



ARTICLE

HiFreq-DETR: A Hierarchical Framework Synergizing High-Resolution Injection and Frequency-Aware Multi-Scale Interaction for Tiny Object Detection

Linyu Dong¹, Tao Li² and Hao Li^{2,*}

¹School of Information Science & Engineering, Yunnan University, Kunming, China

²Yunnan Communications Investment & Construction Group Co., Ltd., Kunming, China

*Corresponding Author: Hao Li. Email: lihao707@ynu.edu.cn

Received: 27 March 2026; Accepted: 18 May 2026; Published: 23 July 2026

ABSTRACT: While Transformer-based detectors excel in global modeling, their efficacy in unmanned aerial vehicle (UAV)-based tiny object detection is limited by information loss during aggressive downsampling and the lack of high-frequency structural cues. To bridge this gap, we propose HiFreq-DETR, a dedicated framework that optimizes the synergy between spatial fidelity and semantic discriminability. The core innovation lies in its hierarchical information preservation strategy, which employs a ResNeSt14d backbone coupled with an S_2 spatial injection path to recover critical high-resolution structural anchors, and introduces a frequency-selective interaction module to decouple target saliency from background noise. Experimental results demonstrate the substantial value of our approach. On the VisDrone dataset, HiFreq-DETR significantly outperforms the baseline RT-DETR, achieving improvements of 3.9% in AP and 4.4% in AP_S, confirming its effectiveness for tiny object detection. Furthermore, an optimized lite variant is evaluated to challenge the limits of high-efficiency processing for resource-constrained scenarios, while superior gains on the HazyDet dataset validate the model's structural robustness in adverse aerial environments. These findings establish HiFreq-DETR as a high-fidelity and versatile solution for complex remote sensing applications.

KEYWORDS: UAV; tiny object detection; multi-scale feature interaction; frequency-selective attention; high-resolution representation learning; DETR

1 Introduction

As a cornerstone task in computer vision, object detection has demonstrated profound practical utility in industrial sectors such as autonomous driving, aerial monitoring, and urban surveillance. With the rapid proliferation of Unmanned Aerial Vehicle (UAV) platforms and high-resolution imaging sensors, acquiring multispectral aerial imagery over expansive geographical areas has become increasingly accessible. This enhanced sensing capability has positioned automated detection as a pivotal driver for critical tasks, including traffic flow analysis, disaster response, and maritime search and rescue [1,2]. However, distinct from natural scene images, UAV imagery presents unique imaging mechanisms that pose severe challenges, namely that targets are typically characterized by extreme distribution density, intricate background clutter, and minute scales. Due to feature sparsity caused by varying Ground Sampling Distance (GSD), existing algorithms encounter significant performance bottlenecks when identifying tiny targets.

Throughout the evolution of object detection, frameworks have transitioned from regional proposal methods to end-to-end regression models, culminating in the recent success of Detection Transformers

(DETR) [3,4]. However, the real-world challenges of UAV imagery often exceed the theoretical capacity of these standard models. Taking the VisDrone dataset at a 640×640 resolution as a representative example, a significant portion of targets, such as pedestrians or distant vehicles, occupy fewer than 10×10 pixels, covering less than 0.02% of the total image area [5]. In such extreme scenarios, even a single-pixel localization error during feature downsampling represents a 10% shift in the bounding box coordinate, rendering the deep semantic interactions at stride-8 or stride-16 stages insufficient for capturing near-vanishing spatial signals. Consequently, traditional single-scale or monolithic interaction paradigms fail to provide the dual requirements of high precision and robustness necessary for practical engineering deployment [6].

Despite these advancements, a deep analysis of RT-DETR and its variants reveals persistent bottlenecks in UAV-captured scenarios. A primary limitation is the spectral bias inherent in deep neural networks, which tend to converge on low-frequency components representing coarse structures while filtering out high-frequency residuals. In UAV imagery, the semantic signature of a tiny object is primarily encoded in its high-frequency spatial abruptness, which corresponds to sharp intensity gradients that define object boundaries. While the weighted-average operations in spatial-domain attention act as adaptive low-pass filters that smooth out these critical perturbations, frequency-domain interaction allows for the explicit decoupling of these signals. By transforming features into the spectral domain, the model can selectively amplify the high-frequency components that encapsulate the structural energy of tiny targets, ensuring that their sparse representations are not annihilated by the dominant low-frequency energy of the surrounding environment [7]. Furthermore, the Attention-based Intra-scale Feature Interaction (AIFI), the hybrid encoder in existing models, relies predominantly on single-scale spatial-domain self-attention. This constrained interaction mode operates on a fixed receptive field, which struggles to adapt to the massive GSD fluctuations caused by varying flight altitudes and sensor specifications. When tiny targets undergo drastic scale shifts, this single-scale mapping lacks the cross-scale semantic flexibility required to maintain representational consistency. Additionally, current end-to-end detectors predominantly follow a serial high-to-low resolution stream, concentrating interactions at the S_3 or deeper layers. This architectural bias often overlooks the signal erosion during successive downsampling [8]. For objects occupying only a few pixels, critical spatial gradients are susceptible to irreversible attenuation before reaching the interaction stage. Consequently, deep semantic features often lack the precise spatial anchors that parallel high-resolution structures are intended to preserve, leading to significant localization instability for tiny target boundaries [9].

