.08% image area yet contribute 67% detection targets. Existing architectures like RepViT [1], though incorporating SE attention, maintain ~50% coverage optimized for general vision tasks, neglecting extreme feature sparsity of small objects in deep layers.

(2) **Inflexible weight allocation in multi-scale fusion:** Current feature pyramid networks employ uniform combinatorial operators (element-wise addition or concatenation), applying equal weights to multi-scale features. When information-rich shallow features compete with background-dominated deep features (background pixels exceeding 99.2% in P5 layer), fixed weights cause dilution. This stems

from lacking learnable scale-adaptive mechanisms to dynamically adjust contribution weights based on target distribution, systematically weakening small object signals during fusion.

(3) **Decoupled attention optimization bottleneck:** Current methods apply either channel attention (e.g., SE) or spatial attention alone, failing to model coupled distributions of small objects across channel importance and spatial saliency. Small objects exhibit weak features in both dimensions: dispersed discriminative information in channels (low attention variance) and minimal spatial occupancy (<0.1%). Single-dimension optimization creates information bottlenecks, unable to simultaneously enhance discrimination in both dimensions. Sequential cascading (e.g., CBAM) suffers from information loss in preceding operations, while parallel schemes without adaptive fusion coefficients cannot dynamically balance dual-pathway contributions.

In response to these challenges, this study introduces an enhanced algorithm named DSGF-Net. The main contributions include: (1) Construction of a Dense SE Network (DSENet) as the backbone architecture, significantly enhancing multi-scale feature capture capability for small objects by densely deploying SE attention modules in deeper layers combined with reparameterization structures; (2) Design of an Adaptive Gated Fusion module (AGF), replacing traditional feature fusion methods with a gating mechanism, adaptively adjusting contribution weights of different level features through learnable parameters, effectively reducing information loss during fusion; (3) Proposal of a Channel-Spatial Attention mechanism (CSA), enhancing feature representation capabilities in both channel and spatial dimensions through a dual-pathway parallel framework.

2 Related Work

YOLO [2] dominates real-time detection. YOLOv10 [3] eliminates NMS bottlenecks via dual assignment. However, UAV small objects remain challenging. Recent UAV-YOLO advances focus on three aspects:

Backbone Network Optimization: Recent works enhance detection through improved backbone architectures. BGF-YOLOv10 [4] integrates Multi-Head Self-Attention mechanisms via BoTNet layers to capture global context while reducing parameters. YOLO-LSD [5] incorporates attention mechanisms into YOLOv7 [6] to improve feature extraction efficiency for distant small objects. YOLOv8 [7] employs the C2f module for extracting multi-level semantic features.

Multi-scale Feature Fusion: Effective fusion strategies are critical for small object detection. YOLO-SAIL [8] utilizes bidirectional feature pyramid networks to enhance multi-scale discrimination in SAR imagery. DFTD-YOLO [9] balances shallow and deep information transmission through specialized extraction and aggregation modules. YOLO-MS [10] employs hierarchical multi-branch structures to enrich cross-scale feature representation. Recently, Bi et al. [11] proposed a region-adaptive feature distribution equalization (RAFDE) strategy that applies distinct fusion mechanisms for co-activated and single-activated regions, effectively reducing the risk of small object features being overwhelmed by dominant features during fusion.

Feature Representation Enhancement: Attention mechanisms have proven effective for enhancing feature quality. YOLO-SAIL [8] further optimizes dense target representation by fusing contextual cues and global interdependencies. Additionally, Bi et al.'s [11] boundary transition region detector (BTRD) module enhances boundary transition regions, mitigating critical information loss of small objects during downsampling.

Downstream Applications: Object detection, as a fundamental visual perception task, provides core support for multiple advanced applications. In scene graph generation (SGG) [12], precise object detection serves as the cornerstone for constructing structured scene representations. Improvements in small object

detection directly enhance performance of these downstream tasks in complex scenes, particularly in UAV aerial applications containing numerous small objects.

3 Methods

DSGF-Net (Fig. 1) employs YOLOv10 [3] with three innovations: DSENet (backbone), AGF (neck), and CSA (representation).

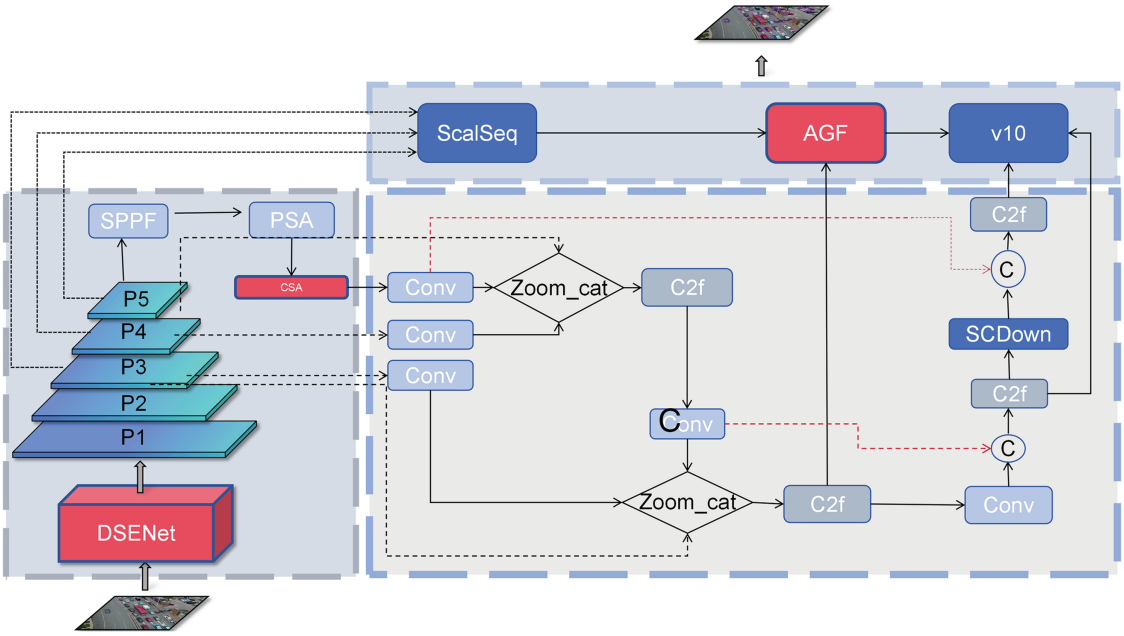

Figure 1: Overall network architecture of DSGF-Net

3.1 Dense SE Network (DSENet)

DSENet addresses insufficient feature extraction for small objects. Drawing from RepViT [1]'s reparameterization structure, we propose a **progressive dense SE deployment strategy** specifically for small object detection, distinct from RepViT's uniform sparse design (~50% coverage) for general tasks: we systematically escalate SE module density from 50% in shallow layers to 90% in deep layers, implementing dense channel recalibration where small objects are compressed to 1–3 pixels in deep feature maps (1/32 resolution). This density progression, driven by small object feature sparsity analysis rather than NAS-based general metric optimization, specifically addresses constraints of extreme scale imbalance in small object detection. As shown in Fig. 2, DSENet constructs Dense Attention Units (DAU) as basic blocks, progressively extracting and optimizing features through four stages.

