remote sensing

MDPI

Article

EFPNet: An Efficient Feature Perception Network for Real-Time Detection of Small UAV Targets

Jiahao Huang [1,2,3], Wei Jin [1,2,3,*], Huifeng Tao [1,2,3], Yunsong Feng [1,2,3], Yuanxin Shang [1,2,3], Siyu Wang [1,2,3] and Aibing Liu [1,2,3]

[1] College of Electronic Engineering, National University of Defense Technology, Hefei 230037, China; huangjiahao@nudt.edu.cn (J.H.); taohf128@nudt.edu.cn (H.T.); fengyunsong17@nudt.edu.cn (Y.F.); shang-yuanxin19@nudt.edu.cn (Y.S.); wangsiyu@nudt.edu.cn (S.W.); liuaibing1017@nudt.edu.cn (A.L.)
[2] Anhui Province Key Laboratory of Electronic Environment Intelligent Perception and Control, Hefei 230037, China
[3] Advanced Laser Technology Laboratory of Anhui Province, Hefei 230037, China
* Correspondence: jinwei17@nudt.edu.cn

Highlights

What are the main findings?

- We propose a novel dual-branch attention mechanism based on RT-DETR and innovatively construct a feature pyramid structure that utilizes deep features to guide the encoding and fusion of shallow features. This design overcomes the limitations of multihead attention mechanisms in unified feature modeling and significantly enhances the perception of fine-grained texture features of small UAV targets.
- On the TIB and drone-vs-bird datasets, EFPNet achieved mAP50 scores of 94.1% and 98.1%, surpassing the baseline model by 3.2% and 1.9%, respectively. Notably, the model achieves these improvements with reduced parameters and computational costs, effectively addressing issues related to feature ambiguity and background interference in complex scenarios.

What are the implications of the main findings?

- The high precision and efficiency of EFPNet provide robust technical support for meeting the practical demands of small UAV detection in complex aerial environments.
- This model establishes a novel, high-precision technical paradigm for real-time small-object detection tasks, offering a reliable solution for all-weather airspace security, intelligent surveillance, and low-altitude defense systems.

Abstract

In recent years, unmanned aerial vehicles (UAVs) have become increasingly prevalent across diverse application scenarios due to their high maneuverability, compact size, and cost-effectiveness. However, these advantages also introduce significant challenges for UAV detection in complex environments. This paper proposes an efficient feature perception network (EFPNet) for UAV detection, developed on the foundation of the RT-DETR framework. Specifically, a dual-branch HiLo-ConvMix attention (HCM-Attn) mechanism and a pyramid sparse feature transformer network (PSFT-Net) are introduced, along with the integration of a DySample dynamic upsampling module. The HCM-Attn module facilitates interaction between high- and low-frequency information, effectively suppressing background noise interference. The PSFT-Net is designed to leverage deep-level features to guide the encoding and fusion of shallow features, thereby enhancing the model's capability to perceive UAV texture characteristics. Furthermore, the integrated DySample dynamic upsampling module ensures efficient reconstruction and restoration of feature representations. On the TIB and Drone-vs-Bird datasets, the proposed EFPNet achieves

Check for updates

Academic Editors: Michal Aibin and Ryan Rad

Received: 7 December 2025
Revised: 16 January 2026
Accepted: 17 January 2026
Published: 20 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

mAP50 scores of 94.1% and 98.1%, representing improvements of 3.2% and 1.9% over the baseline models, respectively. Our experimental results demonstrate the effectiveness of the proposed method for small UAV detection.

Keywords: attention mechanism; DySample; pyramid sparse feature transformer; RT-DETR; small UAV

1. Introduction

With the rapid advancement of unmanned aerial vehicle (UAV) technology, UAVs have demonstrated remarkable advantages—including high maneuverability, compact structures, and low operational costs—in complex environments, hazardous zones, and large-scale missions [1], establishing themselves as indispensable technological platforms. However, as UAV applications diversify and their numbers surge, the accompanying social concerns—such as airspace security threats and potential privacy infringements [2]—have rendered the development of efficient UAV detection systems an urgent necessity.

Currently, detection technologies targeting "low, slow, and small" unmanned aerial vehicle targets primarily encompass radar, radio frequency (RF), acoustic, and visual detection methods [3]. Traditional radar technology offers the distinct advantage of all-weather operation, localizing targets by transmitting electromagnetic waves and receiving return echoes; however, when detecting small UAVs with minimal radar cross sections, the return signals are susceptible to being overwhelmed by complex background clutter, such as that from buildings or trees, resulting in a high missed detection rate. Radio frequency detection functions by intercepting communication link signals between the UAV and the ground control station, yet it frequently proves ineffective against UAVs operating under radio silence or in autonomous flight modes. Although acoustic detection systems are cost-effective, they are highly susceptible to environmental noise interference and are constrained by a limited detection range. In contrast, visual sensors are capable of acquiring rich appearance and texture information about targets, offering significant advantages such as intuitiveness, high observability, and relatively low cost; furthermore, they can effectively complement other sensor modalities or even independently perform high-precision identification tasks in specific close-range scenarios. Consequently, the vision-based detection of unauthorized UAVs remains a prominent research hotspot.

In recent years, deep learning techniques have found widespread application in the field of object detection. UAV detection based on convolutional neural networks (CNNs) has become the mainstream approach, as it enables automatic learning and extraction of complex features from large-scale datasets without manual feature engineering, thereby offering superior adaptability. Despite the remarkable progress achieved by deep-learning-based detectors, their application to UAV-target detection in complex real-world environments still faces several critical challenges. The primary challenge lies in the inherently small sizes of UAV targets, which occupy only a minimal number of pixels within an image. This makes it difficult to effectively extract discriminative features, often resulting in confusion with background clutter or other moving objects such as birds. Furthermore, when UAVs operate within complex backgrounds, the strong similarity between their texture patterns and those of the surroundings substantially increases the difficulty of extracting reliable target features for detection. To address these challenges, previous studies, e.g., [4–7], have primarily focused on enhancing deep-level features to improve detection accuracy. However, during the processes of feature extraction and multilayer transmission, crucial information pertaining to small targets tends to be progressively lost due to the target's limited scale and weak semantic representation. The capability of deep features to represent

the fine-grained details of small targets is inherently constrained, resulting in incomplete preservation of small-target information within deeper layers. Consequently, relying solely on deep feature enhancement is insufficient to compensate for this information loss, thereby limiting further improvement in small-object detection performance.

To address the aforementioned challenges, this paper proposes an efficient feature perception network (EFPNet) tailored for the detection of small UAV targets. First, the proposed approach innovatively integrates the ConvMixer [8] module with the HiLo attention mechanism [9], achieving reduced model complexity while preserving the network's ability to extract discriminative features of small targets. Second, a dynamic upsampling module is introduced to replace conventional upsampling methods, thereby enhancing the restoration of fine-grained feature details. Finally, an enhanced cross-scale feature interaction network is developed, which leverages deep-layer features to guide the fusion of shallow-layer representations, ultimately improving target detection performance. The main contributions of this study are summarized as follows:

(1) A HiLo-ConvMix Attention (HCM-Attn) module is designed, wherein the ConvMix structure replaces the Lo branch of the HiLo attention mechanism to capture global structural information. By leveraging a dual-branch architecture, the module enables effective interaction with high-frequency information, allowing the network to better focus on target details while mitigating background noise interference.

(2) The DySample dynamic upsampling method is employed to replace conventional upsampling operations. By performing adaptive sampling at each spatial position, this approach enables efficient reconstruction and restoration of feature representations.

(3) A multilayer cross-scale feature fusion approach, termed PSFT-Net, is proposed. This method employs a cross-attention mechanism to construct a hierarchically sparse fusion Transformer that effectively preserves shallow feature information. It further utilizes deep-layer representations to identify and select the most salient low-level features, thereby achieving targeted local feature enhancement. This feature fusion strategy substantially strengthens the model's perception of target texture characteristics.

2. Related Work and Challenges

2.1. Advances in Object-Detection Algorithms and UAV Detection

Driven by the rapid advancement of artificial intelligence technologies, numerous deep-learning-based object-detection algorithms have been proposed. Currently, mainstream object detection methods can be broadly categorized into two types: two-stage and one-stage algorithms. Two-stage algorithms, such as Faster R-CNN [10] and Mask R-CNN [11], exhibit superior detection accuracy; however, their relatively high inference latency limits their suitability for applications requiring real-time performance. In contrast, one-stage algorithms, represented by the Single Shot MultiBox Detector (SSD) and You Only Look Once (YOLO) families, emphasize efficiency and real-time detection. SSD extracts multi-scale features from images and generates bounding boxes of various sizes at each feature level, enabling effective cross-scale object detection [12]. Meanwhile, YOLO directly predicts object categories and bounding box coordinates without the need for a proposal generation stage, making it more suitable for real-time applications [13]. Nevertheless, despite the remarkable progress in both detection performance and efficiency, these algorithms share a common drawback—reliance on handcrafted components. For instance, YOLO detectors employ non-maximum suppression (NMS) in the post-processing stage to eliminate redundant bounding boxes, which not only reduces inference speed but also introduces additional hyperparameters that may affect detection accuracy [14].

To mitigate the adverse effects introduced by handcrafted components, Zhao et al. proposed RT-DETR [15], which employs an efficient hybrid encoder to decouple the interaction

and fusion of multi-scale features. This design not only significantly reduces computational redundancy but also effectively eliminates the inference delay caused by non-maximum suppression (NMS). Therefore, this paper improves upon the RT-DETR framework to better adapt it for UAV detection tasks.