To overcome these limitations, we propose HiFreq-DETR: a dedicated framework for aerial scenes that integrates high-resolution spatial details with frequency-aware multi-scale feature interaction. First, we integrate the ResNeSt14d lightweight backbone, utilizing its Split-Attention mechanism to improve cross-channel discriminability while optimizing the parameter budget [10]. Second, inspired by the necessity of preserving spatial fidelity [11], we explicitly construct an S_2 scale feature injection path to introduce high-resolution spatial anchors from the stride-4 stage directly into the interaction network, mitigating geometric structure loss at the source. Finally, we propose the Frequency-selective Attention-based Multi-scale Feature Interaction (FS-AMFI) module. By embedding frequency-selection logic into the multi-scale interaction process, we replace the conventional single-scale attention with a more robust mechanism that simultaneously addresses frequency-domain decoupling and scale-invariant perception, thereby overcoming the limitations of traditional spatial attention in tiny target recognition.

The primary contributions of this work are summarized as follows:

- We propose the HiFreq-DETR architecture, an end-to-end detection framework optimized for UAV imagery. While the base model establishes a robust performance baseline, we further develop a

lightweight variant that maintains competitive precision with greatly reduced computational complexity, ensuring versatility across diverse hardware constraints.

- We design an S_2 scale feature injection mechanism. By explicitly incorporating high-resolution spatial anchors from the stride-4 stage, we provide a high-fidelity grid that manifests frequency-enriched semantic signals, thereby significantly enhancing the localization accuracy of tiny objects occupying only a few pixels.
- We develop the FS-AMFI module, which integrates frequency-selective attention into multi-scale interaction paths. This module adaptively reinforces high-frequency target saliency and compensates for the structural information loss in deep semantic layers, effectively resolving the signal annihilation issues inherent in standard feature transformations.
- We strategically employ ResNeSt14d as the backbone to enhance multi-scale feature extraction without introducing additional parameter overhead. This choice provides superior architectural synergy with our frequency-aware modules, ensuring that the enhanced representational power is fully leveraged by the subsequent interaction layers.

The subsequent sections of the paper are arranged as follows: [Section 2](#) reviews related research; [Section 3](#) details the architectural design of HiFreq-DETR and the mathematical principles of each core module; [Section 4](#) verifies the effectiveness of the algorithm through extensive comparative experiments and ablation studies; [Section 5](#) summarizes the full text and prospects future research directions.

2 Related Work

2.1 CNN-Based Paradigms for Object Detection

Object detection within UAV imagery has transitioned from traditional heuristic-based frameworks to sophisticated deep learning paradigms, characterized by a fundamental shift toward multi-scale representation learning. Driven by the increasing demand for autonomous aerial surveillance, recent research has prioritized the development of holistic sensing solutions; for instance, remote intelligent perception systems [12] and integrated neural network frameworks [13] have been established to provide unified end-to-end architectures for multi-object detection and recognition in complex environments. Within these integrated frameworks, early two-stage detectors exemplified by Faster R-CNN [14] established the utility of Region Proposal Networks but often struggled with the extreme scale variations and high-frequency spatial variances inherent in UAV views. To refine feature granularity, ARFP [15] introduced adaptive receptive field pooling, while SDP [16] leveraged sub-pixel discriminative pooling to recover fine-grained details for tiny instances. Simultaneously, single-stage detectors have gained prominence for their superior inference efficiency. HIC-YOLOv5 [17] and MSFE-YOLO [18] targeted small objects by integrating high-resolution integration layers and specialized multi-scale fusion blocks. Recent variants including Drone-YOLO [19] and YOLOv12 [20] further optimized backbone architectures to enhance representation capability. However, these CNN-based methods predominantly rely on local receptive fields and progressive downsampling, which often lack the global contextual awareness and spectral sensitivity necessary to distinguish tiny targets from intricate backgrounds.

2.2 End-to-End Object Detection

The emergence of the Detection Transformer DETR has revolutionized the field by framing object detection as a direct set prediction problem, effectively eliminating the need for hand-crafted components [3]. Subsequent architectures including Deformable DETR [21] and RT-DETR [4] introduced sparse sampling and Hybrid Encoders to optimize efficiency. This end-to-end paradigm has been further extended by recent state-of-the-art frameworks, including D-FINE [22] and Mamba-YOLO [23]. To bridge the semantic gap

in cluttered scenes, TGCADNet [24] introduces a CLIP-based text-guided method, leveraging linguistic priors from large-scale pre-training to enhance the perception of tiny instances. Despite these architectural advancements, generic detectors encounter a significant information bottleneck during feature transmission. Most DETR-based architectures initiate feature interaction from the S_3 scale or deeper, often overlooking the critical low-level geometric information preserved at shallower stages. Consequently, even advanced global modeling mechanisms struggle to recover high-precision localization cues once the high-resolution spatial anchors are omitted during successive downsampling.

2.3 Tiny Object Detection and Frequency Domain Learning

Detecting tiny objects remains a challenging task due to their limited spatial extent and the interference of complex backgrounds. Current research primarily focuses on feature fusion and localization refinement. Representative studies such as GLSAN [25] adopt feature fusion strategies, while methods like NWD [26] or CCAFNet [27] explore cross-scale contextual alignment via specialized metrics to mitigate feature misalignment. A pivotal emerging trend is the utilization of frequency domain learning to restore weak target signatures that are typically smoothed out in the spatial domain. For instance, YOLO-DCTI [28] incorporates Discrete Cosine Transform coefficients to compensate for geometric ambiguity, while WaveNet [29] leverages wavelet-based saliency detection to isolate high-frequency target signatures from low-frequency background interference. Despite these advancements, two critical challenges remain under explored in UAV-based detection. First, while high-resolution layers improve detail perception, the irreversible loss of structural energy during the early downsampling of tiny targets often occupying only a few pixels has not been fully addressed through explicit feature injection mechanisms. Second, existing frequency domain methods often utilize static spectral coefficients, failing to adaptively filter task irrelevant redundant information during dynamic feature interactions. Therefore, a synergistic approach that explicitly preserves high-resolution spatial fidelity while introducing adaptive frequency-selective perception is essential for achieving robust detection in complex aerial environments.