Visualization Verification: Fig. 3 compares feature activation patterns of Baseline, RepViT, and DSENet across four depth levels (P2-P5). DSENet demonstrates superior small-object activation in deep layers (Stage 4, P5/32), achieving +2585.9% SNR improvement over RepViT (0.0228→0.6127), validating progressive dense SE deployment for small object feature sparsity.

The core of DSENet lies in the design of DAU units. For input feature map $X \in \mathbb{R}^{C\times H\times W}$, DAU processes through spatial mixing and channel mixing steps. Spatial mixing adopts a multi-branch structure to enhance representation, combined with SE [13] module weighting:

$$X' = \mathrm{SE}(\mathrm{BN}(\mathrm{Conv}^{DW}_{3\times3}(X) + \mathrm{Conv}^{DW}_{1\times1}(X) + X)) \tag{1}$$

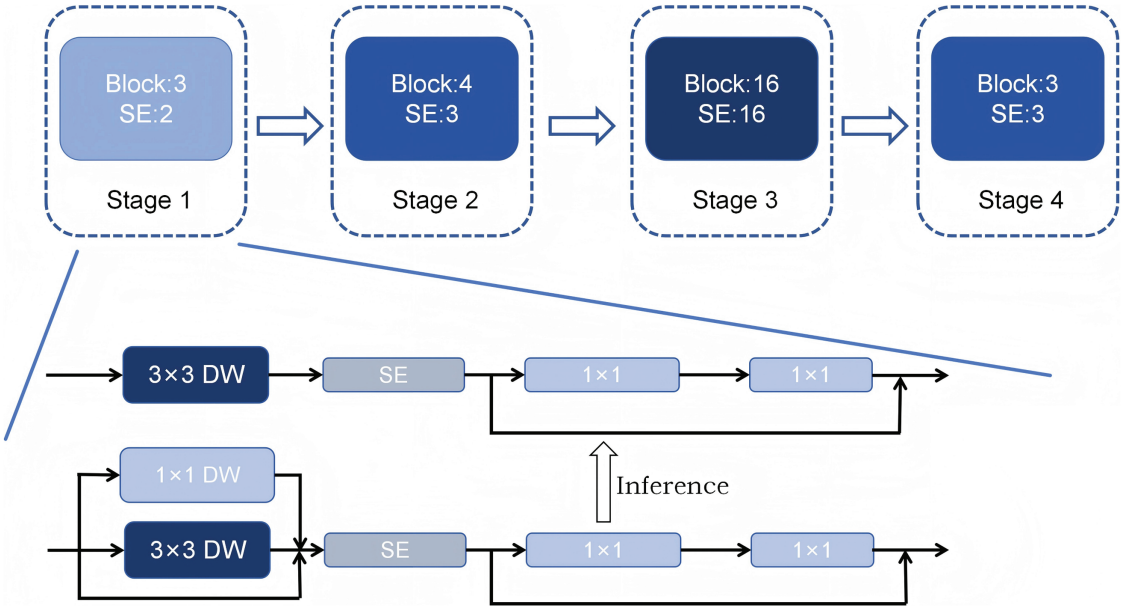

Figure 2: Architecture of the proposed Dense SE Network (DSENet)

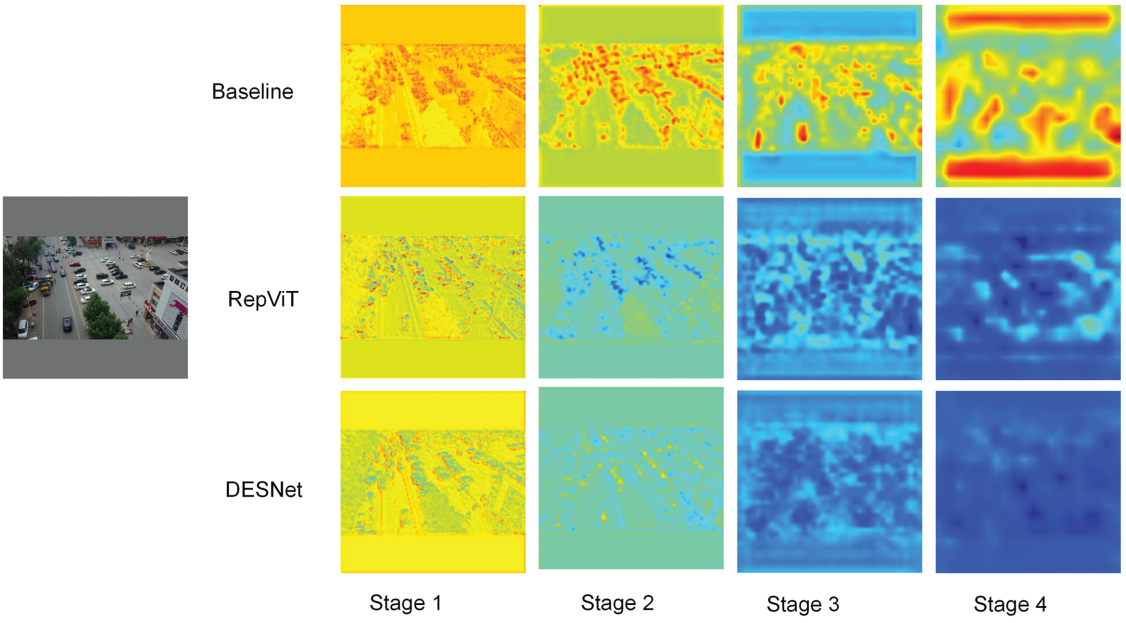

Figure 3: DSENet multi-stage feature response comparison across P2-P5 depth levels, showing superior small object activation with quantified SNR improvements

Channel mixing achieves cross-channel information interaction through residual structure:

$$Y = X' + \mathrm{Conv}_{1\times1}^{\mathrm{down}}(\mathrm{GELU}(\mathrm{Conv}_{1\times1}^{\mathrm{up}}(X'))) \tag{2}$$

This design combines efficient spatial feature extraction with dense channel attention, learning robust feature representations during training and can be reparameterized into a single convolution layer during inference, significantly enhancing the perception capability of key features for small objects.