When UAVs appear at low resolutions in images, effectively extracting features from complex backgrounds and achieving accurate prediction becomes a critical challenge to overcome. To address this, Saribas [16] employed the YOLOv3 detector as a target position initializer. When the Kernelized Correlation Filter (KCF) tracker loses the target, the detector re-localizes it, thereby enhancing UAV tracking robustness. Zhai [17] integrated shallow texture information from low-resolution feature maps into the P2 detection layer of the YOLOv8 network, improving the model's perception of small targets. However, this modification also increased computational cost by more than 20%. Yang [18] utilized a cross-correlation neural network to capture key temporal information in the audio domain and combined it with image feature extraction via a ResNet50 backbone to achieve audio–visual fusion-based UAV detection. Similarly addressing multimodal sensing, Zhang [19] proposed an adversarial-learning-based differential feature perception network; by exploiting the complementary differential features between infrared and visible light images, this approach significantly enhanced detection robustness in complex environments. Hao [20] introduced an OmniKernel module into the YOLOv8 framework to fuse features from the B1 layer of the backbone and the N2 layer of the neck network, endowing the model with enhanced global perception and improving small-target detection performance. Dadrass Javan [21] modified the number of convolutional layers in the YOLOv4 network to extract more accurate and detailed semantic features, thereby achieving improved detection performance. He [22] incorporated the BiFormer attention mechanism into a modified YOLOv7 architecture, aggregating features from large receptive fields to enhance UAV detection under foggy conditions.

Overall, the inherent complexity of aerial scenes often leads to false or missed detections in UAV recognition tasks. Therefore, this study employs datasets collected under diverse and complex scenarios to evaluate the generalization capability of the proposed improved algorithm.

2.2. Attention Mechanisms

Attention mechanisms are commonly employed in object detection tasks to capture discriminative features. By strengthening the focus on key feature regions, attention mechanisms enhance detection accuracy and improve the generalizability of models [23]. For example, Wen [24] incorporated Coordinate Attention (CA) into the YOLOv5 framework, whereas Dai [25] embedded three different attention mechanisms into the detection head, significantly enhancing its feature representation capability. Xin [26] proposed a co-existing attention mechanism for few-shot detection, enabling attention modules to coexist and share information under limited sample conditions. Wang [27] applied the Convolutional Block Attention Module (CBAM) [28] to shallow feature maps in YOLOX, allowing the model to better focus on small-target information. Furthermore, addressing the challenges of low contrasts and complex backgrounds associated with small targets in infrared imagery, Zhang [29] proposed Deep-IRTarget; this approach utilizes a dual-domain feature extraction strategy that integrates channel and spatial attention to significantly enhance target feature representation while effectively suppressing background noise. In this work, a small-object-oriented attention mechanism is adopted to replace the multihead self-attention (MSA) within the attention-based internal scale feature interaction (AIFI) module, aiming to enhance the model's focus on UAV-specific features.

3. Methodology

3.1. Overall Framework of the Network

RT-DETR is a real-time, end-to-end object detection network with a hybrid architecture composed of a backbone network, an efficient hybrid encoder, and an auxiliary prediction decoder. However, this approach presents certain limitations when applied to UAV small-object detection tasks. First, the original backbone network, HgNetv2, demonstrates limited capability in capturing the fine-grained local details of UAV targets, while its relatively high computational cost negatively affects the real-time detection efficiency. Second, the AIFI module operates on high-level feature maps where detailed information is inherently sparse. However, AIFI employs a uniform global attention mechanism, preventing sufficient exploitation and enhancement of the limited detailed features and constraining the model's ability to represent small UAV targets effectively. Finally, although the cross-scale feature-fusion (CCFF) module achieves efficient cross-scale fusion through a lightweight convolutional design, the fine-grained information of small targets gradually attenuates during inter-scale transmission. Moreover, noise and target features may become intertwined in lower-level feature maps, further exacerbating false and missed detections in UAV detection tasks.

To address the aforementioned limitations, this paper proposes an enhanced feature perception network (EFPNet) designed for UAV detection across diverse scenarios. The model improves both detection accuracy and inference speed by replacing the original backbone network. In addition, distinct attention mechanisms are applied to high- and low-frequency information to construct an encoder capable of capturing both local detail and global contextual dependencies. Subsequently, the DySample module is introduced to replace conventional upsampling operations, enabling more accurate reconstruction of fine-grained feature details. Finally, a cross-scale feature fusion structure (PSFT-Net) is proposed to effectively promote multilevel feature integration and enhance the model's focus on small-target characteristics. The overall architecture of EFPNet is illustrated in Figure 1.

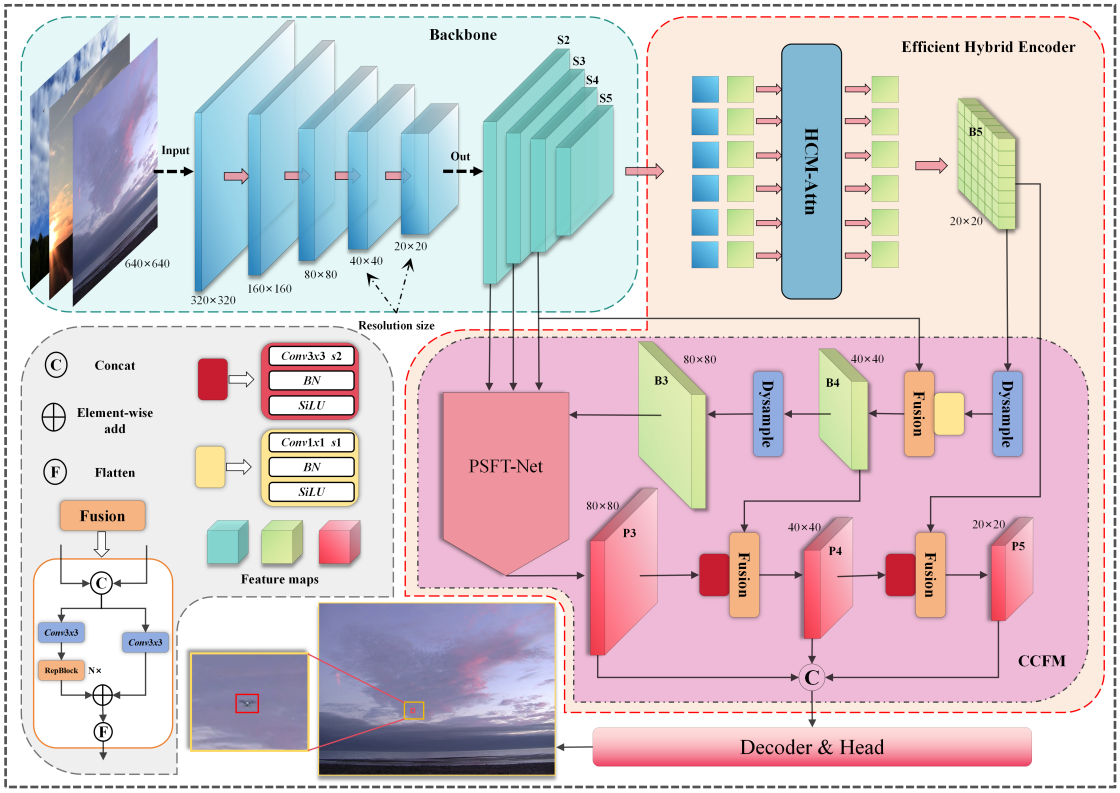


Figure 1. Improved network architecture diagram.

3.2. HiLo-ConvMix Attention (HCM-Attn) Module

According to spatial–frequency and scale–space theories [30], edges and fine-grained texture features in images essentially originate from strong local intensity variations that are primarily concentrated in high spatial frequency components. In contrast, low-frequency components encode the global structure and macroscopic layout information of the image. However, the MSA in the AIFI module adopts a uniform global attention strategy, modeling all spatial locations indiscriminately. In UAV detection scenarios, this mechanism exhibits notable shortcomings: MSA fails to effectively distinguish between high-frequency detail information and low-frequency global structure. This leads to excessive smoothing of local detail features and inefficient allocation of computational resources for global information modeling.

To address the limitations of MSA, specifically its neglect of frequency-diverse features and high computational overhead, Pan introduced the HiLo dual-branch attention mechanism. This architecture decouples high-frequency details from low-frequency contours via a dual-path design: the Hi branch employs local window-based self-attention to finely encode high-frequency texture information, whereas the Lo branch captures global low-frequency structural features by performing average pooling on keys and values. However, although the original Lo branch reduces dimensionality via pooling, it fundamentally remains a global interaction-based attention mechanism, which inevitably incurs a computational complexity of $O(N^2)$ when processing high-resolution feature maps. Furthermore, in the specific context of UAV detection tasks, excessive reliance on global attention mechanisms is prone to introducing background noise. In light of this, we propose replacing the Lo branch with a large-kernel ConvMixer structure; this approach not only efficiently extracts global spatial structures with a linear complexity of $O(N)$ but also leverages inherent convolutional properties to effectively suppress background interference, thereby achieving more robust low-frequency information modeling.

Based on the above analysis, inspired by the HiLo attention mechanism and ConvMixer, this paper proposes the HiLo-ConvMix Attention module. The traditional MSA is decoupled into two functionally complementary branches: the Hi branch performs self-attention within each local window to specifically encode high-frequency detail information, while the Lo branch discards global attention operations and leverages ConvMixer to efficiently capture global structural information. This design preserves the precise encoding of local detail features while significantly reducing the computational overhead of global relation modeling, thereby improving UAV-target detection accuracy. The structure of the proposed module is illustrated in Figure 2.