3 Method

3.1 Overview of HiFreq-DETR

To address the critical challenges of feature degradation and complex background interference in UAV-perspective imagery, we propose the HiFreq-DETR architecture. Grounded in the design philosophy of accuracy-first and lightweight-collaboration, HiFreq-DETR systematically reconstructs the weak responses of tiny targets by enhancing the synergy between shallow spatial features and deep frequency-aware perception. As illustrated in Fig. 1, while retaining the hybrid encoder-decoder structure of the baseline RT-DETR, HiFreq-DETR incorporates three fundamental structural innovations. First, we adopt ResNeSt14d as the backbone, which leverages a split-attention mechanism to enhance fine-grained feature discriminability. This strategic replacement establishes a more robust representation foundation within an optimized parameter budget compared to the standard ResNet-18 [30]. Second, we introduce a stride-4 S_2 injection path to mitigate the spatial information loss inherent in deep hierarchies. Rather than a simple skip-connection, this path functions as a high-resolution spatial grid that manifests the frequency-enriched semantic signals propagated top-down from the encoder, ensuring precise localization anchors for minuscule targets. Finally, we replace the conventional AIFI with the proposed Frequency-selective Attention-based Multi-scale Feature Interaction (FS-AMFI) module at the highest semantic level (S_5). By orchestrating Dynamic Multi-scale Feature Extraction (DMFE) for adaptive receptive field modeling and frequency-selective residual enhancement, FS-AMFI explicitly decouples tiny-object saliency from low-frequency background noise. In this cascaded pipeline, the frequency-enhanced features refined at S_5 are progressively

integrated with the S_2 spatial anchors within the Cross-Scale Feature Fusion (CCFF) network. Ultimately, the refined multi-scale features are processed by the Transformer decoder, enabling HiFreq-DETR to achieve superior detection precision and representational robustness in complex aerial scenes.

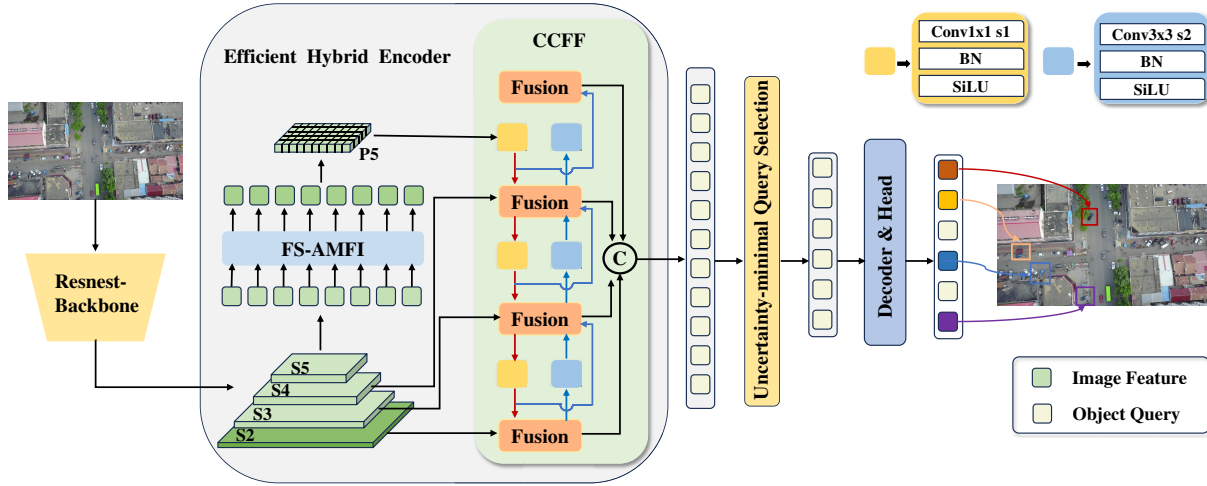


Figure 1: Overall architecture of the proposed HiFreq-DETR. Detailed configurations of the FS-AMFI module are provided in [Section 3.4](#).

3.2 High-Resolution Feature Injection with S_2

In conventional Transformer-based hybrid encoders, the feature pyramid is typically constructed from the $\{S_3, S_4, S_5\}$ stages, corresponding to downsampling strides of 8, 16, and 32, respectively. While this hierarchical structure is efficient for capturing global semantic context, it imposes a significant structural limitation on the representation of minuscule targets in UAV-perspective imagery. We observe that the primary performance bottleneck in tiny object detection is not merely a lack of semantic reasoning, but the irreversible attenuation of shallow spatial gradients during successive downsampling. To address this, HiFreq-DETR establishes a spatial-semantic synergy paradigm, where high-resolution structural anchors from the S_2 stage are explicitly integrated into a cascaded pipeline to manifest the frequency-enriched abstractions refined at deeper levels.

Tiny object detection is fundamentally constrained by spatial sampling density. In aerial imagery, the discriminative features of minuscule targets are inherently encoded in high-frequency spatial patterns. As the downsampling stride increases, the signal energy of a target is dissipated through pooling and convolution processes, leading to spatial saliency nullification. Quantitative analysis of the baseline $\{S_3, S_4, S_5\}$ hierarchy reveals that at the S_3 stage with a stride of 8, a 16×16 pixel tiny target is reduced to a mere 2×2 activation area. For smaller targets of 8×8 pixels, the representation collapses into a single 1×1 pixel response, which suffers from severe information sparsity. Consequently, the target's geometric structure is substantially degraded before reaching the encoder, rendering it nearly impossible for high-level semantic layers to perform precise localization.

