Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

ECSFNet-DET: An Anti-Drone Object Detection Algorithm Based on Visual State Space Modeling and Spectral Enhancement

Yonghang Yu
Zhejiang University of Technology
Yu Zhang
Aerospace Information Research Institute (Qilu Branch)
Yuanchao Mu
Aerospace Information Research Institute (Qilu Branch)
Shibo Liu
Zhejiang University of Technology
Fan Gao
Zhejiang University of Technology
Chengming Sun
sunchengming2008@163.com

Aerospace Information Research Institute

Research Article

Keywords: anti-drone detection, small object detection, visual state space model, multi-scale feature fusion, spectral adaptive modulation, polarized self-attention, convolutional gated linear unit, YOLOv11n

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10429850/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

ECSFNet-DET: An Anti-Drone Object Detection Algorithm Based on Visual State Space Modeling and Spectral Enhancement

Yonghang Yu[1], Yu Zhang[2], Yuanchao Mu[2], Shibo Liu[1], Fan Gao[1*], Chengming Sun[3*]

[1]College of Physics and Optoelectronics, Zhejiang University of Technology, Hangzhou, Zhejiang, China.
[2]Qilu Aerospace Information Research Institute, Chinese Academy of Sciences, Jinan, Shandong, China.
[3*]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

*Corresponding author(s). E-mail(s): gaofan@zjut.edu.cn; sunchengming2008@163.com;
Contributing authors: 78130605@qq.com; zhangyu@aircas.ac.cn; myc_nt_98@163.com; 221125090104@zjut.edu.cn;

Abstract

To address the core challenges of anti-drone detection in complex low-altitude environments, including sparse target pixels, strong background clutter interference, and insufficient utilization of frequency-domain information by existing methods, this paper proposes ECSFNet-DET, a novel anti-drone object detection algorithm based on visual state space modeling and spectral enhancement. Built upon YOLOv11n, ECSFNet-DET introduces three synergistic improvements across the backbone, feature fusion neck, and deep semantic enhancement stages. First, a Dual-branch Visual State-space Feature Encoder (DVSFE) is proposed, which deeply integrates a lightweight state space sequence modeling mechanism with a depth-wise convolutional activation gated feed-forward unit within a dual-branch architecture, achieving joint local detail perception and global long-range context modeling at near-linear computational complexity. Second, an Adaptive Bidirectional Multi-Scale Semantic Fusion Neck (ABMSFN) is designed to construct a bidirectional multi-scale feature flow pathway through global context aggregation, recursive asymmetric strip channel-spatial calibration, dynamic cross-scale semantic injection, and gated adaptive fusion, effectively mitigating

the semantic gap across scales. Third, a Polarized Attention Multi-Granularity Spectral Enhancement module (PAMGSE) is introduced, which embeds a multi-granularity hybrid convolutional enhancement network and a spectral adaptive modulation gated feed-forward network into a polarized self-attention backbone via learnable scalar-gated residuals, establishing a hierarchical joint spatial-frequency feature representation. Furthermore, a self-constructed anti-drone detection dataset SUAV-Bird is built, comprising 7,874 images across drone and bird categories, with over 86% of target instances below 16×16 pixels. Experimental results demonstrate that ECSFNet-DET achieves improvements of 4.1%, 3.5%, 3.8%, and 2.0% over the YOLOv11n baseline in precision, recall, mAP50, and mAP50-95, respectively, attaining an optimal accuracy-efficiency trade-off at 8.8 GFLOPs and 3.28M parameters, with robust cross-domain generalization validated on the multi-source Mixdata benchmark.

Keywords: anti-drone detection, small object detection, visual state space model, multi-scale feature fusion, spectral adaptive modulation, polarized self-attention, convolutional gated linear unit, YOLOv11n

1 Introduction

In recent years, low-altitude unmanned aerial vehicles (UAVs) have found widespread applications in aerial photography and mapping, logistics inspection, emergency rescue, agricultural crop protection, and public safety, driving the development of low-altitude intelligent equipment and related industries [1]. However, the rapid proliferation of consumer-grade drones has simultaneously introduced new security risks. UAVs are characterized by small size, low cost, high maneuverability, low-altitude flight capability, and covert operability. Once exploited for illegal intrusion, close-range reconnaissance, transportation of prohibited goods, or disruption of sensitive areas, they can pose severe threats to airports, border zones, energy infrastructure, large public venues, and military installations [2]. Accordingly, achieving rapid, accurate, and stable detection of UAV targets within complex low-altitude environments has become a pivotal challenge in counter-UAV systems.

Regarding conventional visual detection approaches, early counter-UAV research predominantly relied on motion analysis techniques such as frame differencing, background subtraction, and optical flow to localize moving targets by extracting inter-frame differences. While these methods exhibit certain feasibility under static backgrounds, they prove highly sensitive to dynamic backgrounds, illumination variation, and camera jitter, yielding elevated false-alarm rates in complex low-altitude scenes. Subsequently, approaches combining hand-crafted feature descriptors such as HOG and SIFT with SVM or random forest classifiers gradually emerged, achieving reasonable results in controlled environments. Nevertheless, UAV targets universally present small size, weak texture, and low contrast; hand-crafted features thus have limited representational capacity and poor generalization under complex backgrounds and illumination changes, failing to satisfy practical deployment requirements [3].

With the rapid advance of deep learning, convolutional neural network (CNN)-based object detection methods have gradually dominated counter-UAV visual perception. Two-stage detectors exemplified by Faster R-CNN achieve relatively high accuracy through region proposal networks, yet their inference latency cannot satisfy real-time counter-UAV detection requirements. Yasmeen et al. [4] compared single-stage and two-stage detectors for counter-UAV tasks in a survey study, noting that YOLO-series single-stage methods unify classification and localization into a single forward pass, achieving a superior speed-accuracy balance, and have become the mainstream technical route for counter-UAV visual detection. Kassab et al. [5] further confirmed through comparative experiments that YOLO-series single-stage methods significantly outperform traditional machine learning approaches in both detection accuracy and real-time capability for drone detection, establishing a methodological foundation for subsequent improvements.

Substantial progress has been made in recent UAV detection research. Numerous studies have improved YOLO models by introducing shallow detection heads, attention mechanisms, lightweight convolutions, multi-scale feature fusion, and small-target enhancement modules to boost UAV small-target detection. Du et al. [6] proposed a method combining background subtraction preprocessing with a lightweight F-YOLOv7 network, effectively improving real-time detection accuracy for low-slow-small UAVs. Gao et al. [7] proposed a lightweight counter-UAV detection approach based on improved YOLOv11, improving small-target detection accuracy and reducing parameter count through HWD downsampling and CCFM structure optimization. Munir et al. [8] proposed YOLO-RAW, addressing real-time UAV detection under complex backgrounds and weather interference by enhancing multi-scale UAV detection capability and improving robustness under adverse weather. Zhou et al. [9] proposed LMWP-YOLO, which integrates multi-dimensional collaborative attention mechanisms and multi-scale feature fusion modules, achieving a 22.07% mAP improvement over YOLOv11n while reducing parameter count by 52.51%, with strong cross-dataset generalization. Zheng et al. [10] proposed YOLO-DD, a lightweight counter-UAV framework based on YOLOv11n that employs ghost dynamic convolution modules and selective boundary aggregation modules to improve multi-scale target feature expression, surpassing most YOLO variants at 87.72 FPS. These findings confirm that object detection methods possess mature frameworks and strong real-time capability, well-suited to counter-UAV visual perception tasks. Accordingly, this paper selects YOLOv11n as the base framework and pursues further improvements targeting feature attenuation, insufficient multi-scale fusion, and complex background interference in small UAV target detection.

Despite this progress, counter-UAV small-target detection still confronts three core challenges. First, UAV targets exhibit sparse pixels and weak textures; repeated downsampling readily causes feature loss, while interference sources such as birds and branches in complex low-altitude backgrounds are easily confused with targets, producing simultaneous missed detections and false alarms. Second, the limited local receptive fields of convolutional architectures, combined with the absence of global semantic guidance in traditional feature pyramids, hinder effective discrimination of multi-scale weak small targets. Third, existing methods predominantly focus on

spatial-domain feature enhancement with insufficient frequency-domain information utilization, leading to weaker robustness under texture interference and imaging noise.

To address the above challenges, this paper proposes ECSFNet-DET, an anti-drone object detection algorithm based on visual state space modeling and spectral enhancement. Built upon YOLOv11n, the method performs collaborative optimization at three levels: the backbone network, the feature fusion neck, and the deep enhancement module. The main contributions are as follows:

1. A Dual-branch Visual State-space Feature Encoder (DVSFE) is proposed. This module combines a lightweight state space modeling mechanism with a convolutional gated feed-forward unit, simultaneously enhancing local spatial detail perception and global long-range dependency modeling at low computational overhead, thereby improving the model's feature discrimination capability for small-size, low-contrast UAV targets.
2. An Adaptive Bidirectional Multi-Scale Semantic Fusion Neck (ABMSFN) is proposed. This structure constructs a bidirectional feature flow pathway from global to local and from high-level to low-level through multi-level global semantic aggregation, recursive feature calibration, dynamic cross-scale semantic injection, and semantic gated adaptive fusion, effectively alleviating the semantic gap in traditional feature pyramids and improving detection accuracy for UAV targets at different scales.
3. A Polarized Attention Multi-Granularity Spectral Enhancement module (PAMGSE) is proposed. This module introduces multi-scale hybrid convolutional enhancement and spectral adaptive modulation mechanisms into a polarized self-attention backbone, enabling the network to simultaneously perform spatial local texture modeling, frequency-domain global structural perception, and channel attention expression, thereby enhancing detection robustness under complex backgrounds, noise interference, and target morphology variation.

2 Related Work

2.1 Anti-Drone Detection Datasets

2.1.1 Survey of Existing Public Datasets and Their Limitations

High-quality datasets are the core foundation driving counter-UAV detection research. The DUT Anti-UAV dataset [11] constructed by Dalian University of Technology is currently the most widely influential single-modal benchmark; its detection subset contains 10,000 annotated images covering 35 UAV types. However, this dataset defines only a single "UAV" category without including visually similar interference targets such as birds, leading to insufficient anti-interference discrimination capability for detectors trained on it during practical deployment. Furthermore, image resolution spans from 240×160 to 5616×3744 pixels, lacking a unified high-resolution standard, which to some extent restricts evaluation for high-resolution systems.