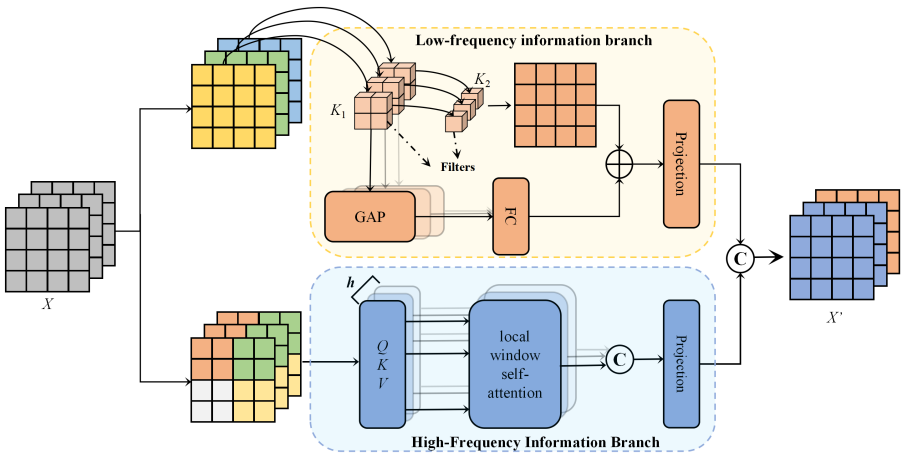


Figure 2. HCM-Attn architecture. Filters denote the convolution kernel type; *h* represents the total number of attention heads.

Given that UAV targets in images are primarily characterized by high-frequency detailed features such as edges and textures, the Hi branch employs a local window attention mechanism to achieve precise extraction of high-frequency information; this design enhances the model's perception capability for fine-grained features while effectively suppressing background noise interference. The specific procedure is as follows: after the feature map X extracted by the backbone network is fed into the HCM-Attn module, the downstream path employs a window-based attention mechanism (with window size w) to mine fine-grained texture and high-frequency information. Assuming that the input feature map is denoted as $X \in \mathbb{R}^{d_{model} \times H \times W}$, the dimensionality of each attention head is $d_{head} = d_{model}/h$. Then, the output can be calculated using the following formula:

$$\begin{cases} H_{out} = Concat(head_1, head_2, \ldots, head_n)W^O, \\ head_i = Merge(Attention_i), \\ Attention_i = softmax(Q_i K_i^T / \sqrt{d_{head}})V_i. \end{cases} \tag{1}$$

where $Q_{w_i}, K_{w_i}, V_{w_i} \in \mathbb{R}^{d_{model} \times w \times w}$ denote the linear matrix corresponding to query, key, and value for the i-th window, while $W^O \in \mathbb{R}^{d_{model} \times d_{model} \times d_{model}}$ represents the output projection weight matrix.

The Lo branch is designed to efficiently model low-frequency global structures by employing a decoupled spatial and channel mixing strategy, serving as an effective alternative to the conventional explicit global attention mechanism. Initially, the feature map X undergoes spatial mixing via a large convolution kernel $K_1 \in \mathbb{R}^{d_{model} \times k \times k}$; according to effective receptive field theory, large-kernel convolutions establish extensive inter-pixel dependencies and function as physical low-pass filters, implicitly aggregating long-range spatial context information while smoothing high-frequency texture noise to capture the macroscopic layout and background structure of the image. Subsequently, point-wise convolution (with a kernel size of $K_2 \in \mathbb{R}^{d_{model} \times 1 \times 1}$) is employed for channel mixing; this step facilitates semantic interaction and feature reorganization across different channels while preserving the spatial structure, ultimately enhancing the model's representational capacity. The computational process is formulated as follows:

$$\begin{cases} S = DWConv(X, K), \\ Y = \sigma(PWConv(S)). \end{cases} \tag{2}$$

in the equation, $S \in \mathbb{R}^{d_{model} \times H_1 \times W_1}$ and $Y \in \mathbb{R}^{d_{model} \times H_1 \times W_1}$ denote the output features and σ represents the activation function. Furthermore, to compensate for the potential limitations of convolution in capturing absolute global information, this branch incorporates a global compensation mechanism based on Global Average Pooling (GAP) to explicitly inject global contextual priors, specifically

$$g = GAP(S). \tag{3}$$

where $g \in \mathbb{R}^{d_{model} \times 1 \times 1}$. The global information is then transformed through a linear layer, ensuring that each feature point acquires the corresponding global contextual information. The global compensation information is added to the convolutional features to obtain the final output L_{out} of this branch

$$L_{out} = Y + \alpha \cdot (W_g + b_g) \tag{4}$$

here, α denotes a learnable scaling factor, W_g represents the linear transformation matrix, and b_g is the bias term.

This design not only circumvents the high computational complexity associated with global attention but also leverages the inductive bias of convolution to effectively enhance the model's robust perception of the holistic UAV structure within complex backgrounds. Compared to MSA, HCM-Attn ensures superior representational capacity while simultaneously achieving enhanced efficiency and scalability.

3.3. Improvement of the Upsampling Structure

In the RT-DETR model, deep features are first processed through the attention-based AIFI module and subsequently undergo cross-scale interaction and fusion with the multi-scale features from the backbone via the CCFF module. During this process, high-level features typically require upsampling operations. Conventional upsampling methods, such as nearest-neighbor and bilinear interpolation, rely on fixed interpolation rules, which often introduce jagged artifacts along edges—particularly when feature maps are magnified—thus degrading the reconstruction and representation of fine details. To address this issue, this paper introduces the DySample [31] dynamic upsampler, which supports rapid backpropagation while incurring only a minimal increase in training time. The dynamic sampling mechanism and structural design of the DySample module are illustrated in Figure 3.

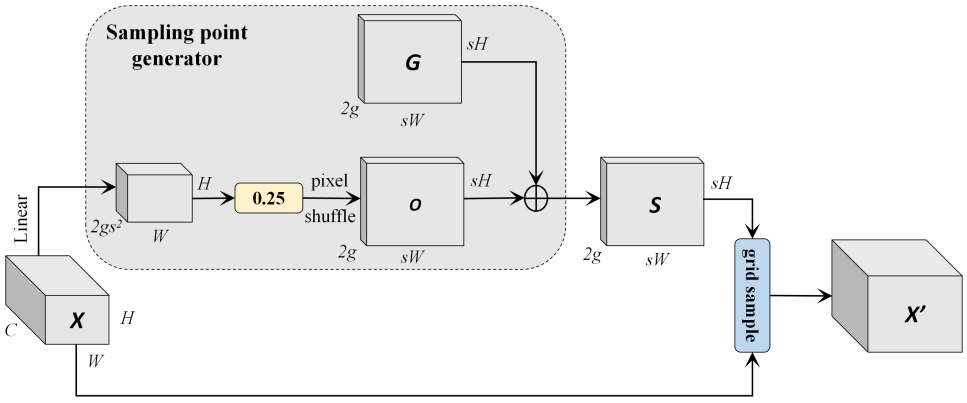


Figure 3. DySample upsampling principle schematic diagram.

Traditional upsampling methods adhere to fixed geometric rules; specifically, for a given position p within the input feature map X, the interpolation result X_p is determined exclusively by the weighted average of pixels located within its spatial neighborhood $N(p)$, which can be formulated as follows:

$$X'_p = \sum_{q \in N(p)} \omega_{pq} \cdot X_q. \tag{5}$$

where the weight ω_{pq} is determined solely by the geometric distance between p and q and remains independent of the underlying feature content. Such a fixed interpolation kernel fundamentally functions as a low-pass filter that tends to smooth feature maps; however, since small UAV targets occupying minimal pixel areas are characterized primarily by high-frequency edge features, this smoothing operation induces severe information attenuation and edge blurring, rendering the targets susceptible to being overwhelmed by background noise during the deep feature recovery process.

Diverging from traditional methods where sampling positions are strictly constrained by geometric grids, DySample dynamically generates sampling offsets contingent upon the content of the input feature map, thereby facilitating content-aware feature reorganization. Specifically, for a task with an upsampling scale factor of s, DySample transcends reliance on fixed neighborhood pixels by employing a lightweight linear layer to predict exclusive

offsets O for each sampling position based on the input features X. Consequently, the resulting set of dynamic sampling positions S can be formulated as follows:

$$S = G + O. \tag{6}$$

Targeting small-scale UAV objects, DySample permits sampling points to actively shift towards high-response target edges or centers; this mechanism effectively circumvents background noise interference during the resampling process and accurately reconstructs fine-grained structures and high-frequency edge information, thereby significantly enhancing the fidelity of feature recovery.

To validate the feature reconstruction capability and robustness of the DySample module under various scale transformations—specifically its effectiveness in recovering fine-grained features at multilevel upscaling rates—we conducted comparative experiments against the traditional bilinear interpolation method. Representative upsampling rates of 2× and 4× were selected to investigate the extent to which different upsampling methods preserve the texture details of small targets when feature map resolutions undergo drastic changes. Figure 4 visually illustrates the feature map visualization results of the two methods under different upscaling rates.