To counteract this, we incorporate the S_2 injection path with a stride of 4 into our feature fusion framework. In our cascaded architecture, the S_2 stage functions as the essential spatial beneficiary of the encoder's output. While it introduces raw physical geometry from the backbone, its primary role within the Cross-Scale Feature Fusion network is to provide a high-resolution grid that manifests the sharp structural cues, such as edges and boundaries, propagated top-down from the FS-AMFI-enhanced S_5 stage. This design establishes a collaborative mechanism: the S_5 path ensures that high-frequency target signals are preserved

and reinforced in the semantic domain, while the S_2 path provides the fine-grained spatial constraints necessary to project these signals for localized boundary regression. Ultimately, this integration ensures that the model maintains the structural integrity of tiny objects across the entire feature flow, establishing a saliency baseline that bridges the gap between raw physical gradients and abstract semantic reasoning.

3.3 Backbone: Efficient ResNeSt14d for Preserving Geometric Fidelity

In the architectural design of HiFreq-DETR, we strategically employ an efficient 14-layer ResNeSt configuration, denoted as ResNeSt14d, to replace the conventional ResNet-18 backbone. Rather than a trivial component substitution, this selection represents a deliberate architectural optimization aimed at maximizing representational capacity within a constrained parameter budget. In this work, ResNeSt14d is defined as a task-oriented configuration implemented via the timm library [31], providing a specific depth-width trade-off that is absent in the standard baseline family officially introduced in [10]. By leveraging the split-attention mechanism, we enhance feature discriminability without the computational redundancy inherent in deeper architectures.

ResNeSt14d incorporates the split-attention mechanism within the ResNet-D structural variant [32], featuring a depth-optimized hierarchy of four stages with a single bottleneck block per stage. Specifically, the d suffix denotes the integration of structural principles from the ResNet-D paradigm, characterized by a deep stem and heuristic downsampling strategies that are critical for preserving tiny object features. The standard 7×7 convolution is replaced by a triplet of 3×3 convolutions to increase the effective receptive field without sacrificing spatial resolution, while 2×2 average pooling is applied before strided convolutions. This approach significantly mitigates the feature aliasing and initial information loss typical of aggressive downsampling in standard ResNet architectures. Such architectural characteristics offer decisive advantages for detecting minuscule instances in cluttered UAV imagery. By eliminating computational redundancy associated with excessive depth, ResNeSt14d maintains an efficient profile while its gradual downsampling effectively preserves the high-frequency geometric boundaries essential for precise boundary regression. Simultaneously, the inherent channel-wise salience concentrates the representational capacity on critical object-level information, thereby enhancing the signal-to-noise ratio. Consequently, this backbone provides a robust, detail-preserving feature stream that serves as a high-fidelity foundation for the subsequent S_2 spatial injection and multi-scale interaction modules.

3.4 Frequency Selective Attention-Based Multi-Scale Feature Interaction

In the original RT-DETR architecture, the Attention-based Intra-scale Feature Interaction (AIFI) module models global context dependencies via a standard self-attention mechanism. While this design enhances high-level semantic representation, it exhibits significant limitations when dealing with dense, tiny objects in UAV imagery. Specifically, the constrained single-scale feature mapping of AIFI lacks the capability for adaptive modeling of drastic scale variations, and the feature smoothing effect inherent in attention computation leads to the dilution of critical high-frequency details. To address these bottlenecks, we propose the Frequency-selective Attention-based Multi-scale Feature Interaction (FS-AMFI) module. As illustrated in Fig. 2, FS-AMFI is strategically positioned at the highest semantic level to transform the raw S_5 feature from the backbone into a frequency-enriched representation P_5 , which serves as the primary semantic source for the subsequent cascaded fusion pipeline.

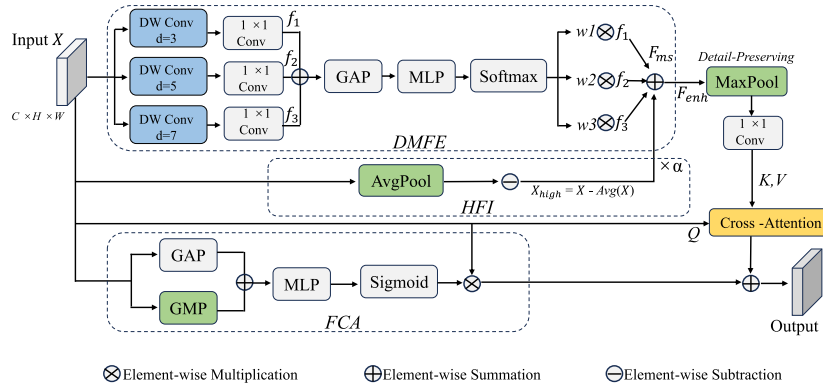


Figure 2: Overview of the proposed FS-AMFI module.

FS-AMFI is constructed along two pivotal dimensions. First, we introduce the Dynamic Multi-scale Feature Extraction (DMFE) as the core of the intra-stage interaction path to capture multi-scale contexts. Second, we integrate a frequency-aware enhancement mechanism underpinned by High-Frequency Injection (HFI), Detail-Preserving Key-Value construction, and Frequency-aware Channel Attention (FCA). This dual-path design enables the model to achieve precise reconstruction of tiny object saliency before it is propagated to shallower scales.