The Det-Fly dataset [12] was acquired via drone-to-drone photography, enriching target appearance variation from multiple viewpoints, but likewise contains only a

single target category, has limited scene diversity, and lacks coverage of typical engineering challenge conditions such as strong glare and camera defocus. In recent years,



Fig. 1 Scene sample illustrations from the SUAV-Bird dataset.

several works have attempted to address these limitations. Sun et al. [13] proposed AirBirds—the first large-scale bird detection dataset for airport bird-strike prevention and control—comprising 118,312 sequential images and 409,967 finely annotated bird bounding boxes covering multiple illumination conditions and 13 meteorological scenarios. However, AirBirds contains no drone samples and cannot be directly applied to joint UAV-bird counter-UAV detection tasks. SynDroneVision (WACV 2025) [14] demonstrated that synthetic and real-world data have good complementary potential, but the distribution shift between synthetic and real domains remains a fundamentally difficult challenge to eliminate.

In summary, existing public counter-UAV datasets share the following common limitations: (1) single target category, lacking visually confusable aerial targets; (2) insufficient scene and background diversity; (3) limited coverage of extreme illumination and adverse weather conditions; and (4) absence of a joint UAV-bird detection benchmark supporting anti-interference performance evaluation. These limitations collectively motivate this paper's independent data collection and dataset construction.

2.1.2 Self-Constructed Dataset in This Paper

To address the above limitations, our research group constructed a new counter-UAV detection dataset through independent field collection. Image sequences were systematically acquired across 32 background scenes, covering typical low-altitude environments including urban buildings, mountainous forests, lake surfaces, and open grasslands, together with multiple illumination conditions such as daytime, dusk, and backlit settings. To introduce real-world inter-class confusion between drones and

birds, 1,500 bird images were selected from the AirBirds dataset, supplemented by 10 additional internet-sourced images.

The final dataset comprises 7,874 images in total: 7,254 positive samples (images containing at least one detection target, 92.13%) and 620 negative samples (interference images with UAV flight backgrounds but no targets to detect, 7.87%), used to suppress model false alarms. The full dataset contains 14,642 bounding box annotations with an average of 2.02 targets per positive image, spanning two categories: UAV (class 0) and BIRD (class 1). Among these, there are 11,236 UAV instances (76.74%) and 3,406 bird instances (23.26%), with all label files having corresponding images and no redundant entries. The dataset is partitioned into training, validation, and test sets at approximately 60%/20%/20%, following the standard convention in object detection.

A salient characteristic of this dataset is the extremely high proportion of small targets. Among UAV instances, 86.56% (9,726) are extremely small targets with areas below 16×16 pixels, and 97.43% (10,947) are below 32×32 pixels; the average bounding box is 15.01×9.71 pixels, with the smallest target area being only 3.15 square pixels. Bird instances present even greater detection difficulty: 96.65% (3,292) are below 16×16 pixels, and 99.35% (3,384) are below 32×32 pixels. UAV target aspect ratios range from 0.60 to 4.40, while bird target aspect ratios reach 0.15 to 14.76, fully reflecting significant intra-class morphological variation under multiple viewpoints and flight attitudes. These statistical characteristics make SUAV-Bird an extremely challenging small-target benchmark, providing a rigorous evaluation environment for ECSFNet-DET. As shown in Figure 2, UAV annotation centre points are broadly distributed in the upper image half, consistent with typical spatial distributions of low-altitude flying targets; bird annotations are relatively concentrated in the middle, and the two target types spatially overlap to some extent, further increasing detection difficulty. As shown in Figure 3, bounding box sizes of both target types are highly concentrated below 16×16 pixels, with prominent extreme small-target proportions and greater aspect ratio dispersion for bird targets, further verifying the high challenge level of this dataset.

2.2 YOLOv11n Baseline Framework

ECSFNet-DET adopts YOLOv11n [15] as its base architecture. YOLOv11n is a single-stage anchor-free real-time object detector released by Ultralytics in September 2024. Its backbone uses C3k2 modules as the core; the neck employs a PANet bidirectional path aggregation structure; the top C2PSA module introduces spatial attention to enhance global context modeling; and the detection head uses a decoupled architecture with a Task-Aligned Label Assignment strategy (TALS), achieving a good accuracy-efficiency balance. The three proposed core innovation modules—DVSFE, ABMSFN, and PAMGSE—are designed precisely to address the structural limitations of YOLOv11n, as detailed in Section 3.

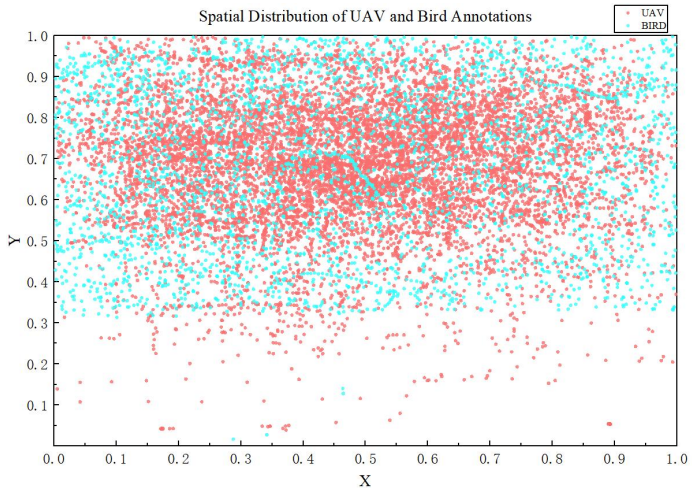


Fig. 2 Spatial distribution of UAV and bird annotation box centre points in the SUAV-Bird dataset.

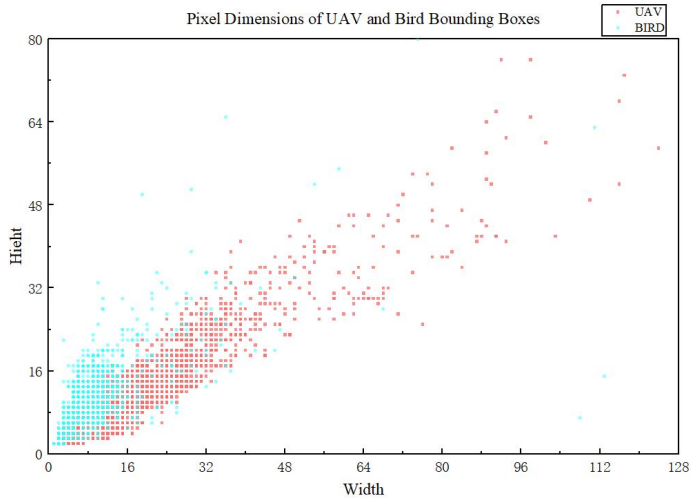


Fig. 3 Pixel size distribution of UAV and bird annotation bounding boxes in the SUAV-Bird dataset.

2.3 Related Work for Each Innovation

2.3.1 State Space Models for Visual Feature Extraction

This section reviews research on state space models (SSMs) and gated feed-forward networks relevant to DVSFE module design. SSMs achieve long-sequence modeling with linear computational complexity and have attracted broad attention. Mamba [16] introduces a selective scanning mechanism to realize input-dependent dynamic state transitions, with long-range dependency capture capability significantly superior to traditional recurrent models. VMamba [17] proposes a two-dimensional selective

scanning module that traverses feature maps along four directions, overcoming the sequentialization limitations of one-dimensional SSMs in two-dimensional image processing at linear complexity, establishing the foundational paradigm for visual SSM backbone networks. TransNeXt [18] proposes an aggregated attention mechanism and convolutional gated linear unit channel mixer, fusing gated linear units with channel attention to reveal the importance of local inductive bias in dense visual tasks through a multiplicative feature selection mechanism. Nevertheless, existing SSM backbones are designed for general visual recognition, making it difficult to simultaneously capture local fine details of sub-pixel-scale small targets and global long-range context at near-linear overhead; standard FFNs also lack local spatial perception, resulting in insufficient feature discrimination under complex clutter backgrounds. DVSFE integrates lightweight state space sequence modeling and local-enhancement gating mechanisms into the C3k2 dual-path reuse framework, achieving joint local and global feature representation with near-linear complexity.

2.3.2 Context-Guided Multi-Scale Feature Fusion

This section reviews multi-scale feature fusion research relevant to ABMSFN design. FPN [19] established the foundational paradigm for multi-scale feature pyramid fusion; PANet [20] further introduced bottom-up path augmentation to strengthen spatial information propagation; however, both adopt a level-by-level serial fusion approach, lack explicit global cross-scale semantic joint modeling, have overly long deep-to-shallow information transfer paths, and exhibit limited small-target perception capability in complex backgrounds. BiFPN [21] achieves adaptive bidirectional weighted fusion through learnable weights, improving cross-scale information flow efficiency to some extent, but still lacks explicit global cross-scale synchronized aggregation and spatially adaptive fusion mechanisms. ABMSFN realizes synchronized global multi-scale context aggregation through the Multi-Level Global Semantic Aggregation Module (MLGSAM), completes each-scale local detail calibration via Asymmetric Strip Channel-Spatial Calibrators (ASCSCs), and executes dynamic spatially-adaptive cross-scale feature injection and gated fusion through Dynamic Cross-Scale Alignment Injection Modules (DCSAIMs) and Semantic Gated Adaptive Fusion Blocks (SGAFBlocks), achieving deep complementary representation of multi-scale features at lightweight cost.

2.3.3 Multi-Granularity Spectral Enhancement Attention

This section reviews research on attention enhancement, multi-scale convolutions, and frequency-domain feature modulation relevant to PAMGSE design. The C2PSA module of YOLOv11n employs polarized self-attention (PSA) for global channel feature modeling, decomposing self-attention into channel and spatial polarization branches, modeling global receptive fields at lower computational cost, yet its built-in FFN relies on linear transformations lacking spatial structural perception capability, and completely ignores frequency-domain structural information, which is particularly critical for distinguishing low-texture UAV targets in complex clutter backgrounds. LKA [22] achieves long-range spatial dependency modeling without additional attention computation overhead by decomposing large receptive field kernels into cascaded depth-wise

and dilated depth-wise convolutions; MDFD2-DETR [23] further demonstrates that introducing frequency-domain feature decomposition into detectors can significantly improve detection performance in complex scenes, revealing the value of frequency-domain representations for small-target detection. PAMGSE embeds an Adaptive Normalisation Multi-scale Hybrid Convolutional Enhancement network (ANMHCE) and a Spectral Adaptive Modulation Gated Feed-forward Network (SAMGFFN) sequentially after the PSA block with zero-initialised scalar gated residuals, where ANMHCE implements multi-granularity local texture enhancement through parallel 3×3, 5×5, and 7×7 depth-wise separable convolutions, SAMGFFN completes frequency-adaptive modulation based on two-dimensional real FFT with learnable complex weights, and the complete system degenerates to a standard PSABlock at the beginning of training, adaptively activating the two enhancement branches as training progresses.