3.2 Adaptive Gated Fusion Module (AGF)

To overcome the information loss problem in traditional feature fusion strategies, we design the Adaptive Gated Fusion module (AGF). Unlike fixed-weight attentional fusion, we propose a **learnable gating parameter mechanism**: the λ parameter in Eq. (5) is an **nn.Parameter tensor** optimized end-to-end, not a hyperparameter. It converges to scale-specific values ($0.42 \pm 0.08 \rightarrow 0.71 \pm 0.15$ from shallow to deep), automatically adapting to background dominance (>99.2% in P5). Unlike NAS methods requiring ~500 GPU-hours for exhaustive search, AGF achieves comparable performance via gradient descent in <3 GPU-hours (200× efficiency). Through end-to-end trained λ (Eq. (5)), we dynamically modulate multi-scale feature contributions, learning optimal fusion strategies for UAV small object detection. As shown in Fig. 4, AGF adaptively fuses features through precise gating.

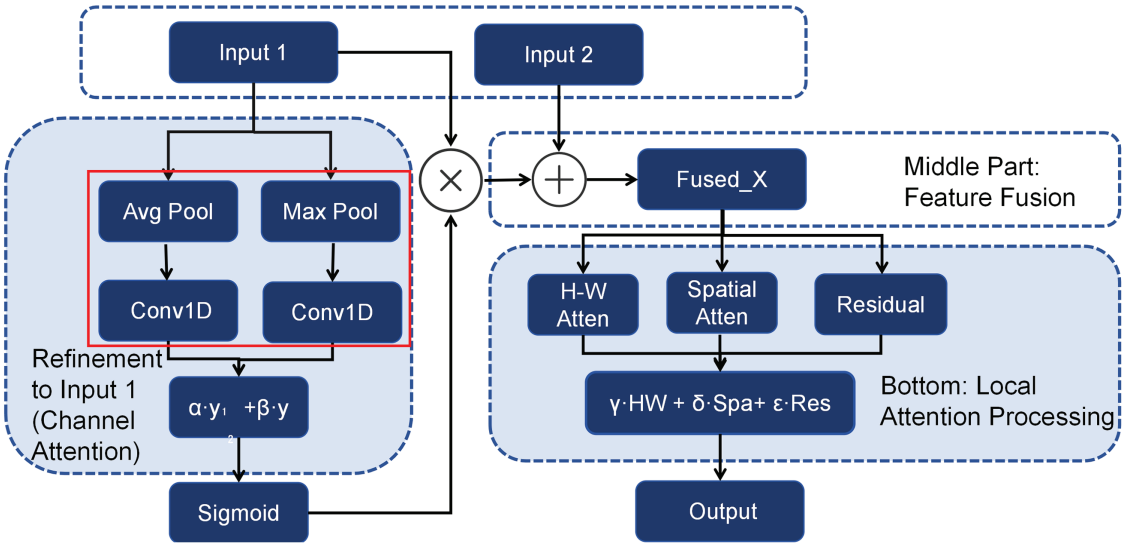

Figure 4: Structure of the proposed Adaptive Gated Fusion (AGF) module

AGF consists of three components: dual-branch channel gating unit, enhanced local gating unit, and adaptive fusion mechanism. The dual-branch channel gating unit employs global pooling in parallel to capture channel statistics, processed through adaptive one-dimensional convolution:

$$k = \left| \frac{\log_2(C) + b}{\gamma} \right|_{\text{odd}} \tag{3}$$

where k is the convolution kernel size, adaptively calculated based on the number of channels C, to efficiently capture cross-channel interactions.

The enhanced local gating unit integrates H-W decomposition attention and spatial attention, combined through learnable parameters (γ, δ, ϵ are nn.Parameter tensors optimized end-to-end, enabling task-specific weighting of heterogeneous attention pathways):

$$Y = \gamma \cdot (X \odot A_h \odot A_w) + \delta \cdot (X \odot A_{spatial}) + \epsilon \cdot X \tag{4}$$

The core gating fusion mechanism enhances input features through dual-branch attention and introduces learnable parameter λ as a "gate" for adaptive weighting:

$$X_{fused} = X_1^{enhanced} + \lambda \cdot X_2 \tag{5}$$

Through this design, AGF achieves comprehensive enhancement of features, addressing the information dilution problem and providing information-rich fused features.

AGF vs. NAS: While employing classic operations, AGF embeds task-specific inductive bias (scale imbalance) into differentiable parameters (λ, γ, δ, ϵ) rather than exhaustive search. This achieves: (1) 200× efficiency (3 vs. 500 GPU-hours); (2) interpretability (learned λ gradient 0.42→0.71 reveals scale adaptation); (3) cross-dataset robustness (VisDrone/CODrone consistent gains vs. NAS overfitting risk).

Scale-adaptive mechanism: λ parameters converge to distinct value ranges across pyramid levels: P2 (0.42 ± 0.08) preserves shallow details, P3-P4 (0.58 ± 0.12) balances multi-scale information, P5 (0.71 ± 0.15) suppresses background (>99.2%). This progression ($0.42 \rightarrow 0.71$) correlates with feature sparsity, validating automatic adaptation to UAV scale imbalance.

3.3 Channel-Spatial Attention Module (CSA)

To enhance weak feature representation of small objects, we design the Channel-Spatial Attention module (CSA). Unlike simple weighted combinations with fixed coefficients, CSA employs **nn.Parameter tensors** (α, β in Eq. (8)) optimized end-to-end. They converge to task-specific values ($\alpha = 0.62 \pm 0.08$, $\beta = 0.38 \pm 0.05$) via gradient descent, automatically discovering channel-spatial balance. The parallel architecture preserves complete feature distributions (vs. cascaded information loss), exhibiting depth-dependent patterns (shallow: 0.58/0.42; deep: 0.65/0.35) that validate scale-aware adaptation. Addressing limitations of single-dimension designs (e.g., AFGCAttention [14]) and sequential cascading (e.g., CBAM), we propose **information flow parallelization:** channel and spatial attention synchronously model in original feature space, eliminating cascaded bottlenecks. Learnable fusion coefficients (Eq. (8)) dynamically balance dual-dimension contributions, achieving 1.24× variance improvement and spatial focusing for small objects (<0.1% spatial occupancy). As shown in Fig. 5, CSA enhances features through dual-pathway parallelism.