Dynamic Multi-scale Feature Extraction (DMFE): The original AIFI module is fundamentally constrained by its homogeneous feature generation, where Query, Key, and Value are derived from identical feature mappings. This paradigm restricts relational modeling to a single receptive field, leading to receptive field homogenization. While suitable for global context extraction, this rigidity becomes a performance bottleneck in UAV-perspective imagery, where the coexistence of minute targets and expansive backgrounds demands a multi-scale spatial perception that a fixed-scale operator cannot provide. To overcome this structural limitation, we propose the Dynamic Multi-scale Feature Extraction (DMFE) module, which elevates feature interaction from a static process to a scale-aware adaptive paradigm.

Given an input $X \in \mathbb{R}^{C \times H \times W}$ from the S_5 stage, DMFE implements a triple-branch decomposition via depthwise separable convolutions (DWConv) with dilation rates $d \in \{3, 5, 7\}$. The selection of these specific dilation rates is strategically designed to construct a hierarchical ensemble of effective receptive fields (ERF). Specifically, with a 3×3 kernel base, these rates correspond to regional perceptions of 7×7 , 11×11 , and 15×15 pixels, respectively. This configuration ensures a comprehensive spatial coverage: the smaller rate ($d = 3$) preserves fine-grained local textures essential for the precise localization of tiny targets, while the larger rates ($d = 5, 7$) aggregate broader environmental semantics to provide the global contextual cues necessary for disambiguating targets in cluttered backgrounds. To move beyond the limitations of naive feature summation, which often introduces scale redundancy and dilutes unique multi-scale characteristics, we introduce an adaptive weighting mechanism governed by global statistical priors. By distilling multi-scale responses into a global descriptor z through Global Average Pooling (GAP), the model gains a holistic view of the feature importance across scales: $z = \text{GAP}(\sum_{i=1}^3 f_i)$. The subsequent Multi-Layer Perceptron (MLP) and Softmax activation function then map this descriptor to a set of normalized scale-specific weights $[w_1, w_2, w_3]$, which dynamically reconfigure the effective receptive field: $F_{ms} = \sum_{i=1}^3 w_i \otimes f_i$. This formulation enables the network to perform dynamic receptive field optimization, allowing the attention mechanism to shift its focus adaptively based on the content density and target distribution of the scene. By replacing the rigid single-scale interaction of AIFI with this multi-scale foundation, DMFE ensures that the subsequent attention matrix is constructed from a discriminative, scale-robust feature base, significantly enhancing the model's robustness to scale variations in complex UAV-captured environments.

Frequency-selective Enhancement (FS): Beyond multi-scale modeling, the FS-AMFI architecture incorporates a frequency-selective enhancement mechanism. In high-resolution aerial imagery, the salient features of tiny objects are typically encoded as high-frequency spatial mutations, whereas the massive backgrounds are characterized by smoother, low-frequency distributions. Conventional AIFI, which relies on vanilla semantic features, inadvertently suppresses these high-frequency signals during feature compression and relational modeling. To address this, we re-engineer the interaction paradigm through three complementary frequency-aware components: (1) High-Frequency Injection (HFI), which explicitly recovers geometric details; (2) Detail-Preserving Key-Value Construction, which safeguards peak responses from downsampling-induced dilution; and (3) Frequency-aware Channel Attention (FCA), which facilitates joint modeling of background semantics and foreground saliency. These components collectively transform FS-AMFI into a frequency-sensitive interaction framework, providing a theoretically robust solution for tiny object representation.

High-Frequency Injection (HFI): Deep networks inherently act as low-pass filters, a property that leads to the suppression of high-frequency components critical for detecting tiny objects. To counteract this, we propose HFI within the FS-AMFI module to explicitly inject structural details into the deep feature stream. Given the input feature X from the backbone, we derive its high-frequency residual X_{high} by subtracting the local average signal: $X_{\text{high}} = X - \text{AvgPool}_k(X)$. Mathematically, this residual functions as a high-pass filter that isolates essential edges and textures. While learnable frequency transformations, such as discrete cosine or wavelet transforms, offer adaptive spectral decomposition, they often entail significant computational overhead and risk overfitting to specific flight altitudes. In contrast, our residual-based approach serves as a scale-invariant structural descriptor. By isolating local intensity variances, this fixed-kernel mechanism provides consistent high-frequency responses regardless of target scale variations caused by varying UAV altitudes. To prevent the amplification of noise, we modulate this residual with a learnable scalar α before integrating it with the multi-scale features: $F_{\text{enh}} = F_{\text{ms}} + \alpha X_{\text{high}}$. This mechanism ensures that the enhanced P_5 output contains preserved structural anchors, which are subsequently propagated through the cascaded pipeline to provide critical semantic guidance for the high-resolution S_2 injection path.

Detail-preserving Cross-attention Construction: AIFI employs a symmetric self-attention mechanism, which proves suboptimal for tiny object detection where the target signal is typically sparse and weak. In such scenarios, target responses are often overwhelmed by the dominant low-frequency background during Softmax normalization. To address this, we reconstruct the interaction into an asymmetric cross-attention paradigm. Specifically, the query (Q) is derived from the high-resolution input X , while the key (K) and value (V) are generated from the enhanced features F_{enh} through a sequence of max-pooling and 1×1 convolutions: $K = W_K \phi(\text{MaxPool}(F_{\text{enh}}))$, $V = W_V \phi(\text{MaxPool}(F_{\text{enh}}))$, where $\phi(\cdot)$ denotes the activation function and W_K , W_V represent the learnable weights of the 1×1 convolutional layers. The choice of MaxPool as the downsampling operator is a deliberate design decision to safeguard the sparse activations of tiny targets. While average pooling (AvgPool) or learnable strided convolutions might appear more balanced, our empirical investigations detailed in [Section 4.3.2](#) reveal that AvgPool tends to dilute the weak signals of minuscule objects by averaging them with surrounding background noise. In contrast, MaxPool functions as a peak-preserving saliency filter, ensuring that the most discriminative high-frequency responses are propagated even after a $16\times$ spatial reduction. Furthermore, this asymmetric construction ensures that while K and V are condensed for computational efficiency and contextual anchoring, the fine-grained structural details required for precise localization are fully preserved in the Q stream, which maintains its original high-resolution spatial grid. This configuration prevents the feature washing-out effect, leading to more precise relational modeling between high-resolution queries and reinforced target features.