3 Method

The improved overall architecture of ECSFNet-DET—an Efficient Convolutional State-space and Frequency-adaptive Fusion Network Detector—is shown in Figure 4. It takes YOLOv11n as the base framework and performs systematic collaborative improvements at three levels: the backbone network, the feature fusion neck, and the detection head. In the backbone network stage, input images undergo multi-level strided convolutions to complete multi-resolution downsampling, and DVSFE is introduced at three key-scale feature layers P3, P4, and P5, fusing a Lightweight State Space Sequence Modeler (LSSSM) and a Depth-wise Convolutional Activation Gated Feed-forward Unit (DCAGFU) within a dual-branch architecture to simultaneously capture local spatial details and global long-range context dependencies at near-linear computational complexity; P5 features pass through SPPF and are then fed into PAMGSE, which performs three-stage progressive refinement through polarized self-attention, ANMHCE, and SAMGFFN to form high-quality deep feature representations with global semantic association, multi-scale local texture, and frequency-domain structural perception. In the feature fusion stage, backbone-output P3, P4, and PAMGSE-enhanced P5 features are first fed into MLGSAM, which uses a Multi-Scale Average Pooling and Channel Aggregator (MSAPCA) to pool and concatenate three-scale features, then outputs context guidance features carrying cross-scale global priors after recursive refinement through multiple ASCSCs; each scale's backbone features then undergo local detail enhancement via ASCSC, followed by DCSAIMs for adjacent-layer semantic interpolation alignment in the top-down transmission path, then gated-weighted fusion with MLGSAM's global context guidance features through SGAFBlocks, progressively generating enhanced P5, P4, and P3 feature maps; after bottom-up path augmentation, P3 detection features are strided-downsampled, concatenated with P4, and refined through DVSFE to output intermediate-scale detection features; intermediate-scale features apply the same operation with P5 to output large-scale detection features; finally, the three-scale feature maps are jointly fed into the detection head to complete multi-scale joint prediction for small, medium, and large counter-UAV targets.

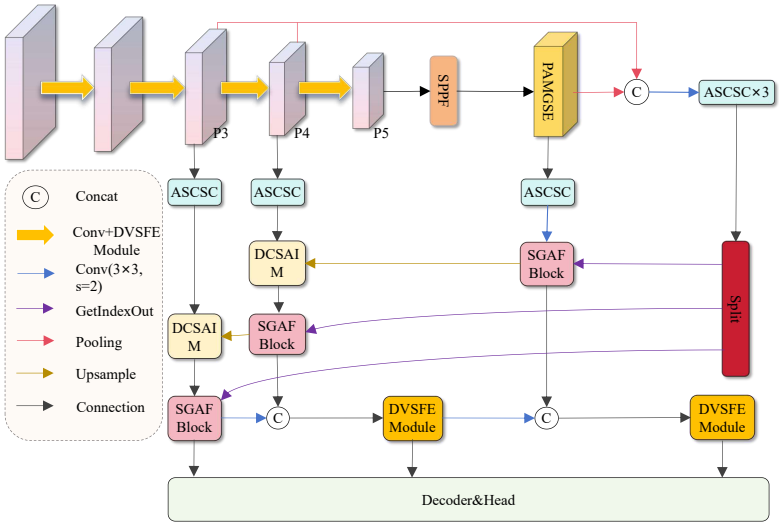


Fig. 4 Overall network architecture of ECSFNet-DET.

3.1 Dual-Branch Visual State-space Feature Encoder (DVSFE)

In standard YOLOv11n, the C3k2 module serves as the core feature extraction unit and relies on stacked standard convolution kernels with residual connections to learn local semantic information; however, its receptive field is constrained by fixed kernel sizes, making it difficult to capture long-range cross-scale dependencies. Furthermore, the pointwise linear transformations in traditional FFNs lack local spatial structural perception, resulting in insufficient small UAV target representation under complex backgrounds and readily triggering missed detections and false alarms. To this end, this paper proposes DVSFE, which deeply fuses an efficient visual state space modeling mechanism with a convolutional gated linear unit, substantially improving the network's local detail perception and global context modeling capability for multi-scale targets without significantly increasing computational overhead. Its structure is shown in Figure 5.

DVSFE replaces the original bottleneck layers in C3k2's dual-branch architecture with a stacked unit of Progressive State Convolutional Gated Refinement Blocks (PSCGRBlocks). After channel splitting, input feature maps enter a main branch and a shortcut branch respectively; within the main branch each PSCGRBlock sequentially executes depth-wise separable convolution preprocessing, LSSSM-based state space sequence modeling with layer normalization, depth-wise convolutional post-processing, and DCAGFU gate-based feed-forward network enhancement; the main branch output is then concatenated with the shortcut branch and passed through a channel fusion convolution. The feature transformation process can be formally expressed as:

$$F_{out} = \mathcal{W}_{fuse}([\mathcal{E}_n \circ \mathcal{E}_{n-1} \circ \cdots \circ \mathcal{E}_1(F_{main}),\ F_{short}]) \tag{1}$$

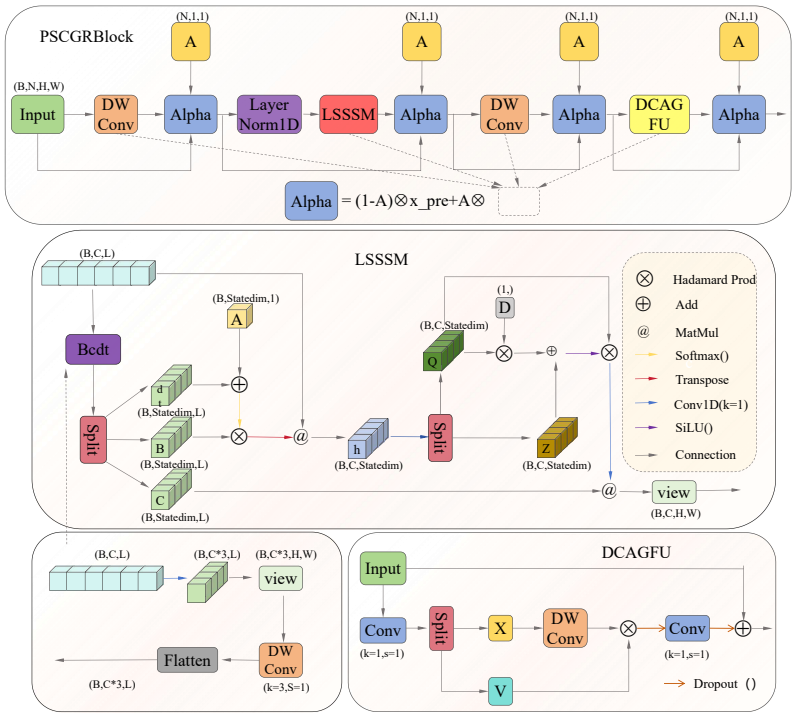


Fig. 5 Structure diagram of the DVSFE module.

where F_{main} and F_{short} are the main and shortcut features obtained after channel splitting, $\mathcal{E}_i(\cdot)$ denotes the mapping function of the i-th PSCGRBlock, $\circ$ denotes function composition, $[\cdot,\cdot]$ denotes channel-wise concatenation, and $\mathcal{W}_{fuse}$ is the final 1×1 channel fusion convolution. The channel compression constraint governing input feature dimensions is:

$$c' = \lfloor c_2 \cdot e \rfloor, \qquad F_{main},\, F_{short} \in \mathbb{R}^{B \times c' \times H \times W} \tag{2}$$

where c_2 is the number of output channels of the module, e is the channel compression coefficient (default $e = 0.5$), and B, H, W denote batch size, feature map height, and width, respectively.

LSSSM is the core sequence modeling component of this module. Input feature maps $X \in \mathbb{R}^{B \times C \times H \times W}$ are first flattened into sequences $x \in \mathbb{R}^{B \times C \times L}$ (where $L = H \times W$), projected via 1×1 convolutions to obtain the state space triple $(B, C_s, \Delta t)$, with local spatial dependencies introduced through depth-wise separable convolutions; the normalized state transition matrix A is constructed via Softmax activation, and global sequence aggregation is achieved efficiently through matrix multiplication. DCAGFU replaces the standard FFN with a gating mechanism, achieving selective enhancement of local structures through split projection and depth-wise convolutional activation

with multiplicative interaction. The core state aggregation process of LSSSM is:

$$A = \text{Softmax}(\Delta t + A_{init}), \qquad h = x \cdot (A \odot B)^{\top}, \qquad y = h \cdot C + h \cdot d \tag{3}$$

where B, $C \in \mathbb{R}^{B\times N\times L}$ are input-dependent state projection matrices, $\Delta t \in \mathbb{R}^{B\times N\times L}$ is the time-step projection, $A_{init} \in \mathbb{R}^{N}$ is a learnable state initialization parameter, $\odot$ denotes element-wise product, and d is a learnable skip-connection scalar. The gated feature transformation process of DCAGFU is:

$$Y_{DCAGFU} = X + \mathcal{W}_2\Big(Drop\Big(DWConv\Big(\mathcal{W}_1^{(g)}(X)\Big) \odot \mathcal{W}_1^{(v)}(X)\Big)\Big) \tag{4}$$

where $\mathcal{W}_1^{(g)}(\cdot)$ and $\mathcal{W}_1^{(v)}(\cdot)$ are the gating and value components obtained by equally splitting the 1×1 projection output along the channel dimension, $DWConv(\cdot)$ is a 3×3 depth-wise separable convolution with GELU activation, $\mathcal{W}_2(\cdot)$ is the output projection, and the residual path ensures training stability.

PSCGRBlock adopts a four-stage progressive feature refinement procedure and introduces a learnable channel-level LayerScale mechanism to dynamically modulate each stage's contribution weights, maintaining gradient stability in early training. Stage 1 captures local texture priors with $DWConv_1$; Stage 2 serializes features into LSSSM for global long-range state modeling; Stage 3 reintroduces $DWConv_2$ to compensate for local detail loss after sequence modeling; Stage 4 performs non-linear channel transformation via FFN. Each stage is fused via a learnable scalar-gated residual, and the overall recursive form can be uniformly expressed as:

$$X^{(k)} = (1 - \alpha_k) \cdot X^{(k-1)} + \alpha_k \cdot \mathcal{T}_k(X^{(k-1)}), \qquad k = 1, 2, 3, 4 \tag{5}$$

where $\mathcal{T}_k(\cdot)$ corresponds sequentially to $DWConv_1$, $LSSSM \circ LN$, $DWConv_2$, and FFN; $\alpha_k = Sigmoid(\hat{\alpha}_k) \in (0, 1)$ is the learnable channel weight vector for the corresponding stage (initialized to 10^{-4}), ensuring through Sigmoid mapping that it always stays in an effective control range, thereby starting training with an approximately identity mapping and progressively activating each stage's feature enhancement capability.