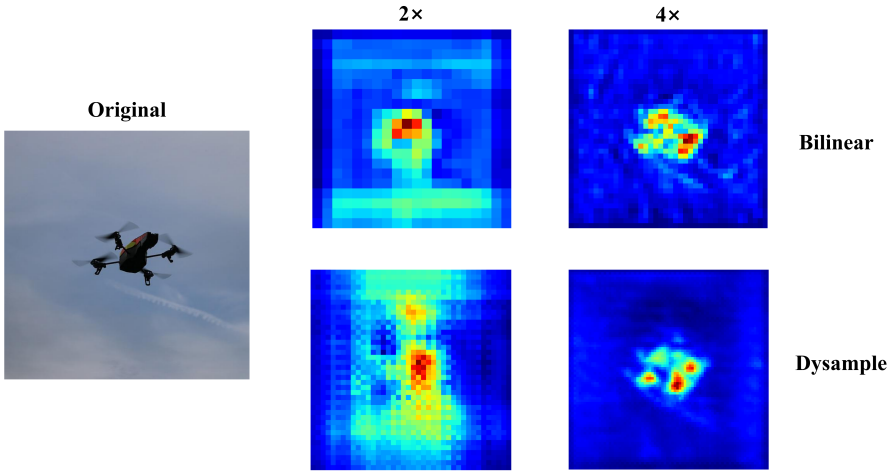


Figure 4. Upsampled feature map comparison.

The visualization results indicate that during the recovery of deep feature maps (at the 2× upsampling stage), the bilinear interpolation method induces distinct blocking artifacts within the UAV target region, primarily due to the low resolution of the feature maps and their richness in highly abstract semantic information. Conversely, even under low-resolution conditions, the DySample method effectively suppresses background noise and concentrates high response values onto the center of the object. Upon further examination of the texture and shape recovery phase (at the 4× upsampling stage), the feature maps generated by the bilinear method exhibit relatively blurred edges, presenting an overall phenomenon of over-smoothing and feature aliasing. In contrast, the DySample method demonstrates a superior boundary preservation capability, generating feature maps with sharp edges that more intuitively resolve the fine-grained structural features of the UAV—preserving structural details such as rotors to a certain extent—thereby validating its robustness in different upsampling tasks.

3.4. Pyramid Sparse Fusion Transformer Network (PSFT-Net)

While RT-DETR achieves an excellent balance between real-time performance and accuracy through the introduction of a hybrid encoder, its feature fusion architecture reveals notable limitations when applied to complex scenarios such as the detection of

small UAVs. Primarily, the existing CCFF module fundamentally relies on convolution operations characterized by local receptive fields and static feature concatenation. This rigid fusion mechanism lacks adaptive capability, rendering it unable to perform dynamic feature recalibration in response to variations in target scale or feature saliency. Secondly, the module exclusively aggregates feature layers S3 through S5, thereby discarding the shallow S2 features that are rich in texture and edge details pertinent to minute targets. This omission of high-resolution information leads to an irreversible loss of critical fine-grained details during the fusion process, severely constraining the model's localization precision for small targets. However, direct integration of the high-resolution S2 feature layer would introduce substantial parameter redundancy and computational overhead. Consequently, to address the aforementioned challenges, this paper designs a Pyramid Sparse Fusion Transformer Network that balances efficiency with accuracy, specifically aiming at efficiently capturing the features of small targets. We process the S2 feature layer with Space-to-Depth Convolution (SPDConv) [32] to obtain features enriched with small-object information, which are then fused with the B3 feature map. This approach mitigates the loss of image details and the inefficiency in feature learning caused by stride convolutions and pooling operations in CNNs. Subsequently, inspired by the concept of cross-attention [33], we construct a Hierarchical Sparse Fusion Transformer (HSFT) that leverages semantic activation from high-level features to guide and enhance shallow feature representation. This design effectively strengthens contextual modeling and feature discriminability while reducing token interaction complexity, enabling efficient cross-scale feature fusion for small-target detection. The detailed architecture of this network is illustrated in Figure 5.

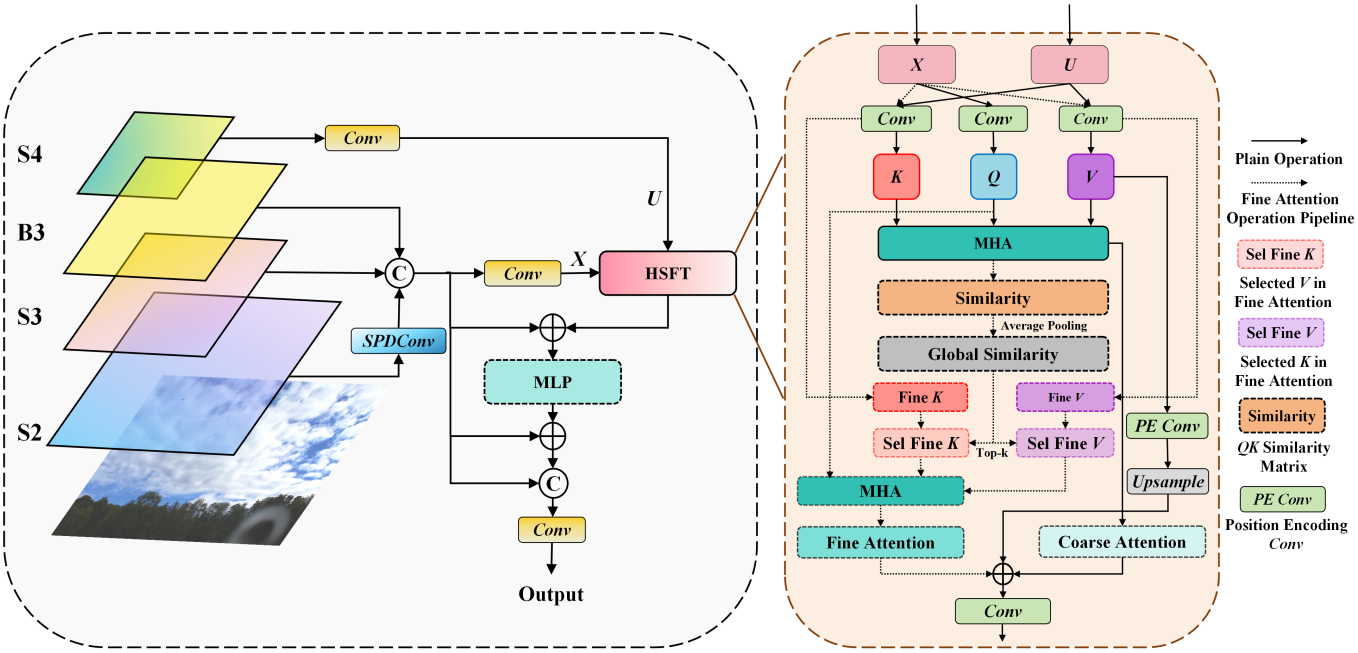


Figure 5. PSFT-Net block diagram. The schematic diagram of HSFT is shown on the right.

Given that the S2 feature layer encapsulates high-resolution spatial details imperative for small-object detection, the direct application of conventional downsampling techniques often precipitates a loss of fine-grained information due to the indiscriminate discarding of pixels, thereby predisposing minute targets to feature blurring. To maximally preserve original pixel information while downscaling feature maps to mitigate computational overhead, we incorporate the SPDConv module. Figure 6 illustrates the architecture of SPDConv, which eliminates the commonly used stride convolution and pooling operations. Instead, it

rearranges spatial feature blocks into the depth dimension, enabling downsampling of the S2 feature map while retaining richer learned information. The resulting features are then concatenated with those from the same hierarchical level to integrate multilevel semantic information, thereby enhancing the network's representational capability.

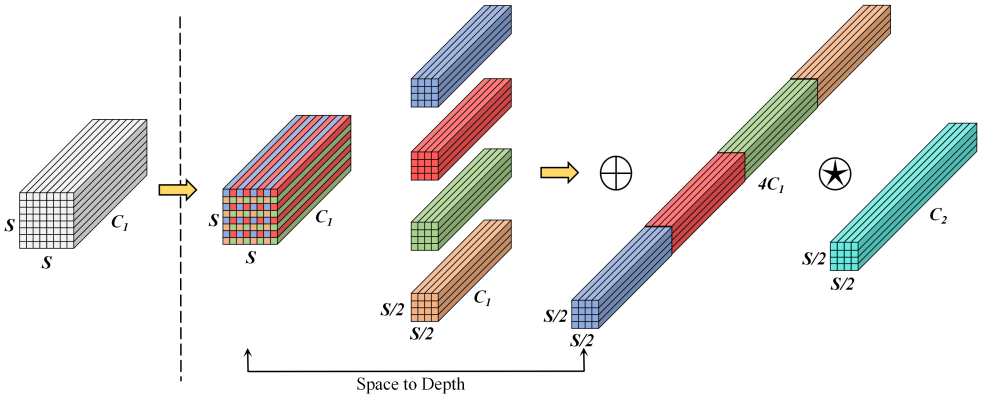


Figure 6. SPDConv downsampling process. The input feature map of size $C_1 \times W \times W$ is divided into four sub-feature maps of size $C_1 \times \frac{W}{2} \times \frac{W}{2}$. These four sub-feature maps are then concatenated to form an intermediate feature representation, which is subsequently refined through a 1×1 convolution layer to further integrate the features and produce the final output feature map.