Frequency-aware Channel Attention (FCA): To optimize feature selection at the channel level, we replace the simple identity residual in AIFI with FCA. This module exploits the frequency-domain complementarity of Global Average Pooling (GAP) and Global Max Pooling (GMP). Specifically, GAP captures the global low-frequency trends representing background semantics, while GMP emphasizes local high-frequency peaks associated with object saliency: $w = \sigma(\text{MLP}(\text{GAP}(X) + \text{GMP}(X)))$. By fusing these statistics, the generated channel weights w are jointly constrained by both the global background distribution and local saliency. This dual-constraint mechanism allows FCA to recalibrate features $F_{ca} = w \odot X$ with higher precision, effectively suppressing background interference while amplifying tiny object signals. This design ensures that the residual connection serves not just as a shortcut, but as a frequency-aware modulator that improves the discriminability of the final representation.

To summarize, the proposed FS-AMFI module redesigns the interaction paradigm for tiny object detection by moving beyond the monolithic, single-scale architecture of the original AIFI. By integrating DMFE with a frequency-selective enhancement branch, the module establishes a dual-stream collaborative framework that harmonizes global semantic modeling with high-frequency spatial saliency. The overall forward propagation of the FS-AMFI module is formulated as:

$$Y = \text{CrossAttn}(Q_X, K_{\Phi(F_{enh})}, V_{\Phi(F_{enh})}) + F_{ca}$$

where $\Phi(\cdot)$ denotes the sequential composition of detail-preserving max-pooling and 1×1 convolution. This design explicitly decouples high-frequency tiny-object features from low-frequency background noise, providing a robust, scale-adaptive, and detail-preserving solution for dense target representation in complex aerial imagery. By transforming S_5 into the frequency-enriched P_5 , FS-AMFI provides the semantic source that underpins the localization capacity of the S_2 , S_3 , and S_4 injection paths during the subsequent top-down fusion process.

3.5 Efficiency-Oriented Optimization for High-Resolution Stages

To further optimize the computational efficiency of the high-resolution S_2 injection path, we introduce a specialized lite configuration (HiFreq-DETR-lite). Our core observation is that the S_2 stage, which operates at a stride of 4, primarily functions as a provider of fine-grained spatial-geometric anchors rather than high-level semantic abstractions. Given a standard input resolution of 640×640 pixels, the S_2 feature maps maintain a high spatial resolution of 160×160 . The computational intensity of standard dense 3×3 convolutions at this scale introduces significant overhead relative to the task of spatial integration. To address this, we replace the standard fusion layers at the S_2 scale with Lightweight Cross Stage Partial Reparameterization Layers (CSPRepLayers), which leverage Depthwise Separable Convolutions (DW-Sep). By decoupling spatial filtering from channel-wise mixing, this targeted substitution significantly mitigates the Floating Point Operations (FLOPs) associated with the 160×160 feature maps while preserving the integrity of the spatial-geometric signals. This optimization ensures that HiFreq-DETR maintains an optimal balance between superior detection accuracy for tiny objects and high inference throughput.

3.6 Loss Function Design

The prediction objective of HiFreq-DETR is formulated as a set prediction problem based on bipartite graph matching, which aligns predicted objects with ground-truth labels. To optimize the detection of minuscule remote sensing targets, we adopt an integrated multi-task loss function. Furthermore, to accelerate model convergence, a deep supervision strategy is employed, where auxiliary losses are computed at each decoder layer l , and the final objective L_{total} is defined as the weighted summation:

$$L_{\text{total}} = \sum_{l=1}^L \left(\lambda_{\text{cls}} L_{\text{cls}}^{(l)} + \lambda_{L1} L_{L1}^{(l)} + \lambda_{\text{GIoU}} L_{\text{GIoU}}^{(l)} \right)$$

where the hyperparameters are set to $\lambda_{\text{cls}} = 1$, $\lambda_{L1} = 5$, and $\lambda_{\text{GIoU}} = 2$. This configuration prioritizes the regression precision of tiny bounding boxes while maintaining a balanced classification objective.

To mitigate the extreme class imbalance and background noise interference inherent in remote sensing imagery, we utilize Varifocal Loss (VFL) as the classification component. Unlike standard cross-entropy loss, VFL incorporates the predicted IoU-aware classification score q to prioritize high-quality positive samples, defined as:

$$\text{VFL}(p, q) = \begin{cases} -q [q \log(p) + (1 - q) \log(1 - p)], & q > 0 \\ -\alpha p^\gamma \log(1 - p), & q = 0 \end{cases}$$

where p is the predicted probability and q represents the IoU between the prediction and ground truth. By directing the optimization toward positive samples that exhibit higher spatial localization quality, VFL prevents the model from being overwhelmed by the high density of background negative samples.