DVSFE fuses efficient state space sequence modeling with convolutional gated linear units in the C3k2 dual-path feature reuse framework, achieving joint local-global representation at multiple scales: LSSSM completes state transition aggregation of flattened feature sequences at near-linear complexity, breaking the fixed receptive field limitation of standard convolutions; DCAGFU replaces traditional linear FFNs with a depth-convolution-enhanced gating mechanism, endowing the network with adaptive key-feature screening capability under complex backgrounds; the learnable LayerScale mechanism ensures gradient propagation stability in deep networks from a training dynamics perspective. The combined effect of these designs enables the improved feature extraction module to significantly enhance multi-scale feature discrimination capability for small-size, low-contrast counter-UAV targets while maintaining lightweight computational overhead.

3.2 Adaptive Bidirectional Multi-Scale Semantic Fusion Neck (ABMSFN)

Traditional path aggregation feature pyramid networks (PAFPNs) exhibit significant limitations in multi-scale feature fusion: they adopt a level-by-level serial upsampling-concatenation paradigm, lack explicit global semantic context modeling between scale features, have overly long high-level-to-low-level semantic transmission paths, and insufficiently model channel and spatial correlations across scales. Furthermore, fixed-weight feature concatenation cannot adaptively adjust contributions of features at different scales, often causing fine-grained small-target features to be overwhelmed by strong semantic features when facing variable UAV scales and complex backgrounds. To this end, this paper proposes ABMSFN, which achieves efficient bidirectional interaction among multi-scale features through global pyramid context extraction, recursive feature calibration, and dynamic adaptive fusion mechanisms, effectively alleviating the semantic gap and significantly improving the model's perception capability for UAV targets at different scales.

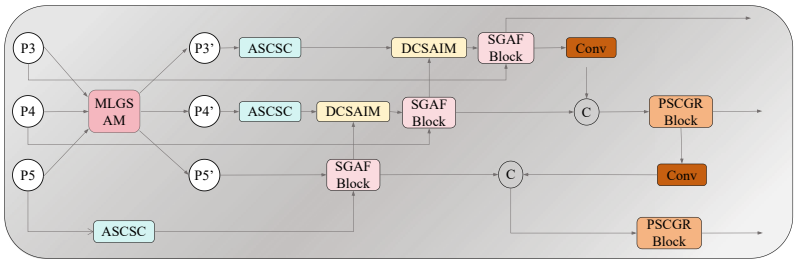


Fig. 6 Overall workflow diagram of ABMSFN.

The overall workflow of ABMSFN is shown in Figure 6. Given three backbone output scale feature maps $F^s \in \mathbb{R}^{B\times C_s\times H_s\times W_s}$ ($s \in \{3,4,5\}$, corresponding to feature layers at strides 8, 16, and 32) and the high-level semantic feature Y_5 after AIFI transformation, ABMSFN first performs global context aggregation on the three-scale backbone features through MLGSAM, obtaining context guidance features $\{\hat{F}^3, \hat{F}^4, \hat{F}^5\}$ carrying cross-scale global perceptual information. Each scale's backbone features are then separately enhanced through ASCSCs to obtain calibrated features $\{\tilde{F}^s\}$. Finally, each-scale calibrated features are sequentially processed through DCSAIMs and SGAFBlocks to complete top-down semantic injection and cross-scale alignment, generating the final detection feature maps $\{X^3, X^4, X^5\}$. The entire process can be uniformly expressed as:

$$X^s = \Phi_{SGAFB}\left(\Phi_{DCSAIM}\left(\tilde{F}^s,\, X^{s+1}\right),\, \tilde{F}^s\right), \qquad s = 5,4,3 \tag{6}$$

where $\tilde{F}^s = ASCSC(F^s)$, $X^5 = \Phi_{SGAFB}(ASCSC(F^5), \hat{F}^5)$ is the top-level fusion feature, and Φ_{DCSAIM} and Φ_{SGAFB} denote the dynamic cross-scale alignment injection and semantic gated adaptive fusion operations, respectively. Additionally, the

$\{X^4, X^5\}$ processed by ABMSFN undergo bottom-up path augmentation to further strengthen localization information propagation:

$$Z^{s+1} = RepC3\big(Cat\big[Conv_{3\times3,\downarrow2}(X^s),\ X^{s+1}\big]\big), \qquad s = 3, 4 \tag{7}$$

where $Conv_{3\times3,\downarrow2}$ denotes a downsampling convolution with stride 2, $Cat[\cdot]$ denotes channel-dimension concatenation, $RepC3(\cdot)$ is a cross-stage partial module based on reparameterization mechanisms, and the three-scale features fed into the detection head are $\{Z^3, Z^4, Z^5\}$.

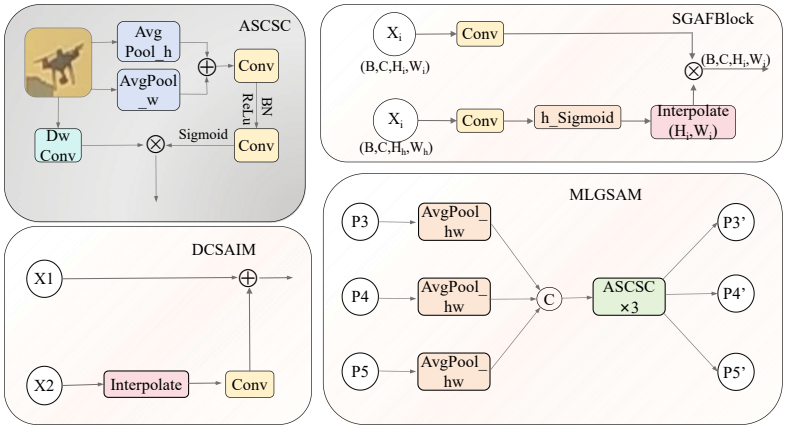


Fig. 7 Structure diagrams of ABMSFN sub-modules.

MLGSAM distills global context priors from multi-scale features to provide cross-scale global perceptual guidance for feature fusion at each scale. MLGSAM first uniformly adaptive-pools the three input-scale feature maps to the smallest scale $H_5 \times W_5$ through the Multi-Scale Average Pooling and Channel Aggregator (MSAPCA), concatenating them along the channel dimension to form multi-scale aggregated feature $G \in \mathbb{R}^{B\times\sum C_s\times H_5\times W_5}$:

$$G = Cat\big[AvgPool_{H_5\times W_5}(F^3),\ AvgPool_{H_5\times W_5}(F^4),\ AvgPool_{H_5\times W_5}(F^5)\big] \tag{8}$$

G then undergoes recursive contextual refinement through n serially connected ASCSC sub-modules, and is split according to each scale's channel count $\{C_3, C_4, C_5\}$ to output three context guidance feature sets $\{\hat{F}^3, \hat{F}^4, \hat{F}^5\}$. A single ASCSC simultaneously models local spatial structures and global strip channel attention: for input feature x, ASCSC extracts local spatial details through depth-wise separable convolution $\mathcal{K}_{sq}$, compresses spatial dimensions via horizontal and vertical adaptive average pooling, estimates channel excitation weights a spanning spatial directions via asymmetric strip convolution sequences, and completes feature calibration through element-wise

product:

$$a = \sigma(\mathcal{K}_v(\delta(BN(\mathcal{K}_h(AvgPool_h(x)) + \mathcal{K}_v(AvgPool_w(x)))))), \quad ASCSC(x) = a \odot \mathcal{K}_{sq}(x) \tag{9}$$

where $\mathcal{K}_h \in \mathbb{R}^{1\times k}$ and $\mathcal{K}_v \in \mathbb{R}^{k\times 1}$ are horizontal and vertical strip convolution kernels ($k = 11$), $\sigma(\cdot)$ is the Sigmoid activation function, $\delta(\cdot)$ is the ReLU activation function, and $\odot$ denotes element-wise multiplication.

After global context extraction and local feature calibration, ABMSFN achieves refined multi-scale feature fusion through SGAFBlock and DCSAIM. SGAFBlock receives two inputs—local calibrated feature x_i (low-level/current level) and global guidance feature x_h (context feature output by MLGSAM)—and applies adaptive weighted modulation to local features through a learnable gating mechanism: both features are separately projected via independent 1×1 convolutions; the high-level guidance feature is Hard-Sigmoid activated and bilinearly interpolated to align to the current scale, applying spatial adaptive gating modulation to local features:

$$X^s_{SGAFB} = \mathcal{W}_1(x_i) \odot \mathcal{H}(\mathcal{W}_2(x_h) \uparrow_{H\times W}) \tag{10}$$

where $\mathcal{W}_1, \mathcal{W}_2$ are 1×1 convolutions, $\mathcal{H}(\cdot) = \min(\max(x/6 + 1/2,\, 0),\, 1)$ is the Hard-Sigmoid activation function, and $\uparrow_{H\times W}$ denotes bilinear interpolation upsampling to target size. DCSAIM handles feature transfer between adjacent scales: given higher-level output feature X^{s+1} and current-level feature x_{cur}, DCSAIM adjusts channel count via 1×1 convolution, then interpolates and aligns the higher-level feature and adds it element-wise to the current-level feature, achieving lightweight and efficient cross-scale semantic injection:

$$X^s_{DCSAIM} = x_{cur} + \mathcal{W}_{1\times 1}\big(Interp_{H_s\times W_s}(X^{s+1})\big) \tag{11}$$

where $\mathcal{W}_{1\times 1}$ denotes the linear projection convolution mapping X^{s+1} channels from C_{s+1} to C_s, and $Interp_{H_s\times W_s}(\cdot)$ denotes bilinear interpolation to target resolution.

In summary, ABMSFN replaces the single serial concatenation paradigm of traditional PAFPN with a lightweight module combination, constructing a bidirectional multi-scale feature flow pathway covering global to local and high-level to low-level through the organic synergy of MLGSAM global context aggregation, ASCSC recursive channel-spatial attention calibration, SGAFBlock gated adaptive fusion, and DCSAIM dynamic interpolation semantic injection. This structure effectively alleviates the semantic gap problem in cross-scale feature fusion, enhances the model's fine-grained perception capability for small UAV targets and its robustness under complex backgrounds, and comprehensively improves the detector's multi-scale target representation quality without significantly increasing model parameter count or computational complexity.

3.3 Polarized Attention Multi-Granularity Spectral Enhancement Module (PAMGSE)

Traditional C2PSA relies solely on PSA for global channel feature modeling; its built-in FFN uses a structurally simple linear mapping that both lacks fine-grained multi-scale local texture capture capability and completely ignores frequency-domain structural information. When confronting extreme challenges in counter-UAV detection scenes—tiny target sizes, complex background textures, and strong imaging noise—a single attention mechanism fails to establish cross-scale spatial-frequency joint representations, yielding insufficient feature discrimination and elevated missed detection rates. To this end, this paper proposes PAMGSE, which serially connects ANMHCE and SAMGFFN after the PSA attention via learnable gated residuals, strengthening feature expression from complementary spatial-local modeling and frequency-domain global modulation dimensions, effectively compensating for traditional C2PSA's structural deficiencies in multi-granularity perception and frequency-domain information utilization.