Although SPDConv effectively preserves the high-resolution spatial information of the S2 layer, subsequent integration with deep features via mere linear concatenation or summation often fails to resolve the semantic and spatial distributional misalignments inherent in cross-scale features, thereby constraining fusion efficacy. To address this, HSFT introduces a semantic-guided "Coarse-to-Fine" dynamic sparse attention mechanism. Initially, HSFT maps deep features into keys and values to serve as high-level semantic guides, performing a global scan over the shallow features that act as queries. This process facilitates the contextual injection of deep semantics into shallow pixels, thereby endowing high-resolution features with global perception capabilities. Subsequently, diverging from static fusion approaches that apply uniform processing across all regions, HSFT leverages generated attention heatmaps to dynamically select the $Top - k$ high-response regions (representing potential small targets); it restricts fine-grained feature interaction and enhancement to these critical areas, thereby automatically filtering background noise during the fusion process. Finally, the module performs adaptive weighted aggregation of the coarse-grained global semantic flow, the fine-grained local texture flow, and spatial positional information. This fusion mechanism not only preserves the geometric details of minute targets from the S2 layer but also integrates deep information, thereby equipping the model with robust semantic discriminability.

Specifically, this module designates feature layer $X \in \mathbb{R}^{C \times H \times W}$ as the query and the adjacent feature layer $U \in \mathbb{R}^{C \times H/2 \times W/2}$ as the key and value; both inputs initially undergo linear transformation to be projected into a shared feature subspace. Subsequently, the module scans the shallow features utilizing the global contextual information from the deep features to inject high-order semantics. Owing to the reduced resolution of the key–value pairs, this step decreases the token interaction complexity to $\frac{1}{4}O(N^2)$, with the coarse-grained attention interaction features calculated as follows:

$$X_{coarse} = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

Subsequently, to achieve dynamic selective fusion, we initially generate a global importance heatmap s by averaging the attention scores along the query dimension, thereby quantifying the semantic contribution of each deep token to the shallow features.

$$s = mean\left(softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)\right) \tag{8}$$

We select the $Top-k$ indices with the highest response values to form the set I. Based on I, we extract the corresponding fine-grained tokens (K^{sel}_{fine}, V^{sel}_{fine}) from the original keys and values via the Gather operation and execute fine-grained sparse attention exclusively on these high-value regions. This procedure ensures that computational resources are highly concentrated on potential small-target regions:

$$X_{fine} = softmax\left(\frac{Q(K^{sel}_{fine})^T}{\sqrt{d_k}}\right)V^{sel}_{fine} \tag{9}$$

Finally, to achieve multi-granularity feature aggregation, we synergistically integrate three distinct information streams: $X_{cosarse}$, which encapsulates global semantics; X_{fine}, which captures local details; and the spatial positional context provided by Convolutional Position Encoding (CPE) [34]. The final fused output X' is obtained as follows:

$$X' = Conv_{1\times1}(X_{coarse} + X_{fine} + Upsample(DConv_{7\times7}(V))) \tag{10}$$

here, CPE utilizes depth-wise convolution to implicitly encode spatial positional information, ensuring strict spatial alignment of the fused features. This furnishes the subsequent detection heads with feature representations that synergize global semantics with local precision.

4. Experiment and Result Analysis

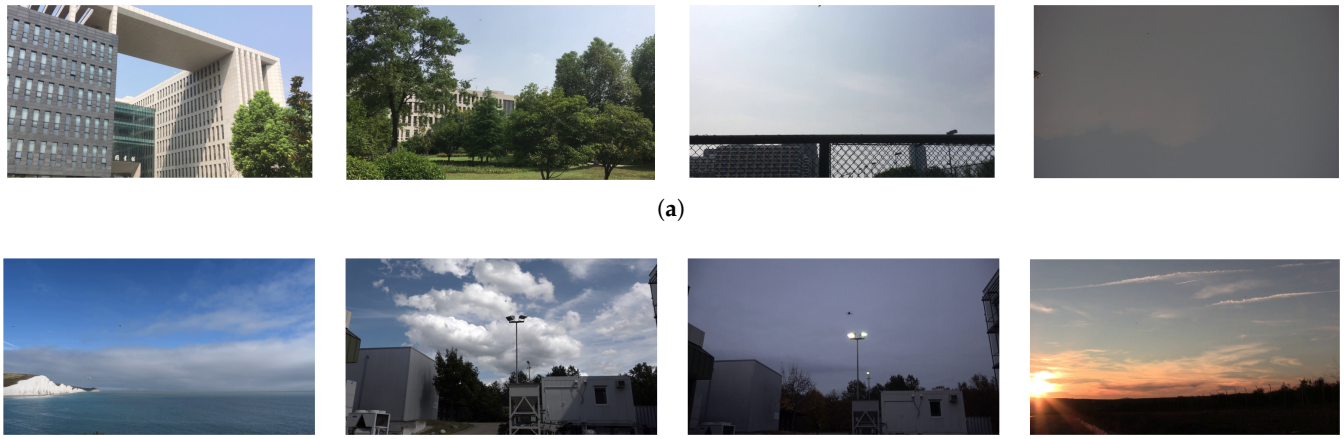
4.1. Dataset Description

The TIB UAV dataset [35] consists of 2850 image samples covering various types of UAVs, primarily including multi-rotor and fixed-wing categories. Representative example images are shown in Figure 7a. The dataset includes image samples captured under various illumination conditions and contains challenging cases such as extremely small targets, motion blur, and complex background environments. The dataset is divided into three subsets: a training set with 1994 samples, a validation set with 428 samples, and a test set with 428 samples.

The Drone-vs-Bird UAV dataset [36] contains 77 video sequences encompassing seven types of UAVs. Sample frames are illustrated in Figure 7b. The dataset covers samples captured under diverse environmental conditions, including sky and vegetation backgrounds, varying weather situations, and complex illumination with direct sunlight and glare interference, as well as data obtained using different camera parameter settings. To reduce computational overhead, we systematically extracted 2168 representative frames from the original video data and divided them into three subsets: 1517 images for training, 325 images for validation, and 325 images for testing. This provides sufficient data support for subsequent algorithm validation and performance evaluation.

In this study, the aforementioned datasets are utilized to validate the effectiveness and generalization capability of the proposed improved model. As shown in Figure 8, the pixel-size distribution of UAVs in the datasets indicates that the vast majority of UAV targets occupy regions smaller than 25 × 25 pixels. According to the MS COCO definition [37], targets occupying an area of 32 × 32 pixels or less in an image are categorized as small

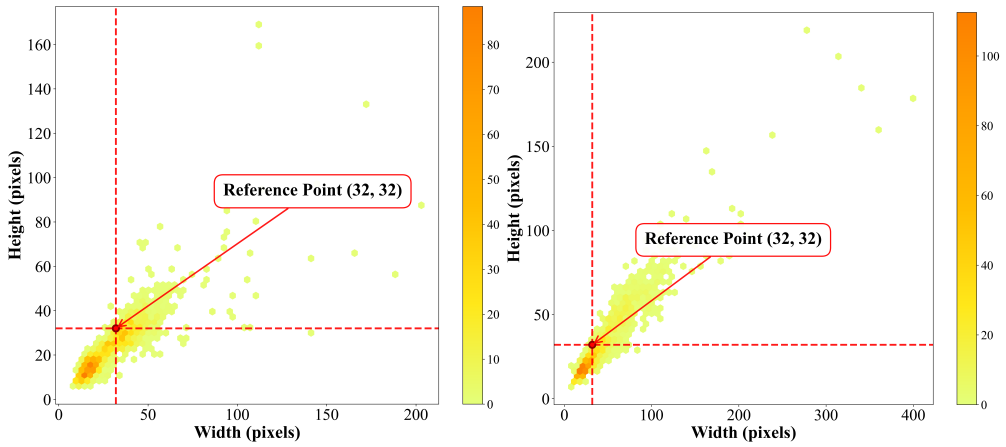
objects. This indicates that under small-object conditions, the model must possess strong recognition and localization capabilities. Therefore, these datasets are particularly suitable for evaluating the proposed method's performance in UAV small-object detection.



(**a**)

(**b**)

Figure 7. Sample images from different scenarios in the TIB and Drone-vs-Bird datasets. (**a**) TIB; (**b**) Drone-vs-Bird.



(**a**) (**b**)

Figure 8. TIB and Drone-vs-Bird Dataset: drone pixel size distribution (darker colors indicate higher quantities). (**a**) TIB; (**b**) Drone-vs-Bird.

4.2. Experimental Setup and Evaluation Metrics

This experiment is conducted on a system equipped with an NVIDIA GeForce RTX 3060 GPU and an Intel Core i7-12700 CPU. All models are trained from scratch. The remaining key constant training parameters are presented in Table 1.

Table 1. Training and model parameters.

Parameter	Value
Epoch	200
Batch size	8
Input size	640 × 640
Optimizer	AdamW
Weight decay coefficient	0.0001
Momentum	0.9
Initial learning rate	0.0001

To evaluate the effectiveness of the EFPNet model, several standard metrics commonly used in object detection tasks were adopted for comparative analysis. Precision (P) represents the ratio of correctly predicted targets to all detected targets. Recall (R) denotes the ratio of correctly detected targets to all actual targets present in the dataset. The calculation formulas are as follows:

$$\begin{cases} P = \frac{TP}{TP+FP}, \\ R = \frac{TP}{TP+FN}. \end{cases} \tag{11}$$

where TP denotes true positives (correct detections), FP denotes false positives (incorrect detections), and FN represents false negatives (missed detections).

The Average Precision (AP) refers to the area under the precision–recall curve. The mean Average Precision (mAP) represents the average of AP values across all categories. Two mAP metrics are employed in this paper: mAP50:50 (with an IoU threshold of 0.5) and mAP50 (averaged across IoU thresholds ranging from 0.5 to 0.95). Here, k denotes the number of evaluated categories, and mAP is computed as

$$mAP = \frac{\sum_{j=1}^{k} AP(j)}{k} \tag{12}$$

GFLOPs and Params are used to assess model complexity, while FPS (Frames Per Second) is employed to evaluate detection speed.