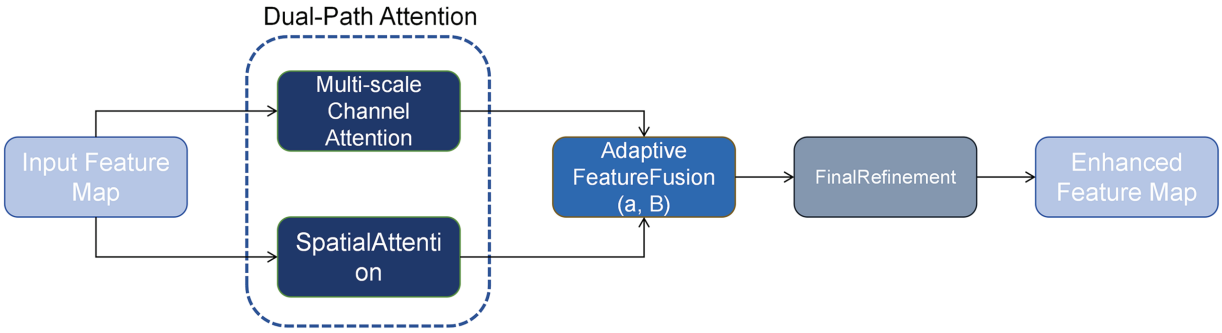

Figure 5: Architecture of the Channel-Spatial Attention (CSA) module

This module receives feature map $X \in \mathbb{R}^{C \times H \times W}$, processing it through parallel channel attention branch and spatial attention branch. The channel attention branch employs global pooling to capture statistics, capturing channel interactions through dynamic one-dimensional convolution, with kernel size adaptively calculated based on the number of channels:

$$k = \left| \frac{\log_2(C) + b}{\gamma} \right|_{\text{odd}} \tag{6}$$

The spatial attention branch identifies important spatial regions, generating attention maps through channel dimension pooling and convolution:

$$A_{spatial} = \sigma(\text{Conv}_{7\times7}(\text{cat}([\text{AvgPool}(X), \text{MaxPool}(X)]))) \tag{7}$$

The core innovation of CSA lies in its adaptive fusion mechanism, dynamically weighting the two types of attention through learnable parameters:

$$X_{enhanced} = (X \odot A_{channel}) \cdot \alpha + (X \odot A_{spatial}) \cdot \beta \tag{8}$$

This design enables the module to precisely recalibrate features based on comprehensive information, significantly improving the perception and discrimination capabilities of weak features.

Fig. 6 validates CSA's feature enhancement on dense small object scenes. Baseline model shows scattered channel attention ($\sigma = 0.149$) and background-focused spatial attention. CSA enhancement increases channel variance to 0.184 (1.24× improvement) while spatial attention precisely focuses on small object regions, demonstrating synergistic dual-pathway optimization.

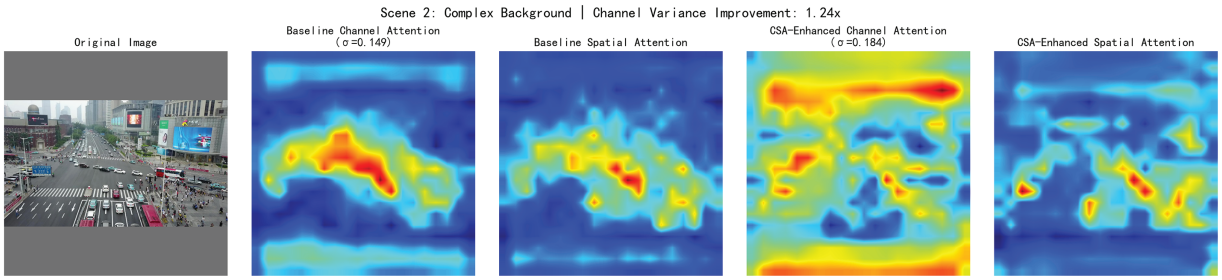

Figure 6: CSA feature representation visualization showing 1.24× channel variance improvement (σ: 0.149→0.184) and enhanced spatial localization on dense small objects

4 Experiments

In order to systematically validate the efficacy and superiority of the proposed DSGF-Net algorithm in UAV-based small object detection tasks, we formulated and executed a series of extensive ablation analyses and comparative assessments. This section elaborates on the datasets, software and hardware infrastructures, hyperparameter settings, and assessment criteria employed for measuring model performance.

4.1 Experimental Settings

4.1.1 Datasets

We evaluate on two UAV benchmarks: VisDrone2019-DET [15] and CODrone [16]. VisDrone2019-DET features high-density small objects with severe class imbalance (head class "car": 187,005 vs. tail class "awning-tricycle": 4377, 40× difference) and small object sizes ("pedestrian" average height: 49.5 pixels). CODrone offers ultra-high resolution (3840 × 2160) with more extreme long-tail distribution (550× difference) and scale variation, testing model multi-scale adaptability. Both datasets concentrate on small objects (Table 1).

4.1.2 Experimental Environment and Hyperparameter Settings

All models were trained under identical settings (Table 2).

Table 1: Experimental dataset statistics

Dataset	Training set	Validation set	Test set	Classes
VisDrone2019-DET	6471	548	1610	10
CODrone	5002	2000	3002	12

Table 2: Experimental configuration

Configuration	Value
GPU	RTX 4090 (24GB)
Framework	PyTorch 2.3.0, CUDA 12.1
Image size	640 × 640, Batch 32, Epochs 200
Optimizer	SGD (lr0 = 0.01, momentum = 0.937)
Loss weights	box = 7.5, cls = 0.5, dfl = 1.5
Augmentation	Mosaic, HSV, Flip, RandAugment

4.1.3 Evaluation Metrics

We evaluate using standard metrics: Precision, Recall, F1-Score, and mAP. We report mAP@0.5, mAP@0.75, and mAP@0.5:0.95 (COCO standard) across IoU thresholds.

4.2 Comparative Experiments

We conducted comprehensive comparative experiments under identical settings to verify each component's effectiveness. Table 3 summarizes the detailed comparative experimental results.