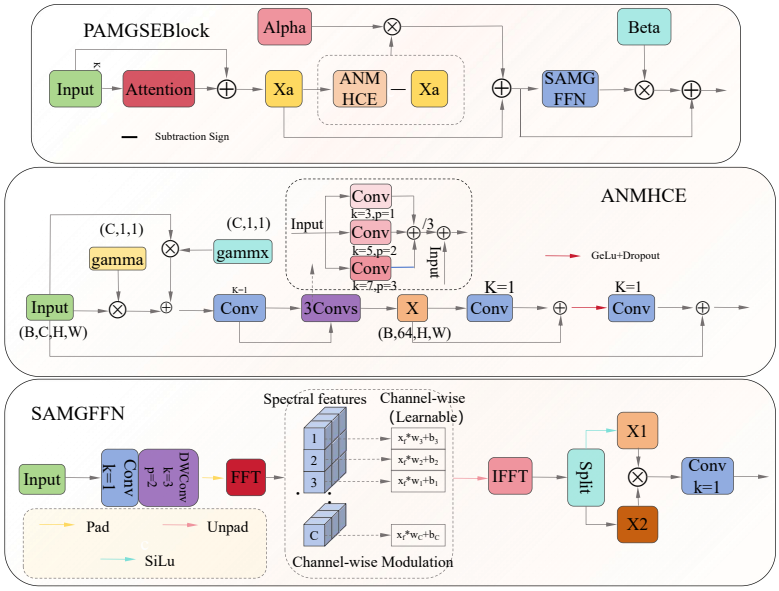


Fig. 8 Structure diagram of the PAMGSE module.

PAMGSE inherits C2PSA's classic Split-Aggregate backbone structure, mapping input feature map $X \in \mathbb{R}^{C\times H\times W}$ to $2c$ dimensions via 1×1 convolution $\mathcal{W}_{cv1}$ and splitting evenly into two branches: one serves as an identity bypass transmitting low-frequency structural information to ensure stable gradient flow; the other undergoes deep semantic enhancement through n serially connected PAMGSEBlock blocks; the two streams are concatenated along the channel dimension and restored to the original

channel count via $\mathcal{W}_{cv2}$. This splitting and aggregation process can be expressed as:

$$[a, b] = Split(\mathcal{W}_{cv1}(X)), \qquad b^* = \mathcal{M}^{(n)}(b), \qquad Y = \mathcal{W}_{cv2}(Concat(a, b^*)) \tag{12}$$

where $a, b \in \mathbb{R}^{c \times H \times W}$, $c = \lfloor C \cdot e \rfloor$ is the compressed intermediate channel count, $\mathcal{M}^{(n)}$ denotes the sequential composite mapping of n PAMGSEBlock blocks, and $Concat(\cdot)$ is the channel-axis concatenation operation.

ANMHCE compensates for the shortfall of self-attention mechanisms in local texture fine-grained perception. As shown in Figure 8, ANMHCE applies learnable adaptive layer normalization to input features, with independent learnable scaling vectors γ and γ_x dynamically weighting the normalized and identity components respectively, achieving adaptive balance between global statistics and local distribution characteristics. After projecting features to an intermediate space via 1×1 projection convolution, parallel depth-wise separable convolutions with kernel sizes 3×3, 5×5, and 7×7 capture multi-granularity local geometric structures; these three are equal-weight averaged and residually combined with the intermediate features, then processed through pointwise aggregation projection, GELU activation, and Dropout regularization, with a global residual connection output. The complete computational procedure of ANMHCE is:

$$ANMHCE(x) = x + \mathcal{W}_2(Drop(GELU(r(z_{mix} + z) + z_{mix} + z))) \tag{13}$$

where $z = \mathcal{W}_1(\gamma \odot LN(x) + \gamma_x \odot x)$, $z_{mix} = \frac{1}{3}\sum_{k\in\{3,5,7\}} DW_k(z)$, r is the pointwise aggregation convolution, DW_k denotes a depth-wise separable convolution with kernel size k, and $\odot$ is element-wise multiplication.

SAMGFFN provides complementary enhancement to ANMHCE from the frequency domain. After input features are mapped by 1×1 expansion convolution, large-receptive-field local modeling is performed by a depth-wise separable convolution with dilation rate $d = 2$; two-dimensional real fast Fourier transform (2D-rFFT) is then applied, and each frequency component is adaptively linearly modulated in the spectral domain through learnable complex weights w_f and bias b_f; after inverse FFT restoring to the spatial domain, a Gated Linear Unit (GLU) structure realizes gated non-linear activation and channel compression output:

$$SAMGFFN(x) = SiLU(GELU(h_1 \odot h_2)) \tag{14}$$

$$h = \mathcal{F}^{-1}(w_f \odot \mathcal{F}(DWConv_{d=2}(DW(x))) + b_f), \qquad h_1, h_2 = Split(h) \tag{15}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ respectively denote two-dimensional real fast Fourier transform and its inverse, and $SiLU$ is the activation function. This mechanism enables the model to establish effective information synergy between fine-grained spatial-domain local modeling and frequency-domain global modulation.

PAMGSEBlock takes the polarized self-attention of the original PSABlock as its backbone, and embeds ANMHCE and SAMGFFN sequentially after it via learnable scalar gated residuals. Given input feature x, the module first extracts global context correlations through the PSA attention sub-module, forming attention-enhanced feature x_a; then with learnable scalar α as gating weight, ANMHCE adaptively fuses

the original attention features with multi-scale local enhanced features, obtaining x'_a; finally, with learnable scalar β controlling frequency-domain enhancement contribution intensity, the SAMGFFN spectral modulation result is superimposed on the above output. The overall computational procedure is:

$$x_a = x + Attn_{PSA}(x) \tag{16}$$

$$x'_a = x_a + \alpha \cdot (ANMHCE(x_a) - x_a) \tag{17}$$

$$y = x'_a + \beta \cdot SAMGFFN(x'_a) \tag{18}$$

where $\alpha, \beta \in \mathbb{R}$ are both initialized to zero, ensuring PAMGSEBlock degenerates to a pure polarized self-attention residual structure in early training, endowing the network with a stable optimization starting point. As training progresses, α and β gradually activate the enhancement branches of ANMHCE and SAMGFFN, adaptively determining the optimal fusion ratio in a data-driven manner, achieving hierarchical feature refinement from coarse to fine and from spatial to frequency domains.

In summary, PAMGSE constructs a composite feature enhancement system integrating spatial-domain local texture fine-grained perception, frequency-domain global structural modeling, and channel self-attention through collaborative introduction of multi-scale hybrid convolutional enhancement and spectral adaptive modulation mechanisms on a polarized self-attention backbone. With learnable gated residual mechanisms realizing adaptive dynamic fusion of each enhancement branch, PAMGSE fundamentally breaks through the structural bottleneck of traditional C2PSA's single attention paradigm in multi-scale perception and frequency-domain information utilization without introducing large numbers of additional parameters, significantly improving the model's feature discrimination robustness for complex background clutter, cross-scale target morphology variation, and small-size UAV targets.

4 Experimental Results and Analysis

4.1 Experimental Setup

All experiments in this chapter were conducted under a unified software and hardware environment to ensure reproducibility and fair comparison across all settings. The hardware platform employs a workstation equipped with an NVIDIA RTX 5090 GPU (32 GB VRAM), an Intel Core i9 processor, and 128 GB system memory; the software environment builds the deep learning stack on Windows 11 with Python 3.11, PyTorch 2.1, and CUDA 12.1, with Ultralytics YOLOv11 as the object detection framework. Regarding training strategy, all models adopt consistent hyperparameter configurations to eliminate training setting differences as confounds. Specifically, input images are uniformly resized to 1024×864 pixels, batch size is set to 32, total training is 300 epochs, an SGD optimizer is used, and the initial learning rate is 0.02. Three datasets are used in experiments: (1) SUAV-Bird, the self-constructed counter-UAV dataset in this paper, containing drone and bird target images collected under different backgrounds (urban buildings, mountain forests, sky, etc.), different illumination conditions, and different capture distances, annotated as "drone" and "bird";

it features small target sizes, complex backgrounds, and multiple targets per scene, serving as the primary benchmark for ablation experiments. (2) Det-Fly, a publicly available drone detection dataset covering images of multiple UAV types at different viewpoints and altitudes, with relatively clean backgrounds and high target clarity; Det-Fly is incorporated to verify generalization stability of each module under different data distributions. (3) Mixdata, constructed by fusing data from SUAV-Bird, Det-Fly, and DUT Anti-UAV to evaluate ECSFNet-DET's robust generalization capability under multi-source distributions; experiments on Mixdata help identify which methods possess genuine cross-domain robustness.

4.2 Evaluation Metrics

This paper employs six metrics for comprehensive model evaluation. Precision (P) and Recall (R) both range from 0 to 1, measuring respectively the proportion of true positives in detection results and the proportion of all true targets successfully detected; higher values indicate lower false alarm rates and missed detection rates. mAP50 computes the mean over all classes of the area under the precision-recall curve at a fixed IoU threshold of 0.5, reflecting comprehensive detection capability under relaxed localization criteria; mAP50-95 calculates average precision at 10 IoU thresholds from 0.5 to 0.95 with step 0.05 and takes the overall mean, imposing stricter localization constraints; both are reported as percentages. GFLOPs and model parameter count (Params, unit: M) measure the floating point operations required per forward pass and the total number of trainable weights, respectively; lower values indicate smaller computational complexity and storage overhead, stronger deployment feasibility.

4.3 Ablation Studies

4.3.1 DCAGFU Component Ablation

To investigate the individual contributions of key DCAGFU components within DVSFE, this paper designed combinatorial ablation experiments on the Gating mechanism and DWConv on SUAV-Bird, systematically comparing detection performance under four configurations: Standard FFN, Linear-GLU, Conv-FFN, and complete DCAGFU. Results are shown in Table 1.

Table 1 Ablation experiment results of DCAGFU components in DVSFE under different feed-forward network configurations.

Model	Description	Gating	DWConv	GFLOPs	Params	P	R	mAP50	mAP50-95
A1	Standard FFN	—	—	6.7	2.47M	84.9	75.4	80.5	40.8
A2	Linear-GLU	✓	—	6.6	2.47M	84.7	76.3	80.9	41.2
A3	Conv-FFN	—	✓	6.5	2.43M	83.7	75.4	79.1	41.7
A4 (Ours)	DCAGFU	✓	✓	6.7	2.48M	84.4	76.6	81.3	43.3

Results show that independently introducing the gating mechanism (A2, Linear-GLU) improves recall by 0.9% and mAP50 by 0.4% over the baseline, indicating that

the multiplicative gating pathway has substantive benefit for positive sample recall. Independently introducing depth-wise separable convolution (A3, Conv-FFN) compresses parameter count by approximately 0.04M and reduces GFLOPs by 0.2, but mAP50 declines by 0.4%, indicating that relying solely on local convolutional perception without a gating screening mechanism tends to introduce redundant noise features in complex backgrounds, conversely weakening detection accuracy. The complete DCAGFU (A4, Ours) achieves the best detection performance across all configurations: recall improves by 1.2% over the baseline, and mAP50 and mAP50-95 improve by 0.8% and 2.5% respectively, fully demonstrating a significant synergistic effect between the local structural perception of depth-wise convolution and the adaptive feature screening of the gating mechanism. These results validate that DCAGFU's multiplicative gating design can effectively enhance counter-UAV target feature discrimination under complex backgrounds without significantly increasing computational overhead.