The regression objective integrates L_1 loss and Generalized Intersection over Union (GIoU) loss to ensure robust coordinate refinement. The L_1 loss provides direct supervision for the normalized center coordinates (c_x, c_y) and dimensions (w, h) , defined as $L_{L1} = |\mathbf{b} - \hat{\mathbf{b}}|_1$. While this ensures stable convergence during early training, it is sensitive to scale variations. To counteract this, we incorporate the GIoU loss, which is defined as:

$$L_{\text{GIoU}} = 1 - \left(\frac{|\mathbf{b} \cap \hat{\mathbf{b}}|}{|\mathbf{b} \cup \hat{\mathbf{b}}|} - \frac{|C \setminus (\mathbf{b} \cup \hat{\mathbf{b}})|}{|C|} \right)$$

where C is the smallest convex hull encompassing both the predicted box $\hat{\mathbf{b}}$ and the ground-truth box $\mathbf{b}$. Because tiny targets exhibit extreme sensitivity to pixel-level deviations, the scale-invariance of GIoU is essential. It provides meaningful gradients even when there is no initial spatial overlap between the boxes, effectively bridging the localization gap for minuscule targets where initial predictions often exhibit significant displacement.

In summary, the synergistic design of these loss components addresses the primary bottlenecks in remote sensing tiny object detection. The combination of VFL for high-quality sample re-weighting and the complementary L_1 and GIoU losses for scale-invariant coordinate regression establishes a robust training framework. This approach enables HiFreq-DETR to maintain high localization precision under extreme spatial constraints, ensuring that the model preserves the structural integrity of tiny objects while maximizing classification accuracy.

4 Experiments

4.1 Experimental Setup

1) Dataset: *VisDrone-DET2019*: VisDrone-DET2019 [5] is a large-scale benchmark for UAV-based visual analysis. It contains 10,209 high-resolution images from various drone-mounted cameras, covering diverse urban and suburban environments across 14 cities in China. It is split into training (6471 images), validation (548 images), and test-dev (3190 images) sets, with approximately 2.5 million annotated bounding boxes across 10 categories, including pedestrians, vehicles, and bicycles. Its core challenge is small and densely distributed targets in complex scenes, making it a rigorous benchmark for aerial object detection.

SIMD: The Satellite Imagery Multi-scale Object Detection dataset [33] comprises 5000 1024×768 RGB images from 79 locations. It contains 45,096 instances across 15 categories, primarily vehicles, aircraft, and watercraft. Characterized by significant scale variations and complex backgrounds, SIMD provides a rigorous benchmark for evaluating multi-scale detection and algorithmic robustness in high-resolution satellite imagery.

HazyDet: HazyDet [34] is a benchmark specifically for object detection in adverse atmospheric conditions. It consists of 11,000 images split into training, validation, and test subsets with 8000, 1000, and 2000 images, respectively, together with 383,000 annotations across three vehicle categories: car, truck, and bus. By integrating real-world captures with physics-based haze simulations, it effectively evaluates model robustness in low-visibility aerial surveillance scenarios.

2) Evaluation Metrics: Following the standard ms COCO evaluation protocol [35], we adopt Mean Average Precision (mAP) as the core performance indicator, including AP (at IoU thresholds from 0.5 to 0.95), AP_{50} , and scale-specific metrics AP_S , AP_M , and AP_L . To evaluate model complexity and real-time viability, we further incorporate Parameters, FLOPs, Latency, and FPS, which collectively reflect the model's computational efficiency and inference speed in practical deployment scenarios.

3) Implementation Details: All experiments are conducted on two NVIDIA GeForce RTX 4090 GPUs using the RT-DETR framework as the baseline. The efficient ResNeSt14d backbone is initialized with ImageNet-1k pretrained weights via the timm library, while other components are trained from random initialization using a standard end-to-end fine-tuning protocol. We train all models for 300 epochs with a global batch size of 8, employing the AdamW optimizer with a learning rate of 10^{-4} and a weight decay of 10^{-4} . To ensure convergence stability and enhance generalization, we incorporate a gradient clipping max norm of 0.1 and apply an Exponential Moving Average (EMA) with a decay rate of 0.9999. Input images are uniformly resampled to a resolution of 640×640 pixels for both training and inference. To prevent overfitting, an early stopping strategy with a 20-epoch patience is implemented based on the Average Precision (AP) metric, ensuring that the reported results represent optimal steady-state performance. All comparative evaluations are performed under strictly identical hyperparameter configurations to maintain a rigorous and fair assessment.

4.2 Comparison With the State-of-the-Arts

To verify the effectiveness and superiority of the proposed HiFreq-DETR, we conduct a comprehensive comparative analysis against 13 state-of-the-art (SOTA) methods on the VisDrone-DET2019 validation set. These methods span three major categories: two-stage, one-stage, and end-to-end detectors. The quantitative results are summarized in Table 1. Furthermore, we evaluate the performance of HiFreq-DETR across multiple benchmarks, including the SIMD and the large-scale HazyDet datasets, with the comparative results detailed in Tables 2 and 3, respectively.

Table 1: Comparison of detection results of various algorithms on the VisDrone-DET2019 val set.

Model	Publication	Inputsize	Params	FLOPS	AP (%)	AP50 (%)
<i>Two-stage methods</i>						
Faster R-CNN [14]	NeurIPS 2015	640×640	41.2M	127.0G	18.2	31.0
ARFP [15]	Appl Intell	1333×800	42.7M	193.8G	20.4	33.9
EMA [36]	ICASSP 2023	640×640	91.2M	–	30.4	49.7
SDP [16]	TGRS	800×800	96.7M	–	30.2	52.5
<i>One-stage methods</i>						
YOLOv12-M [20]	arXiv 2025	640×640	20.2M	67.5G	26.3	43.1
HIC-YOLOv5 [17]	ICRA 2024	640×640	10.5M	31.2G	26.0	44.3
MSFE-YOLO-L [18]	GRSL	640×640	41.6M	160.2G	29.0	46.8
YOLO-DCTI [28]	Remote Sensing	1024×1024	37.7M	–	27.4	49.8
Drone-YOLO-L [19]	Drones	640×640	76.2M	–	31.9	51.3
<i>End-to-end methods</i>						
DETR [3]	ECCV2020	1333×750	40.0M	187.1G	24.1	40.1
Deformable DETR [21]	ICLR2022	1333×800	40.2M	172.5G	27.1	43.1
RT-DETR-R18 [4]	CVPR2024	640×640	20.1M	60.0G	28.9	48.8
RT-DETR-R50 [4]	CVPR2024	640×640	41.8M	133.2G	30.9	50.7
HiFreq-DETR-base	–	640×640	21.1M	156G	32.8	54.1
HiFreq-DETR-lite	–	640×640	19.7M	121G	32.2	53.7

Note: Red and blue values represent the best and second-best performance.