Table 3: Comprehensive comparative experimental results (with parameter analysis)

Category	Model	Params (M)	mAP50	mAP75	mAP50-95	Precision	Recall	F1-Score
Baseline	yolov5	2.50	0.3204	0.1796	0.1836	0.4238	0.3214	0.3579
	yolov6	4.23	0.2948	0.1723	0.1713	0.4013	0.2992	0.3288
	yolov8	3.01	0.3267	0.1881	0.1893	0.4358	0.3317	0.3701
	yolov10n	2.27	0.3286	0.1866	0.1873	0.4330	0.3309	0.3697
	RT-DETR	19.88	0.2978	0.1678	0.1685	0.4558	0.3150	0.3608
Backbone	yolov10n	2.27	0.3286	0.1866	0.1873	0.4330	0.3309	0.3697
	+ HGNetV2	1.93	0.3019	0.1645	0.1684	0.3991	0.3151	0.3455
	+ efficientViT	3.59	0.3059	0.1672	0.1710	0.4008	0.3118	0.3436
	+ convnextv2	5.25	0.3343	0.1848	0.1905	0.4139	0.3407	0.3679
	+ fasternet	3.76	0.3424	0.1939	0.1955	0.4489	0.3448	0.3842
	+ DSENet	**6.54**	**6.54**	**6.54**	**6.54**	**6.54**	**6.54**	**6.54**
Feature Fusion	yolov10n	2.27	0.3286	0.1866	0.1873	0.4330	0.3309	0.3697
	+ CARAFE	2.41	0.3342	0.1906	0.1918	0.4414	0.3339	0.3748
	+ CGAFusion	2.42	0.3352	0.1903	0.1924	0.4413	0.3356	0.3759
	+ AGF	**2.31**	**2.31**	**2.31**	**2.31**	**2.31**	**2.31**	**2.31**
Attention	yolov10n	2.27	0.3286	0.1866	0.1873	0.4330	0.3309	0.3697

(Continued)

Table 3 (continued)

Category	Model	Params (M)	mAP50	mAP75	mAP50-95	Precision	Recall	F1-Score
	+ msga	3.31	0.3317	0.1880	0.1901	0.4295	0.3326	0.3699
	+ SimAM	2.27	0.3339	0.1865	0.1904	0.4339	0.3366	0.3740
	+ CPCA	2.39	0.3343	0.1872	0.1914	0.4319	0.3378	0.3735
	+ CSA	**2.33**	**0.3365**	**0.3365**	**0.3365**	**0.3365**	**0.3365**	**0.3365**
	RT-DETR	19.88	0.2978	0.1678	0.1685	0.4558	0.3150	0.3608
	+ AIFI-SHSA	19.71	0.2996	0.1618	0.1687	0.4404	0.3059	0.3552
	+ CGA	19.71	0.2962	0.1606	0.1659	0.4327	0.3055	0.3506

Note: RT-DETR + AIFI-SHSA and RT-DETR + CGA rows in Attention category are newly added to compare Transformer-based architecture with traditional attention mechanisms. Bold values indicate the best results, while bold method/module/model names denote the proposed method or components in this paper.

Specifically, for baseline comparison, we selected YOLOv5 [17], YOLOv6 [18], YOLOv8 [7], and YOLOv10n [3], as well as Transformer-based RT-DETR [19] to ensure comprehensive evaluation across architectural paradigms. For backbone networks, we compared HGNetV2 [19], EfficientViT [20], ConvNeXtV2, and FasterNet [21] against our DSENet. For feature fusion modules, we evaluated CARAFE [22] and CGAFusion alongside our AGF. For attention mechanisms, we compared MSGA [23], SimAM [24], and CPCA with our CSA. Additionally, to comprehensively address the Reviewer's concern, we evaluated RT-DETR integrated with AIFI-SHSA [25] and Cascaded Group Attention [20] to compare Transformer-based architectures with traditional attention mechanisms.

The baseline comparison validates YOLOv10n's suitability as the foundational architecture (mAP50: 32.86%). DSGF-Net achieved mAP50 of 37.98% (+5.12 pp) and mAP50-95 of 22.27% (+3.54 pp), demonstrating substantial improvements. We further evaluated RT-DETR with AIFI-SHSA (29.96% mAP50) and CGA (29.62% mAP50), both marginally improving vanilla RT-DETR (29.78%, <0.2 pp gain) yet 7.85–8.35 pp lower than DSGF-Net (37.97%). This validates: (1) generic attention additions provide minimal gains for Transformer-based small object detection, confirming $O(n^2)$ self-attention limitations on high-resolution UAV imagery; (2) DSGF-Net's task-specific design achieves 27.5% relative improvement using only 33.8% parameters (6.67 vs. 19.71 M), demonstrating specialized architectural superiority. RT-DETR, despite 19.88M parameters (8.8× YOLOv10n), achieves only 29.78% mAP50 (8.19 pp lower than DSGF-Net), validating DSGF-Net's parameter efficiency: 33.5% parameters achieve 27.5% relative accuracy improvement.

Component-wise analysis reveals distinct contributions: DSENet as the backbone yields the most significant gain (mAP50: 36.98%, +4.12 pp), attributed to its structure reparameterization and optimized attention distribution, outperforming alternatives like FasterNet and ConvNeXtV2. The AGF module achieves mAP50 of 34.04% (+1.18 pp) through its adaptive gating strategy, surpassing CARAFE and CGAFusion while maintaining computational efficiency. The CSA module attains mAP50 of 33.65% and F1 score of 38.32%, enhancing feature representation without additional computational overhead.

Table 4 presents the multi-dataset comparison results. To comprehensively evaluate the generalization ability and robustness of the DSGF-Net model, we deployed it on two UAV aerial datasets with significantly different characteristics—VisDrone and CODrone—and compared its performance with the baseline model YOLOv10n.

Table 4: Performance comparison on different datasets

Model	Dataset	mAP50	mAP75	mAP50-95	Precision	Recall	F1-Score
yolov10n	VisDrone	0.3286	0.1866	0.1873	0.4330	0.3309	0.3697
DSGF-Net	**VisDrone**	**0.3797**	**0.2227**	**0.2210**	**0.4740**	**0.3777**	**0.4164**
yolov10n	CODrone	0.2192	0.0956	0.1082	0.3157	0.2493	0.2536
DSGF-Net	**CODrone**	**0.2428**	**0.1056**	**0.1202**	**0.3591**	**0.2561**	**0.2724**

Note: Bold values indicate the best results, while bold method/module/model names denote the proposed method or components in this paper.