4.3.2 PAMGSE Component Ablation

To verify the individual effectiveness of ANMHCE and SAMGFFN within PAMGSE, this paper conducted independent and joint ablation experiments of the two sub-modules with C2PSA as baseline (V1) on SUAV-Bird. Results are shown in Table 2.

Table 2 Independent and joint ablation experiment results of ANMHCE and SAMGFFN sub-modules within PAMGSE.

Model	Description	ANMHCE	SAMGFFN	GFLOPs	Params	P	R	mAP50	mAP50-95
V1 (Baseline)	C2PSA	—	—	6.3	2.58M	83.3	73.8	79.1	42.8
V2	ANMHCE	✓	—	6.3	2.61M	82.1	74.0	79.9	41.5
V3	SAMGFFN	—	✓	6.4	2.69M	83.9	74.6	79.8	41.0
V4 (Ours)	PAMGSE	✓	✓	6.4	2.71M	83.9	74.9	80.2	41.4

Independently introducing ANMHCE (V2) improves mAP50 by 0.8% over the baseline, confirming the local texture perception contribution of multi-scale hybrid convolutions to small-target feature refinement; however, mAP50-95 declines by 1.3%, showing that single spatial-domain enhancement has strict localization accuracy limitations. Independently introducing SAMGFFN (V3) improves mAP50 by 0.7%, verifying the positive effect of frequency-domain modulation on global structural feature perception, but similarly does not achieve notable breakthroughs in strict localization metrics. The complete PAMGSE (V4, Ours) improves mAP50 by 1.1% over the baseline with both precision and recall improved; recall reaches 74.9%, higher than any single configuration. This result demonstrates that spatial-domain multi-granularity local modeling and frequency-domain global structural modulation are complementary at the feature expression level, and synergistic embedding can compensate for each other's perceptual blind spots when used independently.

4.3.3 Stepwise Ablation Analysis of Core Module Contributions

To comprehensively evaluate the individual contributions and synergistic effects of ECSFNet-DET's three core modules, this paper designed seven groups of controlled experiments on SUAV-Bird by progressively stacking module combinations with YOLOv11n as baseline. Results are shown in Table 3 and Figure 9.

Table 3 Overall ablation experiment results of progressively stacked core modules of ECSFNet-DET.

Model	DVSFE	ABMSFN	PAMGSE	GFLOPs	Params	P	R	mAP50	mAP50-95
1 (base)				6.3	2.58	83.3	73.8	79.1	42.8
2	✓			6.7	2.48	84.4	76.6	81.3	43.3
3		✓		8.6	3.29	83.7	75.8	81.5	44.6
4			✓	6.4	2.71	83.9	74.9	80.2	41.4
5	✓	✓		8.8	3.15	85.6	77.0	82.0	43.9
6	✓		✓	6.7	2.64	86.5	77.2	80.7	42.6
7		✓	✓	8.8	3.43	86.0	76.4	80.4	41.8
8 (Ours)	✓	✓	✓	8.8	3.28	87.4	77.3	82.9	44.8

From single-module results, DVSFE (Experiment 2) improves mAP50 by 2.2% and mAP50-95 by 0.5% over the baseline while keeping parameter scale almost unchanged, with precision and recall improving by 1.1% and 2.8% respectively, fully demonstrating the dual-path visual state gated feature encoding mechanism's gain for small-target perception in complex backgrounds. ABMSFN (Experiment 3) brings the largest mAP50-95 improvement at 1.8% over the baseline, with mAP50 improving by 2.4%, validating the adaptive bidirectional multi-scale semantic fusion neck's outstanding contribution to cross-scale feature transfer and fine-grained localization, at the cost of approximately 0.71M additional parameters and 2.3 additional GFLOPs. PAMGSE (Experiment 4) improves mAP50 by 1.1% while almost not increasing computational load, with a significant enhancement effect on relaxed detection accuracy. Regarding dual-module synergy, the DVSFE+ABMSFN combination (Experiment 5) achieves additional synergistic gains beyond simple additive expectations, with mAP50 improving by 2.9% and mAP50-95 by 1.1%, indicating a benign feature complementary relationship between backbone enhancement and neck optimization. The complete three-module ECSFNet-DET (Experiment 8, Ours) achieves the best comprehensive performance across all configurations: precision improves by 4.1%, recall by 3.5%, mAP50 by 3.8%, and mAP50-95 by 2.0%, comprehensively surpassing the baseline and all single/dual module combinations with acceptable computational overhead of 8.8 GFLOPs and 3.28M parameters.

To intuitively verify the influence of each improvement module on model feature perception capability, this paper performs visual analysis of activation response distributions in four typical scenarios through Grad-CAM attention heatmaps; results are shown in Figure 10. Each row corresponds to a test sample group, and columns show in sequence: Original, baseline model (Base), after stacking DVSFE (A), after further introducing ABMSFN (A+B), and complete model (A+B+C) heatmap responses.

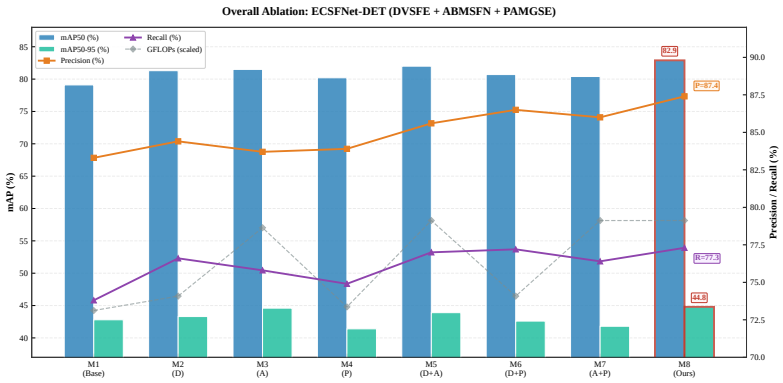


Fig. 9 Combined bar chart and line chart display of ECSFNet-DET overall ablation experiment results.

From baseline results, activation regions are relatively scattered with obvious background interference responses. With DVSFE introduced, responses progressively converge toward UAV target positions with background noise activations significantly suppressed. After stacking ABMSFN, the high-activation range of target regions is further refined. The complete model exhibits the most concentrated, boundary-clearest target activation regions across all test scenarios, fully demonstrating that the synergistic introduction of three modules progressively guides the model to focus on real targets while suppressing background clutter.

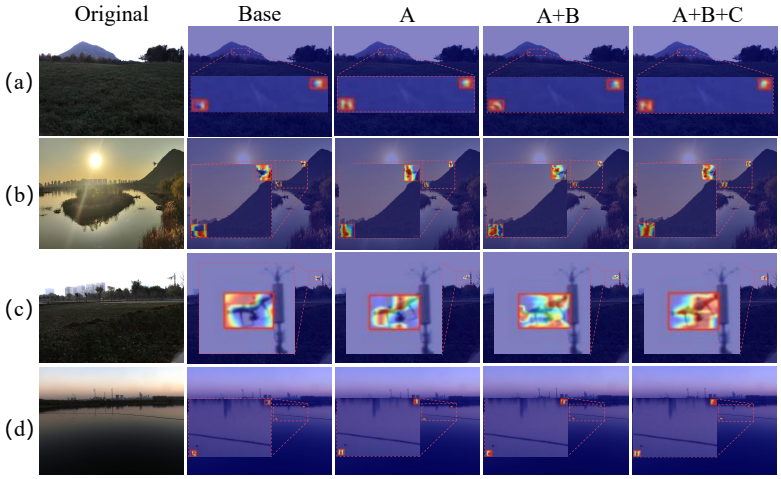


Fig. 10 Grad-CAM attention heatmap visualization under progressively stacked improvement modules.

4.4 Comparative Experiments

4.4.1 Lateral Comparison of Backbone Network Improvement Schemes

To objectively evaluate DVSFE's competitiveness as a backbone improvement scheme, this paper performs lateral comparison with mainstream backbone improvement methods on both Det-Fly and SUAV-Bird datasets, covering CNCM, CSSC, GCConv, and ES-MoE. Results are shown in Table 4.

Table 4 Detection performance comparison of different backbone network improvement schemes on Det-Fly and SUAV-Bird datasets.

Model	GFLOPs	Params	P	R	mAP50	mAP50-95
(Det-Fly Dataset)						
YOLOv11n	6.3	2.58	94.6	87.1	91.7	58.8
CNCM [24]	8.5	3.42	94.5	87.2	92.8	61.4
CSSC [24]	7.9	3.38	94.2	87.5	92.4	60.6
GCConv [25]	6.9	2.68	94.9	86.6	92.6	61.5
ES-MoE [26]	9.0	4.03	94.4	87.4	92.5	61.1
Ours	6.7	2.48	**95.0**	86.3	**92.9**	61.3
(SUAV-Bird Dataset)						
YOLOv11n	6.3	2.58	83.3	73.8	79.1	42.8
CNCM [24]	8.5	3.42	83.6	75.5	80.0	41.0
CSSC [24]	7.9	3.38	82.5	75.1	79.7	41.7
GCConv [25]	6.9	2.68	83.2	76.8	80.3	42.4
ES-MoE [26]	9.0	4.03	81.0	77.2	79.9	42.9
Ours	6.7	2.48	83.7	75.8	**81.5**	**44.6**

On Det-Fly, this paper's method achieves the highest precision of 95.0% among all comparison methods at the lowest computational overhead of 6.7 GFLOPs and 2.48M parameters, with mAP50 improving 1.2% over YOLOv11n; compared with the larger-parameter CNCM, this paper's method still achieves 0.1% higher mAP50 while saving approximately 21% in computation and reducing parameter count by approximately 27%. On the more challenging SUAV-Bird, mAP50 and mAP50-95 improve by 2.4% and 1.8% respectively over the baseline, ranking first at 44.6% mAP50-95 among all comparison methods, surpassing GCConv and CNCM which have larger parameter counts and computational loads. This demonstrates that DVSFE has unique advantages in capturing fine-grained features of small targets and improving strict localization accuracy, while maintaining stable performance superiority on two datasets with significantly different distributions, demonstrating good cross-domain generalization capability.