4.3. Backbone Network Selection

Feature extraction plays a crucial role in determining model performance. Therefore, in this section, the original backbone network (HgNetv2) of RT-DETR is replaced with several popular architectures—FasterNet [38], EfficientViT [39], ResNet34, ResNet50, and ResNet18—and the model evaluated on the TIB UAV dataset.

As shown in Table 2, when ResNet50 is used as the backbone, it achieves the highest detection accuracy but also exhibits the largest GFLOPs and parameter count, resulting in reduced detection efficiency. In contrast, FasterNet and EfficientViT demonstrate clear advantages in terms of computational cost and model size; however, their mAP50 and mAP50:95 scores are relatively lower. Considering both accuracy and efficiency, ResNet18 was ultimately selected as the backbone network for RT-DETR, serving as the baseline model for subsequent experiments.

Table 2. Comparison of different backbone networks on the TIB dataset.

Backbone	P (%)	R (%)	mAP50 (%)	mAP50:95 (%)	GFLOPs	Params	FPS
HgNet	88.7	88.3	84.7	33.7	103.8	32,148,396	46.2
FasterNet	87.9	86.9	81.5	31.7	28.8	9,138,516	126.2
EfficientVit	89.8	88.5	83.5	33.4	27.6	10,804,048	107.2
ResNet50	**92.8**	**92.9**	**91.4**	38.6	129.9	42,118,508	39.0
ResNet34	91.1	91.8	88.9	37.5	89.1	31,227,972	54.9
ResNet18	92.0	91.8	90.9	**40.1**	57.3	**19,974,480**	84.7

Note: The bold values indicate the best results.

4.4. Validation of the Effectiveness of HCM-Attn

To rigorously validate the rationale behind substituting the conventional global self-attention mechanism with ConvMixer in the Lo branch of the HCM-Attn module, and to quantitatively evaluate the design's superiority regarding computational efficiency and feature representation capability, this section presents a dedicated comparative experimental analysis. Adopting the baseline model as a benchmark, we incorporate both the HiLo

attention mechanism and the improved scheme proposed herein to conduct experimental testing on the TIB dataset.

As illustrated in Table 3, while the incorporation of the standard HiLo mechanism yielded a marginal improvement in detection accuracy, it was concurrently accompanied by a marked decline in inference speed. In contrast, the improved scheme proposed herein demonstrates significant superiority; relative to the standard HiLo mechanism, it not only achieves simultaneous enhancements in both precision and recall but also substantially recovers inference speed while reducing the parameter count. The experimental results demonstrate that substituting the Lo branch of the HiLo attention mechanism with ConvMixer not only augments the model's perceptual capability regarding global structures and its robustness against noise but also mitigates the computational redundancy associated with the attention mechanism, thereby achieving a significant optimization of detection efficiency without compromising high precision.

Table 3. Comparison of Different Attention Mechanisms on the TIB Dataset.

Model	P (%)	R (%)	mAP50 (%)	mAP50:95 (%)	GFLOPs	Params	FPS
Baseline	92.0	91.8	90.9	40.1	57.3	19,974,480	84.7
+HiLo	92.2	91.9	91.0	**40.3**	57.8	19,965,422	78.3
+HCM-Attn	**92.5**	**92.3**	**91.0**	40.0	57.4	**19,940,944**	**87.4**

Note: The bold values indicate the best results.

4.5. Ablation Experiment

To investigate the impact of each improvement on model performance, computational complexity, and parameter size, ablation experiments were conducted on the TIB UAV dataset using the EFPNet model. The model's accuracy on the test set was recorded under non-pretrained conditions. The experiments examined three factors: (A) replacing the AIFI module in RT-DETR with the proposed high–low frequency interaction HCM-Attn module; (B) introducing the DySample dynamic upsampler; and (C) incorporating the PSFT-Net cross-scale feature fusion structure. The results are presented in Table 4 and Figure 9a.

The results indicate that the baseline model achieved a precision of 92.0%, a recall of 91.8%, mAP50 of 90.9%, mAP50:95 of 40.1, GFLOPs of 57.3, a parameter count of 19,974,480, and FPS of 84.7. In Experiments 2, 3, and 4, where only one module was added individually, the detection accuracy improved significantly, confirming the feasibility and effectiveness of each proposed component. For combinations of two modules, B + C achieved the highest precision (94.6%), recall (94.8%), and mAP50 (93.9%); A + C obtained the best mAP50:95 (43.3%); and A + B improved the FPS to 86.1. When all three modules were integrated, the model achieved optimal performance. Compared to the baseline, P, R, mAP50, and mAP50:95 improved by 2.8%, 3.1%, 3.2%, and 3.0%, respectively. Meanwhile, the GFLOPs decreased by 1.4 and the number of parameters was reduced by approximately 0.2 M. It is noteworthy that the joint deployment of modules A and B resulted in a slight increase in GFLOPs. This is primarily attributed to the feature dimension discrepancy between the output of HCM-Attn and the input of DySample, which necessitated the introduction of additional convolutional layers at the implementation level for channel alignment and feature mapping.

Supplementary experiments were conducted on the Drone-vs-Bird UAV dataset. (The results are shown in Table 5 and Figure 9b.) The results demonstrated a performance trend consistent with previous findings, further highlighting the robustness of the proposed method. The experiments confirmed the synergistic effect of the HCM-Attn, DySample, and PSFT-Net modules. When detecting UAV targets, these modules enable effective

integration of deep and shallow features while suppressing interference from complex backgrounds, thereby achieving more accurate detection results.

Table 4. Results of TIB dataset ablation experiments.

Test.No	A	B	C	P (%)	R (%)	mAP50 (%)	mAP50:95 (%)	GFLOPs	Params	FPS
1	×	×	×	92.0	91.8	90.9	40.1	57.3	19,974,480	84.7
2	✓	×	×	92.5	92.3	91.0	40.0	57.1	19,940,944	87.4
3	×	✓	×	93.2	93.2	91.4	41.6	57.3	19,990,928	85.0
4	×	×	✓	93.5	93.4	92.1	41.6	55.6	19,852,144	45.8
5	✓	✓	×	93.2	92.2	90.8	40.2	57.5	19,957,392	86.1
6	✓	×	✓	94.1	94.0	92.5	**43.3**	56.2	19,900,656	62.6
7	×	✓	✓	94.6	94.8	93.9	43.0	56.1	19,950,640	61.5
8	✓	✓	✓	**94.8**	**94.9**	**94.1**	43.1	**55.9**	19,818,608	62.5

Note: The bold values indicate the best results. The ✓ symbol indicates that this module is in use, while × means it is not.

Table 5. Results of Drone-vs-Bird dataset ablation experiments.

Test.No	A	B	C	P (%)	R (%)	mAP50 (%)	mAP50:95 (%)	GFLOPs	Params	FPS
1	×	×	×	97.3	96.6	96.2	55.3	57.3	19,974,480	85.9
2	✓	×	×	97.6	96.3	97.3	56.1	57.1	19,940,944	87.8
3	×	✓	×	97.3	96.5	97.2	55.5	57.3	19,990,928	85.3
4	×	×	✓	97.0	96.3	97.5	**58.7**	55.6	19,852,144	47.5
5	✓	✓	×	97.9	97.0	98.1	56.4	57.5	19,957,392	80.8
6	✓	×	✓	97.5	97.0	97.6	57.9	56.2	19,900,656	57.5
7	×	✓	✓	97.3	**97.8**	97.9	57.3	56.1	19,950,640	59.9
8	✓	✓	✓	**98.3**	97.6	**98.1**	58.3	**55.9**	19,818,608	60.1

Note: The bold values indicate the best results. The ✓ symbol indicates that this module is in use, while × means it is not.

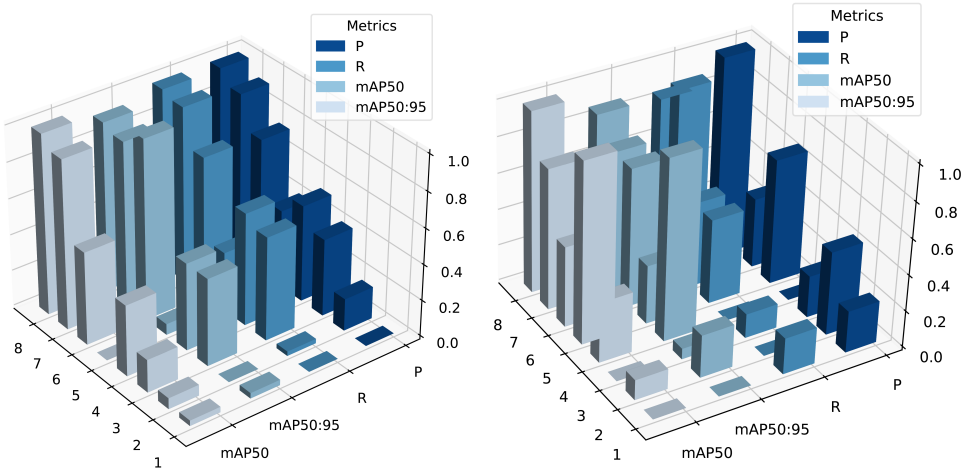


Figure 9. Visualization of selected parameters from ablation experiments. (**a**) TIB; (**b**) Drone-vs-Bird.