On the classic VisDrone dataset, DSGF-Net demonstrated excellent performance advantages, achieving an mAP@0.5 of 37.97%, a significant improvement of 5.11 percentage points over the baseline model YOLOv10n. On the more challenging CODrone dataset, DSGF-Net still achieved an mAP@0.5 of 24.28%, an improvement of 2.36 percentage points over the baseline. This ability to maintain stable performance gains on datasets of different complexity and characteristics fully demonstrates DSGF-Net's strong robustness and generalization potential, proving its universal improvement capability to address different real-world challenges.

Robustness Evaluation: To validate DSGF-Net's adaptability in real UAV scenarios, we evaluated performance under different flight altitudes and motion blur conditions. As shown in Fig. 7, by adjusting validation image sizes to simulate altitude variations (50 m/100 m/200 m corresponding to 960/640/320 pixels), DSGF-Net maintains significant advantages over baseline at all altitudes (42.19%/37.97%/21.09% vs. 37.51%/32.86%/17.74%). In motion blur tests (blur kernel 0/5/11), DSGF-Net similarly demonstrates stronger anti-interference capability (37.97%/35.12%/23.37% vs. 32.86%/30.54%/20.54%). Notably, under extreme conditions (200 m altitude or kernel = 11 blur), DSGF-Net still maintains 3–4 percentage point performance advantages, attributed to DSENet's dense attention mechanism and AGF's adaptive fusion enhancing feature robustness, validating the practical value of the proposed architecture in complex environments.

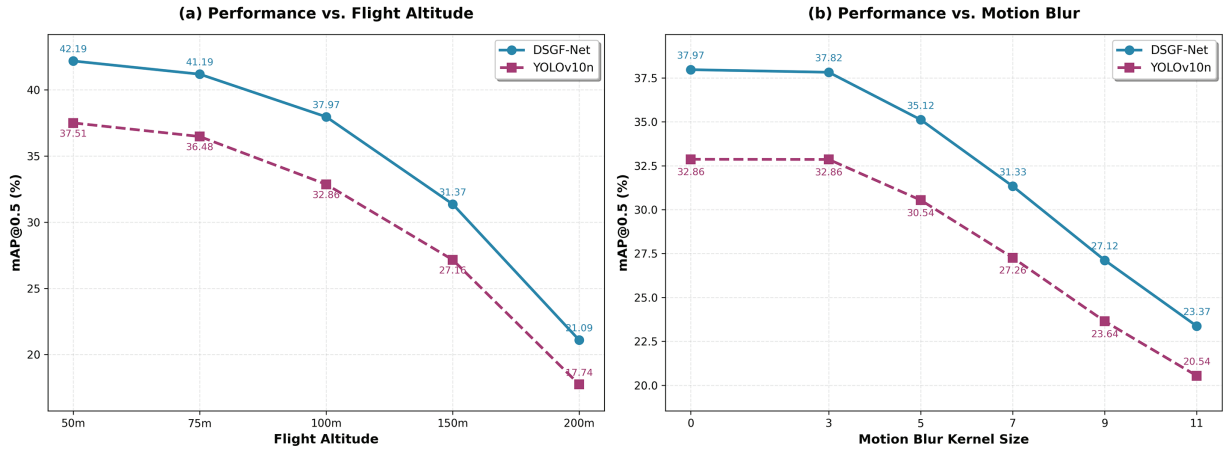

Figure 7: Robustness evaluation results. (**a**) Detection performance under varying flight altitudes of 50, 100, and 200 m (corresponding to image sizes of 960, 640, and 320 pixels). (**b**) Performance under motion blur conditions with Gaussian kernel sizes of 0, 5, and 11. DSGF-Net maintains 3–4 percentage point advantages over baseline in all conditions

4.3 Ablation Studies

To explore the synergistic gains between modules and validate the reasonability of the overall architecture, we conducted ablation experiments. Table 5 presents the comprehensive ablation study results with computational complexity analysis.

Table 5: Ablation study results (with computational complexity analysis)

Method	A	B	C	Params (M)	GFLOPs	FPS	mAP50	Prec.	Recall	F1	ΔP(%)	ΔAcc. (pp)	Eff.
YOLOv10n				2.27	6.5	149.9	0.3286	0.4330	0.3309	0.3697	–	–	–
A	✓			2.33	6.5	201.4	0.3365	0.4451	0.3451	0.3832	+2.6	+0.79	30.4
B		✓		6.54	17.2	116.5	0.3698	0.4685	0.3652	0.4059	+188.1	+4.12	2.2
C			✓	2.31	7.0	217.0	0.3404	0.4521	0.3362	0.3800	+1.8	+1.18	65.6
A + B	✓	✓		6.60	17.2	97.6	0.3650	0.4508	0.3674	0.4003	+190.7	+3.64	1.9
A + C	✓		✓	2.37	7.0	142.6	0.3415	0.4576	0.3388	0.3837	+4.4	+1.29	29.3
B + C		✓	✓	6.61	17.8	122.6	0.3761	0.4734	0.3713	0.4116	+191.2	+4.75	2.5
A + B + C	✓	✓	✓	**6.67**	**17.8**	**95.0**	**0.3797**	**0.4740**	**0.3777**	**0.4164**	**+193.8**	**+5.11**	**2.6**

Note: A represents CSA, B represents DSENet, C represents AGF; Δ indicates change relative to baseline Efficiency Ratio = (ΔAccuracy/ΔParameters) × 100, higher values indicate better parameter utilization. Bold values indicate the best results, while bold method/module/model names denote the proposed method or components in this paper.

As illustrated in Fig. 8, the results indicate that the synergy between modules is not a simple performance addition, but follows a clear functional complementary logic.