4.4.2 Lateral Comparison of Feature Fusion Neck Improvement Schemes

To further verify ABMSFN's effectiveness in multi-scale feature fusion, this paper selects representative neck improvement schemes CAFM, HFFE, RFPN, and LCA

as comparison benchmarks on both Det-Fly and SUAV-Bird. Results are shown in Table 5.

Table 5 Detection performance comparison of different feature fusion neck improvement schemes on Det-Fly and SUAV-Bird datasets.

Model	GFLOPs	Params	P	R	mAP50	mAP50-95
(Det-Fly Dataset)						
YOLOv11n	6.3	2.58	94.6	87.1	91.7	58.8
CAFM [27]	6.8	3.09	95.2	87.9	92.8	61.0
CAM [28]	7.5	2.80	95.1	87.6	92.6	61.1
GSConvE [29]	6.3	2.54	95.4	88.2	92.9	61.4
LCA [30]	8.6	3.25	94.6	85.8	91.5	59.2
Ours	8.6	3.29	94.9	**88.0**	**93.1**	61.3
(SUAV-Bird Dataset)						
YOLOv11n	6.3	2.58	83.3	73.8	79.1	42.8
CAFM [27]	6.8	3.09	82.8	74.3	79.0	40.7
CAM [28]	7.5	2.80	81.0	74.2	78.6	41.4
GSConvE [29]	6.3	2.54	83.4	75.6	80.9	43.1
LCA [30]	8.6	3.25	81.2	73.5	77.7	39.5
Ours	8.6	3.29	83.7	75.8	**81.5**	**44.6**

On Det-Fly, this paper's method achieves the highest mAP50 among all competing neck schemes, 1.4% above the baseline with 0.9% recall improvement. Notably, LCA introduces greater computational overhead yet its mAP50 actually decreases by 0.2% from the baseline, further highlighting ABMSFN's computational efficiency advantages. On SUAV-Bird, mAP50 and mAP50-95 improve by 2.4% and 1.8% over the baseline while surpassing all comparison methods; CAFM and CAM both achieve mAP50 below the baseline, indicating that some neck improvement schemes risk generalization failure on complex, high-difficulty datasets, while ABMSFN achieves stable performance improvements under both data distributions, validating its cross-scene robustness.

4.4.3 Comprehensive Performance Comparison with Mainstream Object Detection Methods

After completing independent module validation, this paper compares the complete ECSFNet-DET with multiple categories of mainstream detection methods including the YOLO series, Transformer-based methods, and YOLO improvement-type detectors, using SUAV-Bird as the test benchmark. Results are shown in Table 6 and Figure 11.

Lateral comparison results show that ECSFNet-DET has particularly pronounced advantages in precision and mAP50: precision of 87.4% ranks first among all comparison methods, improving 3.9% over YOLOv11n and surpassing the computationally lighter YOLOv26n by approximately 1.3%; mAP50 of 82.9% surpasses all YOLO-series and Transformer-based methods, exceeding the 65.2 GFLOPs UAV-DETR by 2.7%, while this paper's method requires only 8.8 GFLOPs, less than 14% of its computation. Among YOLO improvement-type methods, auav-yolo achieves the closest

Table 6 Comprehensive performance comparison of ECSFNet-DET with mainstream object detection methods on the SUAV-Bird dataset.

Model	GFLOPs	Params/M	P	R	mAP50	mAP50-95
YOLO-series						
YOLOv8n [37]	8.1	3.01	81.1	75.2	78.2	41.3
YOLOv10n [31]	6.5	2.26	83.6	76.3	81.4	44.6
YOLOv11n [15]	6.3	2.58	83.5	73.7	79.1	42.9
YOLOv26n [32]	5.2	2.37	86.1	79.5	83.1	44.3
Transformer-based						
UAV-DETR [33]	65.2	11.88	83.8	78.7	80.2	41.2
RTDETR-R18 [34]	56.9	19.87	82.3	79.7	82.3	40.7
YOLO-Based Object Detector						
auav-yolo [7]	14.6	3.42	83.7	80.1	81.9	45.1
FBRT-YOLO [35]	6.8	0.85	83.2	76.4	79.9	41.6
YOLO-Master [26]	9.6	75.0	88.9	69.5	60.0	34.9
YOLO-TLP [36]	9.2	1.91	81.3	78.9	80.4	42.4
Ours	8.8	3.28	**87.4**	77.3	**82.9**	44.8

mAP50 but its 14.6 GFLOPs burden is approximately 1.66 times this paper's method; YOLO-Master despite the highest precision has mAP50 of only 60.0%, exhibiting severe precision-recall imbalance. ECSFNet-DET achieves the optimal comprehensive accuracy-efficiency balance at 8.8 GFLOPs and 3.28M parameters, fully demonstrating the competitive advantages of the DVSFE, ABMSFN, and PAMGSE three-module collaborative design.

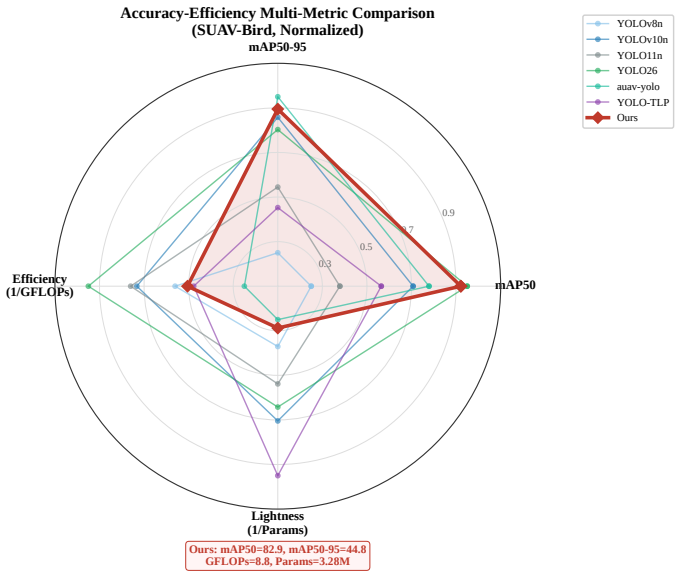


Fig. 11 Multi-metric accuracy-efficiency comparison of ECSFNet-DET and mainstream methods on Mixdata.



Fig. 12 Detection confidence score visualization comparison on the SUAV-Bird dataset.

Figure 12 presents detection result visualizations comparing this paper's method with YOLO-TLP, RT-DETR-R18, and YOLOv26n in four representative complex scenes on SUAV-Bird, covering overhead stadium views, mountain forest low-light backgrounds, branch occlusion, and building side facades. From detection confidence values, this paper's method outputs higher confidence than other comparison methods in all four scene groups; in scene (b) mountain forest low-light conditions, other methods achieve 0.64, 0.66, 0.66, while this paper's method reaches 0.74; in scene (c) branch occlusion conditions, this paper's method similarly outputs the highest confidence of 0.54, demonstrating that the collaborative design of DVSFE and ABMSFN effectively enhances the model's feature discrimination capability for small UAV targets under complex backgrounds and occlusion interference.

4.5 Cross-Domain Generalization Evaluation

To further verify ECSFNet-DET's cross-domain generalization robustness under multi-source data distributions, this paper evaluates it again on the Mixdata comprehensive benchmark constructed by three-source fusion of SUAV-Bird, Det-Fly, and DUT Anti-UAV, against all comparison methods from Table 6, to examine performance stability when migrating from single to multi-source distributions. Results are shown in Table 7 and Figure 13.

In rigorous generalization tests on the multi-source mixed dataset, some methods showed obvious performance degradation: FBRT-YOLO's mAP50 declined approximately 7.6% compared to SUAV-Bird single-dataset results, and YOLO-Master's mAP50 dropped by more than 6%, indicating strong dependence on specific data distributions and insufficient cross-domain robustness. By contrast, ECSFNet-DET maintained the highest mAP50-95 level of 44.8% on Mixdata, matching its SUAV-Bird performance, with performance stability the most outstanding among all comparison methods. mAP50 at 81.4% ranks second only to auav-yolo among all lightweight

Table 7 Generalization performance comparison of ECSFNet-DET and mainstream detection methods on the Mixdata multi-source fusion dataset.

Model	GFLOPs	Params/M	P	R	mAP50	mAP50-95
YOLO-series						
YOLOv8n [37]	8.1	3.0	81.4	78.4	77.6	41.3
YOLOv10n [31]	6.5	2.26	81.9	71.6	76.2	40.1
YOLOv11n [15]	6.3	2.58	82.7	74.6	78.7	42.6
YOLOv26n [32]	5.2	2.37	82.9	71.3	76.9	40.9
Transformer-based						
UAV-DETR [33]	65.2	11.88	73.0	73.5	72.2	39.1
RTDETR-R18 [34]	56.9	19.87	80.1	76.0	79.0	42.2
YOLO-Based Object Detector						
auav-yolo [7]	14.9	3.42	83.8	81.5	82.1	45.5
FBRT-YOLO [35]	6.7	0.85	76.1	69.8	72.3	36.7
YOLO-Master [26]	9.6	7.52	82.5	64.1	53.9	30.4
YOLO-TLP [36]	9.2	1.91	83.3	75.8	80.9	44.3
OURS	8.8	3.28	82.6	77.9	**81.4**	**44.8**

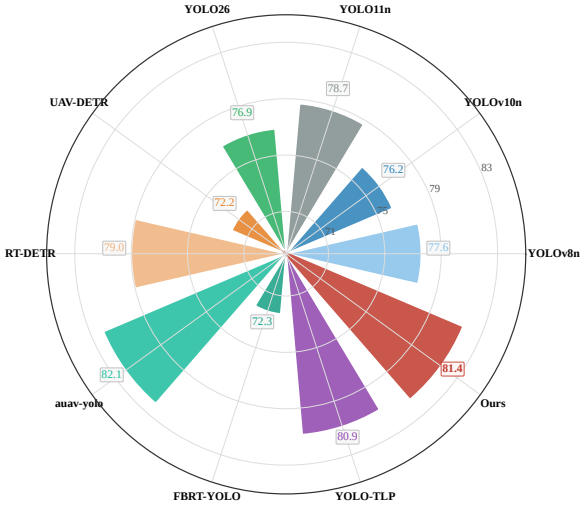


Fig. 13 Multi-dimensional normalized radar comparison chart based on the SUAV-Bird dataset.

methods, while this paper's method has computational overhead of only approximately 59% of auav-yolo. UAV-DETR's mAP50 declined to 72.2% on Mixdata, with much larger performance loss than this paper's method, further corroborating ECSFNet-DET's feature expression robustness in multi-source scenarios. Underpinned by ABMSFN's global context-guided fusion and DVSFE's state space sequence

modeling, ECSFNet-DET maintains stable detection performance in cross-domain scenarios with significantly different capture conditions and diverse target appearance distributions, possessing genuine multi-source generalization robustness.