To comprehensively evaluate the performance of EFPNet in various UAV small-object-detection scenarios, HiResCAM [40] was employed to generate and compare heatmaps before and after applying the proposed mechanism. Three representative scenarios were selected for experimentation: (1) scenes with complex background interference, (2) scenes involving confusion with similar targets, and (3) scenes containing multiple detection targets. The comparative visualization results are presented in Figure 10.

In Scenario 1, due to the highly complex background and the strong texture similarity between UAV targets and the surrounding environment, the baseline model struggled to

extract discriminative features. As a result, its attention dispersed over irrelevant background regions, often leading to false detections involving non-target areas (see Figure 10d). In Scenario 2, objects with highly similar texture details to UAVs (e.g., birds) were present. The baseline model, lacking sufficient discriminative capacity, mistakenly identified such objects as UAVs (see Figure 10e). In Scenario 3, which involves multiple UAV targets, the original model's multihead attention mechanism was relatively diffuse due to the spatial separation of targets. Consequently, it struggled to distinguish multiple target regions effectively within the same feature map, resulting in overlapping or missed detection boxes (see Figure 10f). When EFPNet was applied, the introduction of an improved attention mechanism and a detail-oriented feature fusion structure enabled the model to effectively suppress background interference and enhance its sensitivity to small targets and local details, thereby achieving higher detection recall and better target differentiation. As shown in Figure 10g–i, the model demonstrated a stronger ability to precisely focus on target regions within complex images, eliminating interference from the background and pseudo-targets (e.g., birds). This significantly reduced false and missed detections, thereby improving the overall detection accuracy.



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

(**g**) (**h**) (**i**)

Figure 10. Visualization of heatmaps generated by ablation experiments. (**a**–**c**) Original images; (**d**–**f**) Heatmap visualization before applying EFPNet; (**g**–**i**) Heatmap visualization after applying EFPNet.

4.6. Comparison of Different Detectors

To verify the performance superiority of the proposed EFPNet, we compared it against a series of classical and state-of-the-art detectors on the TIB and Drone-vs-Bird datasets. The models used in the experiments include Faster R-CNN [15], YOLOv5-m [41], YOLOv8-

m [42], YOLOv11-m [43], Deformable DETR [44], UAV-DETR [45], and the proposed EFPNet, labeled as models A–H, respectively. Each model's performance was evaluated using the aforementioned metrics. The results are presented in Tables 6 and 7.

Table 6. Comparison results on the TIB dataset.

Model	P (%)	R (%)	mAP50 (%)	mAP50:95 (%)	GFLOPs	Params	FPS
A	83.7	77.6	84.7	41.5	207.9	41,348,181	13.1
B	87.1	69.2	77.9	34.8	64.0	25,045,795	102.9
C	82.6	73.3	79.5	35.6	78.7	25,840,339	88.0
D	84.5	73.6	77.6	34.1	68.2	20,053,779	99.2
E	93.2	92.5	92.2	35.5	192.9	40,098,631	78.4
F	93.5	94.6	93.7	40.9	72.8	21,345,096	49.5
G	**94.8**	**94.9**	**94.1**	**43.1**	**55.9**	**19,818,608**	62.5

Note: The bold values indicate the best results.

Table 7. Comparison of different backbone networks on the Drone-vs-Bird dataset.

Model	P (%)	R (%)	mAP50 (%)	mAP50:95 (%)	GFLOPs	Params	FPS
A	94.5	91.7	95.5	57.5	207.9	41,348,181	13.7
B	93.2	88.7	92.8	52.8	64.0	25,045,795	92.3
C	94.2	89.0	93.9	53.7	78.7	25,840,339	81.9
D	91.9	91.8	94.5	53.2	68.2	20,053,779	87.1
E	95.6	93.6	95.2	51.9	192.9	40,098,631	77.9
F	96.3	93.5	96.6	53.5	72.8	21,345,096	49.8
G	**98.3**	**97.6**	**98.1**	**58.3**	**55.9**	**19,818,608**	60.1

Note: The bold values indicate the best results.

On the TIB dataset, model G (EFPNet) demonstrated the best overall performance, achieving a precision of 94.8%, a recall of 94.9%, mAP50 of 94.1%, and mAP50:95 of 43.1%—the highest across all tested models. Compared with the next-best, model F (UAV-DETR), EFPNet improved the mAP50 and mAP50:95 by 0.4% and 2.2%, respectively, while reducing GFLOPs and parameter count by 23.2% and 7.2%. Additionally, the classical YOLO series models exhibited excellent performance in FPS, with model B (YOLOv5) achieving the highest speed of 102.9 FPS. However, these models showed lower detection accuracy and higher GFLOPs and parameter counts. These results demonstrate that EFPNet effectively enhances UAV-target detection performance.

On the Drone-vs-Bird dataset, EFPNet again outperformed all other networks across all metrics, achieving mAP50 of 98.1% and mAP50:95 of 58.3%. This indicates that EFPNet maintains superior detection accuracy and robustness across different UAV scenarios, providing outstanding precision and recall performance.

Figure 11 illustrates the detection performance of seven algorithms across typical complex scenarios, including urban architecture, strong illumination, low-light conditions, vegetation interference, and cloudy weather. Compared with other algorithms that exhibit missed detections or false positives in various scenarios, the proposed EFPNet consistently extracts effective features and performs accurate detections across all environments, while maintaining high confidence levels under identical conditions.

Overall, the experimental results demonstrate that EFPNet achieves superior detection accuracy while maintaining a high processing speed. These findings fully validate the effectiveness and applicability of EFPNet for UAV-target detection tasks in complex background environments. In summary, EFPNet achieves a well-balanced trade-off between performance and efficiency, demonstrating competitive overall results across both datasets.



Figure 11. Comparison of detection results across models.

4.7. Comparative Evaluation of EFPNet Across Diverse Datasets

To comprehensively evaluate the generalization performance and robustness of EFPNet across varying scene distributions and verify its adaptability to diverse environmental backgrounds, this section extends the experimental scope by incorporating two challenging public UAV detection datasets—DUT Anti-UAV [46] and UETT4K Anti-UAV [47]—in addition to the existing TIB and Drone-vs-Bird datasets. We conducted a quantitative comparison between EFPNet and the baseline model across these four datasets under consistent experimental settings.

The experimental results are presented in Table 8. EFPNet exhibits consistent performance gains over the baseline model across all four datasets, providing robust evidence of its generalization capability across diverse data distributions. Specifically, on the highly challenging UETT4K Anti-UAV dataset, EFPNet achieved a significant improvement of 2.2% in mAP50, accompanied by increases of 1.7% and 1.3% in precision and recall, respectively. It is noteworthy that on the DUT Anti-UAV dataset, while precision remained stable, recall experienced a substantial increase of 2.7%; this indicates a significant enhancement in the model's target acquisition capability within complex backgrounds, effectively mitigating the missed detection rate. Although the introduction of feature perception modules resulted in a slight decrease in inference speed, the average FPS of EFPNet across all datasets remained above 60, substantially exceeding the standard threshold for real-time detection (30 FPS). In conclusion, EFPNet not only demonstrates superior performance in specific scenarios but also possesses broad adaptability to other environments, significantly enhancing UAV target acquisition capabilities while maintaining real-time performance.

Table 8. Comparison of EFPNet and the baseline model across four datasets.

Dataset	Model	P (%)	R (%)	mAP50 (%)	mAP50:95 (%)	FPS
TIB	Baseline	92.0	91.8	90.9	40.1	84.7
	EFPNet	**94.8 (+2.8)**	**94.9 (+3.1)**	**94.1 (+3.2)**	**43.1 (+3.0)**	62.5
Drone-vs-Bird	Baseline	97.3	96.6	96.2	55.3	85.9
	EFPNet	**98.3 (+1.0)**	**97.6 (+1.0)**	**98.1 (+1.9)**	**58.3 (+3.0)**	60.1
DUT Anti-UAV	Baseline	97.4	86.2	91.8	59.7	80.9
	EFPNet	97.3	**88.9 (+2.7)**	**92.8 (+1.0)**	**59.9 (+0.2)**	60.7
UETT4KAnti-UAV	Baseline	92.8	92.5	91.3	55.3	83.5
	EFPNet	**94.5 (+1.7)**	**93.8 (+1.3)**	**93.5 (+2.2)**	**57.5 (+2.2)**	61.3

Note: The bold values indicate the best results.

5. Discussion

Although EFPNet demonstrates superior performance across multiple datasets, it is acknowledged that the robustness of vision-only approaches is constrained by inherent physical limitations under extreme lighting (e.g., nocturnal environments) or adverse weather (e.g., dense fog) conditions. The integration of multimodal sensor fusion technologies represents a pivotal strategy for overcoming the perceptual bottlenecks associated with a single visual modality. Specifically, infrared thermal imaging technology can detect thermal radiation signatures generated by UAV motors or batteries, maintaining efficacy even in nocturnal environments; meanwhile, radar possesses superior penetrability through rain, fog, and smoke, alongside its high sensitivity towards moving targets. Future research will aim to extend EFPNet into a multimodal perception framework by fusing thermal features from the infrared spectrum or spatiotemporal positional information from radar; this integration seeks to leverage complementary strengths, thereby constructing a robust detection system capable of true all-weather operation.