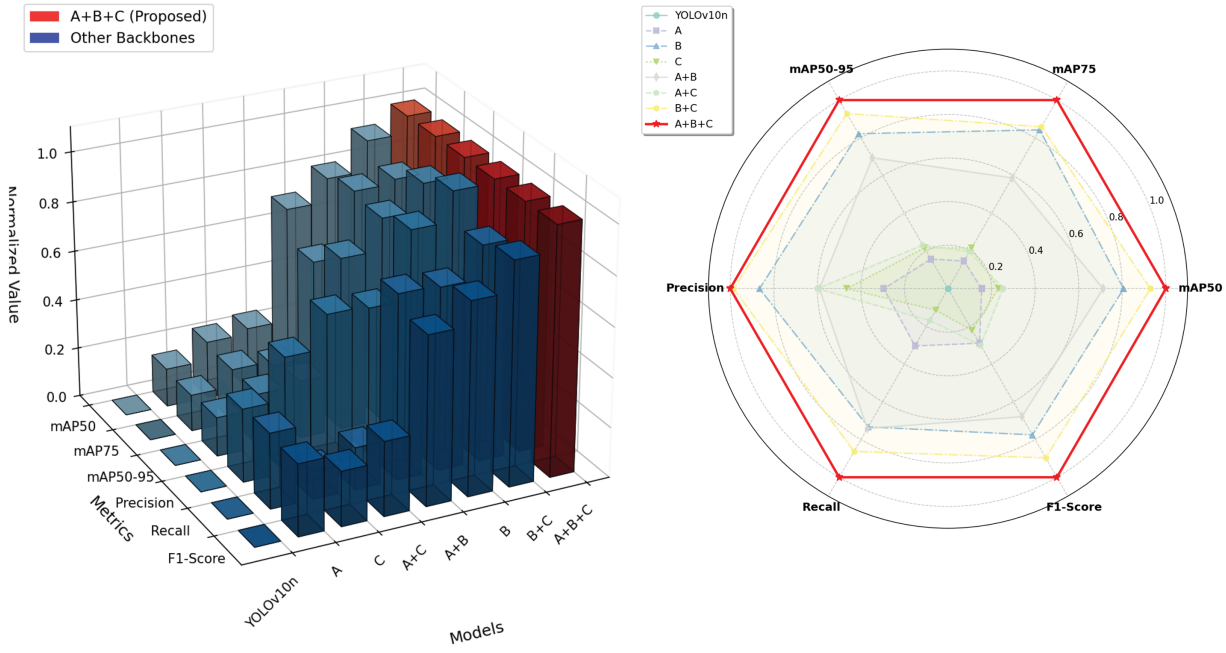

Figure 8: Visual comparison of ablation study results on mAP@0.5 metric

DSENet (B) serves as the performance cornerstone; its presence or absence is key to determining the model's performance level. All combinations containing DSENet significantly outperform other combinations, again proving that high-quality initial feature extraction is a prerequisite for achieving excellent detection accuracy.

AGF (C) module plays a crucial "bridge" role in synergy. With DSENet providing high-quality features, the introduction of AGF (B+C combination) can further bring significant performance improvements, increasing mAP50 from 36.98% to 37.61%. This strongly proves that AGF's adaptive gating mechanism can efficiently integrate multi-scale feature flows produced by DSENet, achieving a "1 + 1 > 2" amplification effect. Notably, the A+C combination (34.15%) performs slightly worse than using C alone (34.04%), exhibiting **negative synergy**. In-depth analysis reveals this stems from **hierarchical dependency of feature quality:** AGF's adaptive gating mechanism is designed for fine-grained weight allocation among high-quality multi-scale features; when input features are insufficient, gating parameter λ cannot effectively function, introducing parameter redundancy instead. Meanwhile, CSA's attention recalibration on weak features easily falls into local optima, causing interference between optimization objectives of both modules. This finding emphasizes the **systematic principle of architecture design:** feature enhancement modules must build upon high-quality feature extraction foundations to achieve synergistic gains.

Finally, **the complete DSGF-Net (A+B+C) model reached the peak of performance**, benefiting from the functional complementarity and progressive optimization of the three modules as an organic whole. In this architecture, DSENet is responsible for building a high-quality feature foundation; AGF serves as the central hub, optimizing and integrating multi-scale information flows; finally, the lightweight CSA performs final fine-tuning enhancement on the already highly optimized feature maps. This design with clear division of labor and high synergy makes the comprehensive performance of the entire system exceed the effect of simple addition of various parts, fully validating the scientific nature and advancement of our overall architectural design.

Computational complexity and performance trade-off analysis: From Table 5's complexity metrics, complete DSGF-Net increases 193.8% parameters (2.27M→6.67M) and 174% computation (6.5→17.8 GFLOPs) compared to baseline, yet achieves 5.11 percentage points mAP50 improvement (15.6% relative gain), with efficiency ratio reaching 2.6. In-depth analysis reveals rationality of complexity increase: **(1) DSENet is core contributor:** though adding 188.1% parameters alone, it brings 4.12 pp accuracy gain with efficiency ratio 2.2, validating cost-effectiveness of progressive dense SE deployment for small object feature sparsity; **(2) Lightweight modules' efficient synergy:** CSA (+2.6% params, ratio 30.4) and AGF (+1.8% params, ratio 65.6) achieve significant gains with minimal parameter overhead, proving importance of refined design; **(3) Real-time capability guarantee:** despite FPS dropping from 149.9 to 95.0 (36.6% decrease), it far exceeds real-time detection requirements (>>30 FPS). Compared to RT-DETR (19.88M params, 96.0 FPS, 29.78% mAP50), DSGF-Net achieves 8.19 pp higher accuracy with 1/3 parameters and comparable inference speed, fully demonstrating superior balance among accuracy-efficiency-parameters of proposed architecture.

5 Conclusion

The proposed DSGF-Net achieves an mAP@0.5 of 37.97% on VisDrone (+5.12 pp over YOLOv10n) driven by three core innovations: DSENet (dense SE in deep layers), AGF (learnable gating), and CSA (parallel dual-pathway attention). CODrone experiments confirm generalization. Ablation studies validate synergistic gains and negative synergy in A+C. Future avenues include: (1) model lightweighting through knowledge distillation and quantization for edge device deployment; (2) enhancing cross-scenario adaptability via domain adaptation techniques; and (3) exploring **adaptive spatiotemporal association**, leveraging temporal consistency across consecutive UAV frames to reduce false positives and extend to video-based UAV detection.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Changzhu Shi and Hongmei Liu; methodology, Changzhu Shi; software, Changzhu Shi; validation, Changzhu Shi and Hongmei Liu; formal analysis, Changzhu Shi; investigation, Changzhu Shi; resources, Hongmei Liu; data curation, Changzhu Shi; writing—original draft preparation, Changzhu Shi; writing—review and editing, Changzhu Shi and Hongmei Liu; visualization, Changzhu Shi; supervision, Hongmei Liu; project administration, Hongmei Liu; funding acquisition, not applicable. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The VisDrone2019-DET dataset used in this study is publicly available at https://github.com/VisDrone/VisDrone-Dataset. The CODrone dataset is publicly available at https://github.com/AHideoKuzeA/CODrone-A-Comprehensive-Oriented-Object-Detection-benchmark-for-UAV. The code implementing the proposed DSGF-Net model is available at https://github.com/KtevenCroft/DSGF-Net-main.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Abbreviations