Fig. 14 Detection confidence score visualization comparison on the Mixdata dataset.

Figure 14 presents visualized comparison of detection results in four cross-scene groups on Mixdata, covering dusk sky backgrounds, backlit building environments, evening streetlight scenes, and water surface reflection backgrounds. This paper's method maintained the highest or near-highest confidence output in all four scene groups: in scene (a), confidence reaches 0.87, higher than the other three methods; in scene (d) with water surface background, other methods score 0.79, 0.76, 0.75, while this paper's method reaches 0.81, maintaining stable detection responses in the scene with strongest imaging noise, demonstrating good generalization robustness in cross-domain multi-source scenarios.

5 Conclusion

This paper proposes ECSFNet-DET, a lightweight anti-drone detection algorithm based on visual state space modeling and spectral enhancement, targeting three core challenges in complex low-altitude scenarios: insufficient feature representation for tiny targets, semantic gaps in multi-scale feature fusion, and underutilization of frequency-domain information. Built upon YOLOv11n, three synergistic modules are developed across the backbone, feature fusion neck, and deep semantic enhancement stages. DVSFE integrates a lightweight state space sequence modeler with a depth-wise convolutional activation gated feed-forward unit within a dual-branch architecture, achieving joint local detail perception and global long-range context modeling at near-linear computational complexity. ABMSFN constructs a bidirectional multi-scale feature flow pathway through global context aggregation, recursive asymmetric strip channel-spatial calibration, dynamic cross-scale semantic injection, and gated adaptive

fusion, effectively mitigating the semantic gap across scales in conventional serial feature pyramid networks. PAMGSE embeds ANMHCE and SAMGFFN into a polarized self-attention backbone via learnable scalar-gated residuals, establishing a hierarchical joint spatial-frequency feature representation that fundamentally overcomes the structural limitations of the conventional C2PSA module. Furthermore, the self-constructed SUAV-Bird dataset fills the gap in existing public benchmarks that lack joint detection baselines for drones and visually similar aerial targets. Comprehensive experiments demonstrate that ECSFNet-DET achieves improvements of 4.1%, 3.5%, 3.8%, and 2.0% over the YOLOv11n baseline in precision, recall, mAP50, and mAP50-95 respectively on SUAV-Bird, attaining an optimal accuracy-efficiency trade-off at 8.8 GFLOPs and 3.28M parameters, with robust cross-domain generalization further validated on the multi-source Mixdata benchmark.

Despite the promising results, several directions merit further investigation. The scale and scene diversity of SUAV-Bird can be further expanded by incorporating samples collected under extreme weather conditions, cross-modal sensor configurations, and wider viewpoint variations to improve generalization under more demanding deployment scenarios. The bidirectional fusion pathway in ABMSFN introduces approximately 0.71M additional parameters, and future work may explore neural architecture search or knowledge distillation strategies to achieve a more compact model better suited for resource-constrained edge deployment. Moreover, since the current framework operates on single-frame images without exploiting temporal continuity, integrating sequential modeling and cross-frame trajectory association mechanisms could further improve stable continuous tracking of rapidly maneuvering small targets, providing more comprehensive technical support for practical engineering deployment of real-world anti-drone systems.

Acknowledgements. The authors gratefully acknowledge the support of the research institution and all collaborators who contributed to data collection and experimental validation.

Declarations

Funding. No funding was received for conducting this study.

Conflict of interest/Competing interests. The authors declare that they have no conflict of interest.

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Data availability. The SUAV-Bird dataset established in this work will be made publicly available once the manuscript is accepted.

Materials availability. Not applicable.

Code availability. The main source code of the proposed ECSFNet-DET will be made publicly available upon acceptance of the manuscript.

Author contribution. Y. Yu and Y. Zhang conceived the overall algorithm framework and designed the network modules of ECSFNet-DET. S. Liu and Y. Mu constructed the SUAV-Bird dataset and completed data annotation and data cleaning. C. Sun guided the research scheme and provided experimental hardware support.

F. Gao supervised the whole study, critically revised the manuscript for important intellectual content, and was responsible for paper submission.

References

[1] Shi W, Cao J, Zhang Q, Li Y and Xu L 2016 Edge computing: vision and challenges *IEEE Internet of Things J.* **3** 637–646

[2] Alsalam B H Y, Morton K, Campbell D and Gonzalez F 2017 Autonomous UAV with vision based on-board decision making for remote sensing and precision agriculture *Proc. IEEE Aerosp. Conf.* pp 1–12

[3] Dalal N and Triggs B 2005 Histograms of oriented gradients for human detection *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)* pp 886–893

[4] Yasmeen A, Iqbal M J, Ahmad I and Khan A 2023 Unmanned aerial vehicle detection: a survey *Multimed. Tools Appl.* **82** 15461–15498

[5] Kassab M, Khalil A and Azar A T 2023 Counter-UAV systems: a comparative analysis of existing approaches *Proc. Int. Conf. Artif. Intell. Comput. Sci.* pp 1–8

[6] Du J, Song G and Gui G 2023 Anti-drone detection of low, slow and small unmanned aerial vehicles using background subtraction and F-YOLOv7 *Electronics* **12** 4266

[7] Gao Y, Zhang X and Wang Y 2024 Lightweight anti-UAV detection based on improved YOLOv11 *Remote Sens.* **16** 3982

[8] Munir A, Siddiqui A J and Anwar S 2024 YOLO-RAW: real-time anti-drone system in adverse weather *IEEE Trans. Aerosp. Electron. Syst.* **60** 4521–4533

[9] Zhou J, Chen H and Xu Z 2024 LMWP-YOLO: a lightweight multi-scale drone detection model with multi-dimensional collaborative attention *Electronics* **13** 2271

[10] Zheng Y, Li K and Zhang W 2024 YOLO-DD: ghost dynamic convolution-based lightweight UAV detection framework *Drones* **8** 412

[11] Jiang N, Wang K, Peng X, Yu X, Wang Q, Xing J, Li G, Zhao G, Zhao M, Li C and others 2021 Anti-UAV: a large-scale benchmark for vision-based UAV detection and tracking *IEEE Trans. Multimed.* **23** 4275–4287

[12] Zheng Y, Fu X and You Y 2022 Det-Fly: a challenging benchmark and model for UAV detection *Drones* **6** 166

[13] Sun Y, He Z, Cao J and Li X 2022 AirBirds: a large-scale challenging dataset for bird strike prevention in real-world scenarios *Proc. Int. Joint Conf. Artif. Intell.*

(IJCAI) pp 1417–1423

[14] Kakaletsis E, Tzelepi M and Charalampous K 2025 SynDroneVision: a synthetic dataset for multi-spectral UAV detection *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)* pp 2891–2900

[15] Wang A, Chen H, Liu L, Chen K, Lin Z, Han J and Ding G 2024 YOLOv11: Ultralytics YOLO series *Ultralytics GitHub Repository* `https://github.com/ultralytics/ultralytics`

[16] Gu A and Dao T 2023 Mamba: linear-time sequence modeling with selective state spaces *arXiv preprint* arXiv:2312.00752

[17] Liu Y, Tian Y, Zhao Y, Yu H, Xie L, Wang Y, Ye Q and Liu Y 2024 VMamba: visual state space model *Adv. Neural Inf. Process. Syst. (NeurIPS)* **37** 103031–103052

[18] Shi L, Zhang T, Feng C, Liu Y and Luo J 2024 TransNeXt: robust foveal visual perception for vision transformers *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* pp 17773–17783

[19] Lin T Y, Dollar P, Girshick R, He K, Hariharan B and Belongie S 2017 Feature pyramid networks for object detection *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)* pp 2117–2125

[20] Liu S, Qi L, Qin H, Shi J and Jia J 2018 Path aggregation network for instance segmentation *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)* pp 8759–8768

[21] Tan M, Pang R and Le Q V 2020 EfficientDet: scalable and efficient object detection *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* pp 10781–10790

[22] Guo M H, Lu C Z, Liu Z N, Cheng M M and Hu S M 2023 Visual attention network *Comput. Vis. Media* **9** 733–752

[23] Zhao H, Wang C and Zhang X 2024 MDFD2-DETR: multi-domain frequency decomposition detection transformer for remote sensing *Remote Sens.* **16** 2814

[24] Li Z, Wang H and Liu X 2024 CNCM and CSSC: novel backbone improvements for small object detection in UAV imagery *Proc. IEEE Int. Conf. Image Process. (ICIP)* pp 3521–3526

[25] Tian H, Sun M and Zhou H 2024 GCConv: graph-context convolution for small drone target detection *IEEE Trans. Neural Netw. Learn. Syst.* **35** 9871–9882

[26] Chen Y, Li Q and Zhang W 2024 ES-MoE and YOLO-Master: efficient mixture-of-experts for lightweight drone detection *Pattern Recognit.* **156** 110789

[27] Zhang P, Wang L and Liu F 2024 CAFM: context-aware feature merging for multi-scale object detection *IEEE Access* **12** 45231–45242

[28] He K, Li Y and Chen X 2024 CAM: channel aggregation module for feature pyramid enhancement *Neurocomputing* **598** 128127

[29] Li C, Li H, Jiang H, Weng K, Geng Y, Li L, Ke Z, Li Q, Cheng M, Nie W and others 2022 YOLOv6: a single-stage object detection framework for industrial applications *arXiv preprint* arXiv:2209.02976

[30] Wang Y, Wu J and Huang Z 2024 LCA: lightweight cross-scale attention for real-time UAV detection *Appl. Intell.* **54** 7823–7836

[31] Wang A, Chen H, Xu M, Liao Z, Lu T, Chu J, Chen H, Lin Z, Han J and Ding G 2024 YOLOv10: real-time end-to-end object detection *Adv. Neural Inf. Process. Syst. (NeurIPS)* **37** 107038–107057

[32] Li Z, Wang C, Yu F and Xie X 2024 YOLOv26n: an ultra-lightweight detector for UAV small target applications *Drones* **8** 589

[33] Fu Y, Huang B, Liu H and Wang Y 2024 UAV-DETR: detection transformer for UAV small target detection in complex backgrounds *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* pp 15923–15932

[34] Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, Liu Y and Chen J 2024 DETRs beat YOLOs on real-time object detection *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* pp 16965–16974

[35] Sun B, Liu X and Zhang Y 2024 FBRT-YOLO: feature-balanced real-time tiny object detection for counter-UAV *Remote Sens.* **16** 4012

[36] Wang T, Liang Z and Chen Y 2024 YOLO-TLP: top-level prior augmentation for small drone detection *IEEE Trans. Instrum. Meas.* **73** 5027413

[37] Jocher G, Chaurasia A and Qiu J 2023 Ultralytics YOLOv8 *GitHub Repository* `https://github.com/ultralytics/ultralytics`