6. Conclusions

This study proposes EFPNet, an innovative object detection framework based on the RT-DETR algorithm, designed to tackle the challenges of detecting small UAV targets under complex environmental conditions, where weak feature representation and background interference can hinder detection performance. The core innovations of this approach include the introduction of a High–Low Frequency Interaction Enhancement Module, a Dynamic Sampling Module, and a Cross-Scale Feature Fusion Sparse Pyramid Network. Together, these components enhance the model's sensitivity to UAV target features across diverse scenarios, effectively suppress background interference, and improve detection accuracy. Experimental results on benchmark datasets (TIB and Drone-vs-Bird) validate the feasibility and superiority of the proposed method. Without any pretraining, EFPNet achieved 94.1% mAP50 and 43.1% mAP50:95 on the TIB dataset, and 98.1% mAP50 and 58.3% mAP50:95 on the Drone-vs-Bird dataset—substantial improvements over the original RT-DETR model. These results highlight the performance advantages of the proposed model in UAV-target detection tasks.

Author Contributions: Conceptualization, J.H. and Y.F.; methodology, J.H. and W.J.; software, J.H. and H.T.; validation, J.H., H.T. and Y.S.; investigation, J.H. and W.J.; resources, Y.F. and H.T.; data curation, J.H., W.J. and H.T.; writing—original draft preparation, J.H., S.W. and A.L.; writing—review and editing, W.J., H.T. and Y.F.; visualization, J.H.; project administration, H.T. and Y.F.; funding acquisition, W.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62501623.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Acknowledgments: The authors thank the editor, associate editor, and anonymous reviewers for their valuable time and constructive feedback, which greatly improved the quality of this work. The authors also thank the providers of the public TIB and Drone-vs-Bird UAV datasets, as well as the authors of the comparison methods used in this article, for generously sharing their data and code.

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

References

1. Sun, S.; Chen, C.; Yang, B.; Xu, Y.; Zhao, L.; He, Y.; Jin, A.; Li, L. ALM-LED: Autonomous LiDAR mapping in underground space with Luojia explorer anti-collision drone. *ISPRS J. Photogramm. Remote Sens.* **2025**, *230*, 346–373. [CrossRef]
2. Siddiqi, M.A.; Iwendi, C.; Jaroslava, K.; Anumbe, N. Analysis on security-related concerns of unmanned aerial vehicle: Attacks, limitations, and recommendations. *Math. Biosci. Eng.* **2022**, *19*, 2641–2670. [CrossRef]
3. Taha, B.; Shoufan, A. Machine Learning-Based Drone Detection and Classification: State-of-the-Art in Research. *IEEE Access* **2019**, *7*, 138669–138682. [CrossRef]
4. Luo, H.; Wang, P.; Chen, H.; Kowelo, V.P. Small object detection network based on feature information enhancement. *Comput. Intell. Neurosci.* **2022**, *2022*, 6394823. [CrossRef] [PubMed]
5. Shi, D.; Zhao, C.; Shao, J.; Feng, M.; Luo, L.; Ouyang, B.; Huang, J. Context-Aware Enhanced Feature Refinement for small object detection with Deformable DETR. *Front. Neurorobot.* **2025**, *19*, 1588565. [CrossRef] [PubMed]
6. Zhao, F.; Zhang, J.; Zhang, G. FFEDet: Fine-grained feature enhancement for small object detection. *Remote Sens.* **2024**, *16*, 2003. [CrossRef]
7. Cui, L.; Ma, R.; Lv, P.; Jiang, X.; Gao, Z.; Zhou, B.; Xu, M. MDSSD: Multi-scale deconvolutional single shot detector for small objects. *arXiv* **2018**. [CrossRef]
8. Trockman, A.; Kolter, J.Z. Patches are all you need? *arXiv* **2022**. [CrossRef]
9. Pan, Z.; Cai, J.; Zhuang, B. Fast vision transformers with hilo attention. *arXiv* **2022**. [CrossRef]
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef]
11. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969. [CrossRef]
12. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37. [CrossRef]
13. Terven, J.; Córdova-Esparza, D.M.; Romero-González, J.A. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1680–1716. [CrossRef]
14. Liu, S.; Huang, D.; Wang, Y. Adaptive NMS: Refining Pedestrian Detection in a Crowd. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 6452–6461. [CrossRef]
15. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974. [CrossRef]
16. Saribas, H.; Uzun, B.; Benligiray, B.; Eker, O.; Cevikalp, H. A hybrid method for tracking of objects by UAVs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019. [CrossRef]
17. Zhai, X.; Huang, Z.; Li, T.; Liu, H.; Wang, S. YOLO-Drone: An optimized YOLOv8 network for tiny UAV object detection. *Electronics* **2023**, *12*, 3664. [CrossRef]
18. Yang, Y.; Yuan, S.; Yang, J.; Nguyen, T.H.; Cao, M.; Nguyen, T.M.; Wang, H.; Xie, L. AV-FDTI: Audio-visual fusion for drone threat identification. *J. Autom. Intell.* **2024**, *3*, 144–151. [CrossRef]
19. Zhang, R.; Li, L.; Zhang, Q.; Zhang, J.; Xu, L.; Zhang, B.; Wang, B. Differential Feature Awareness Network Within Antagonistic Learning for Infrared-Visible Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 6735–6748. [CrossRef]
20. Hao, H.; Peng, Y.; Ye, Z.; Han, B.; Zhang, X.; Tang, W.; Kang, W.; Li, Q. A High Performance Air-to-Air Unmanned Aerial Vehicle Target Detection Model. *Drones* **2025**, *9*, 154. [CrossRef]

21. Dadrass Javan, F.; Samadzadegan, F.; Gholamshahi, M.; Ashatari Mahini, F. A modified YOLOv4 Deep Learning Network for vision-based UAV recognition. *Drones* **2022**, *6*, 160. [CrossRef]
22. He, X.; Fan, K.; Xu, Z. Uav identification based on improved YOLOv7 under foggy condition. *Signal Image Video Process.* **2024**, *18*, 6173–6183. [CrossRef]
23. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.H.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
24. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717. [CrossRef]
25. Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Yuan, L.; Zhang, L. Dynamic head: Unifying object detection heads with attentions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7373–7382. [CrossRef]
26. Xin, Z.; Wu, T.; Chen, S.; Zou, Y.; Shao, L.; You, X. ECEA: Extensible co-existing attention for few-shot object detection. *IEEE Trans. Image Process.* **2024**, *33*, 5564–5576. [CrossRef] [PubMed]
27. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19. [CrossRef]
28. Wang, Y.; Li, J.; Chen, Z.; Wang, C. Ships' small target detection based on the CBAM-YOLOX algorithm. *J. Mar. Sci. Eng.* **2022**, *10*, 2013. [CrossRef]
29. Zhang, R.; Xu, L.; Yu, Z.; Shi, Y.; Mu, C.; Xu, M. Deep-IRTarget: An Automatic Target Detector in Infrared Imagery Using Dual-Domain Feature Extraction and Allocation. *IEEE Trans. Multimedia* **2022**, *24*, 1735–1749. [CrossRef]
30. Sporring, J.; Nielsen, M.; Florack, L.; Johansen, P. *Gaussian Scale-Space Theory*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013; Volume 8.
31. Liu, W.; Lu, H.; Fu, H.; Cao, Z. Learning to upsample by learning to sample. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 6027–6037. [CrossRef]
32. Sunkara, R.; Luo, T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Grenoble, France, 19–23 September 2022; pp. 443–459. [CrossRef]
33. Lin, H.; Cheng, X.; Wu, X.; Shen, D. Cat: Cross attention in vision transformer. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18–22 July 2022; pp. 1–6. [CrossRef]
34. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 22–31. [CrossRef]
35. Sun, H.; Yang, J.; Shen, J.; Liang, D.; Ning-Zhong, L.; Zhou, H. TIB-Net: Drone detection network with tiny iterative backbone. *IEEE Access* **2020**, *8*, 130697–130707. [CrossRef]
36. Coluccia, A.; Fascista, A.; Sommer, L.; Schumann, A.; Dimou, A.; Zarpalas, D. The drone-vs-bird detection grand challenge at icassp 2023: A review of methods and results. *IEEE Open J. Signal Process.* **2024**, *5*, 766–779. [CrossRef]
37. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755. [CrossRef]
38. Chen, J.; Kao, S.-h.; He, H.; Zhuo, W.; Wen, S.; Lee, C.H.; Chan, S.H.G. Run, do not walk: Chasing higher FLOPS for faster neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 12021–12031. [CrossRef]
39. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 14420–14430. [CrossRef]
40. Draelos, R.L.; Carin, L. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. *arXiv* **2020**. [CrossRef]
41. Khanam, R.; Hussain, M. What is YOLOv5: A deep look into the internal features of the popular object detector. *arXiv* **2024**. [CrossRef]
42. Yaseen, M. What is YOLOv9: An in-depth exploration of the internal features of the next-generation object detector. *arXiv* **2024**. [CrossRef]
43. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**. [CrossRef]
44. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**. [CrossRef]

45. Zhang, H.; Liu, K.; Gan, Z.; Zhu, G.N. UAV-DETR: Efficient end-to-end object detection for unmanned aerial vehicle imagery. *arXiv* **2025**. [CrossRef]
46. Zhao, J.; Zhang, J.; Li, D.; Wang, D. Vision-Based Anti-UAV Detection and Tracking. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 25323–25334. [CrossRef]
47. Sarwar Awan, M.; Azhar Ali Zaidi, S.; Mir, J. UETT4K Anti-UAV: A Large Scale 4K Benchmark Dataset for Vision-Based Drone Detection in High-Resolution Imagery. *IEEE Access* **2025**, *13*, 73553–73568. [CrossRef]

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