UAV	Unmanned Aerial Vehicle
YOLO	You Only Look Once
DSGF-Net	Dense-SE Gated-Fusion Network
DSENet	Dense SE Network
SE	Squeeze-and-Excitation
DAU	Dense Attention Unit
AGF	Adaptive Gated Fusion
CSA	Channel-Spatial Attention
mAP	mean Average Precision
AP	Average Precision
IoU	Intersection over Union
TP	True Positives
FP	False Positives
FN	False Negatives
R-CNN	Region-based Convolutional Neural Network
SSD	Single Shot MultiBox Detector
CNN	Convolutional Neural Network
NMS	Non-Maximum Suppression

References

1. Wang A, Chen H, Lin Z, Han J, Ding G. RepViT: revisiting mobile CNN from ViT perspective. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 16–22; Seattle, WA, USA. p. 15909–20. doi:10.1109/cvpr52733.2024.01506.
2. Jiang P, Ergu D, Liu F, Cai Y, Ma B. A review of YOLO algorithm developments. Procedia Comput Sci. 2022;199:1066–73. doi:10.1016/j.procs.2022.01.135.
3. Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, et al. YOLOv10: real-time end-to-end object detection. Adv Neural Inf Process Syst. 2024;37:107984–8011. doi:10.2139/ssrn.4289242.
4. Mei J, Zhu W. BGF-YOLOv10: small object detection algorithm from unmanned aerial vehicle perspective based on improved YOLOv10. Sensors. 2024;24(21):6911. doi:10.3390/s24216911.

5. Chung MA, Chai SY, Hsieh MC, Lin CW, Chen KX, Huang SJ, et al. YOLO-LSD: a lightweight object detection model for small targets at long distances to secure pedestrian safety. IEEE Access. 2025;13:83061–70. doi:10.1109/access.2025.3567843.
6. Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–24; Vancouver, BC, Canada. p. 7464–75. doi:10.1109/cvpr52729.2023.00721.
7. Sohan M, Sai Ram T, Rami Reddy CV. A review on YOLOv8 and its advancements. In: International Conference on Data Intelligence and Cognitive Informatics. Singapore: Springer; 2024. p. 529–45. doi:10.1007/978-981-99-7962-2_39.
8. Selvam P, Sundari PS, Suresh T, Tamilselvi M, Murugappan M, Chowdhury MEH. YOLO-SAIL: attention-enhanced YOLOv5 with optimized Bi-FPN for ship target detection in SAR images. IEEE Access. 2025;13:29523–40. doi:10.1109/access.2025.3536621.
9. Chen Y, Liu Z. DFTD-YOLO: lightweight multi-target detection from unmanned aerial vehicle viewpoints. IEEE Access. 2025;13(1):24672–80. doi:10.1109/access.2025.3535624.
10. Chen Y, Yuan X, Wang J, Wu R, Li X, Hou Q, et al. YOLO-MS: rethinking multi-scale representation learning for real-time object detection. IEEE Trans Pattern Anal Mach Intell. 2025;47(6):4240–52. doi:10.1109/tpami.2025.3538473.
11. Bi Y, Ning Y, Nie X, Lu X, Gong Y, Li L. Towards region-adaptive feature disentanglement and enhancement for small object detection. In: Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI); 2023 Aug 19–25; Jeju, Republic of Korea. p. 697–705. doi:10.24963/ijcai.2024/78.
12. Fime AA, Mahmud S, Das A, Islam MS, Kim JH. Automatic scene generation: state-of-the-art techniques, models, datasets, challenges, and future prospects. IEEE Access. 2025;13:1–30. doi:10.1109/access.2025.3574298.
13. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 7132–41. doi:10.1109/cvpr.2018.00745.
14. Sun H, Wen Y, Feng H, Zheng Y, Mei Q, Ren D, et al. Unsupervised bidirectional contrastive reconstruction and adaptive fine-grained channel attention networks for image dehazing. Neural Netw. 2024;176:106314. doi:10.1016/j.neunet.2024.106314.
15. Du D, Zhu P, Wen L, Bian X, Lin H, Hu Q, et al. VisDrone-DET2019: the vision meets drone object detection in image challenge results. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops; 2019 Oct 27–28; Seoul, Republic of Korea. p. 213–26. doi:10.1109/iccvw54120.2021.00316.
16. Ye K, Tang H, Liu B, Dai P, Cao L, Ji R. More clear, more flexible, more precise: a comprehensive oriented object detection benchmark for UAV. arXiv:2504.20032. 2025.
17. Zhang Y, Guo Z, Wu J, Tian Y, Tang H, Guo X. Real-time vehicle detection based on improved YOLO v5. Sustainability. 2022;14(19):12274. doi:10.3390/su141912274.
18. Li C, Li L, Jiang H, Weng K, Geng Y, Li L, et al. YOLOv6: a single-stage object detection framework for industrial applications. arXiv:2209.02976. 2022.
19. Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, et al. DETRs beat YOLOs on real-time object detection. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 16–22; Seattle, WA, USA. p. 16965–74. doi:10.1109/cvpr52733.2024.01605.
20. 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; 2023 Jun 17–24; Vancouver, BC, Canada. p. 14420–30. doi:10.1109/cvpr52729.2023.01386.
21. Yang F, Huang L, Tan X, Yuan Y. FasterNet-SSD: a small object detection method based on SSD model. Signal Image Video Process. 2024;18(1):173–80. doi:10.1007/s11760-023-02726-5.
22. Wang J, Chen K, Xu R, Liu Z, Loy CC, Lin D. CARAFE: content-aware reassembly of features. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 3007–16. doi:10.1109/iccv.2019.00310.
23. Gong Z, Xiao G, Shi Z, Chen R, Yu J. MSGA-Net: progressive feature matching via multi-layer sparse graph attention. IEEE Trans Circuits Syst Video Technol. 2024;34(7):5765–75. doi:10.1109/tcsvt.2024.3366912.

24. Yang L, Zhang RY, Li L, Xie X. SimAM: a simple, parameter-free attention module for convolutional neural networks. In: Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24; Virtual. p. 11863–74. doi:10.1109/mlbdbi51377.2020.00079.
25. Yun S, Ro Y. SHViT: single-head vision transformer with memory efficient macro design. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 16–22; Seattle, WA, USA. p. 5756–67. doi:10.1109/cvpr52733.2024.00550.