Table 2: Comparative results before and after model improvement on SIMD dataset.

Model	AP_S	AP_M	AP_L	AP	AP_{50}
RT-DETR-R18	15.1	64.1	74.2	68.6	83.7
HiFreq-DETR	21.3	65.8	76.8	70.9	85.7

Note: All our models default to the base version unless the Lite variant is explicitly stated in the text.

Table 3: Comparative results before and after model improvement on HazyDet dataset.

Model	AP_S	AP_M	AP_L	AP	AP_{50}
RT-DETR-R18	20.8	55.8	76.8	54.9	76.0
HiFreq-DETR	25.7	61.3	77.5	58.8	79.8

1) Results on VisDrone-DET2019: As summarized in Table 1, the proposed HiFreq-DETR establishes a new state-of-the-art benchmark on the VisDrone-DET2019 validation set. The HiFreq-DETR-base configuration achieves a peak performance of 32.8% AP and 54.1% AP_{50} . Compared to traditional two-stage detectors such as Faster R-CNN and ARFP, our model exhibits a decisive advantage. In the realm of one-stage and specialized aerial detectors, HiFreq-DETR surpasses competitive models including MSFE-YOLO-L and Drone-YOLO-L with superior precision. Notably, when compared with Transformer-based end-to-end

architectures, HiFreq-DETR demonstrates substantial representational gains over the baseline RT-DETR-R18 and even the deeper RT-DETR-R50. To further optimize real-time processing and deployment efficiency, we introduce HiFreq-DETR-lite, which optimizes the high-resolution stages using depthwise separable operators. This variant successfully reduces the computational workload from 156 to 121G FLOPs and the parameter count to 19.7M, while still maintaining a high precision of 32.2% AP. As illustrated in the qualitative comparison in Fig. 3, HiFreq-DETR significantly suppresses both false negatives and false positives in high-density and cluttered environments. Compared to the baseline, our model demonstrates superior detection robustness and localization precision for tiny objects, effectively mitigating the missed detections of minuscule targets and confirming its robust potential for complex aerial surveillance tasks.

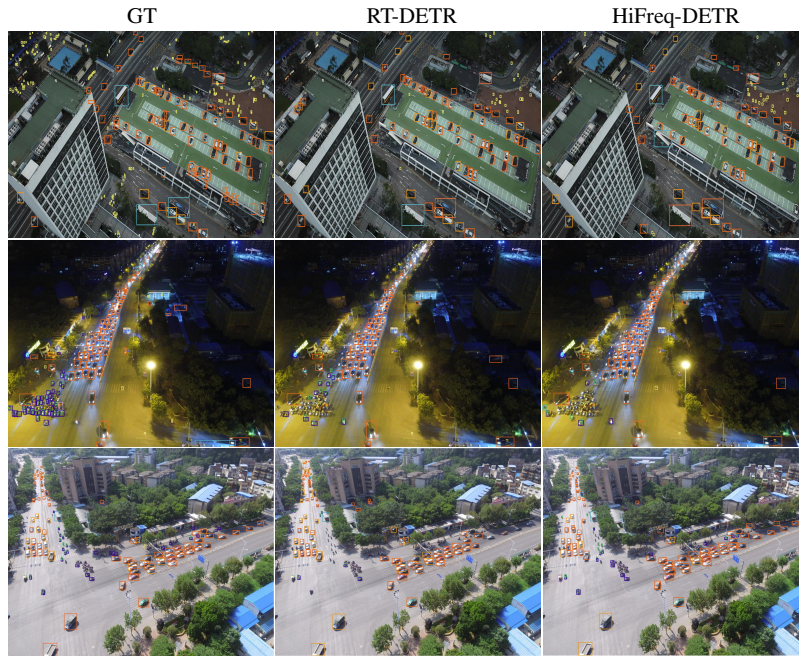


Figure 3: Visualization of detection results on VisDrone-DET2019. We compare the predictions of RT-DETR and HiFreq-DETR against the Ground Truth (GT) to highlight localization accuracy and robustness in high-density scenarios.

2) Results on SIMD: To further verify the generalization and robustness of the proposed architecture, we evaluate HiFreq-DETR on the SIMD dataset. As detailed in Table 2, HiFreq-DETR maintains a consistent performance advantage over the baseline RT-DETR-R18 across all evaluation metrics. Notably, our model achieves a substantial improvement in the detection of small-scale objects, where the AP_s increases from 15.1% to 21.3%. This significant jump underscores the model's superior sensitivity to minuscule targets amidst complex terrestrial backgrounds. Furthermore, the simultaneous improvements in medium and large-scale categories, with respective gains of 1.7% and 2.6%, indicate that the proposed structural optimizations enhance the overall representational capacity without compromising the localization of larger structures. Collectively, HiFreq-DETR elevates the overall AP from 68.6% to 70.9% and AP_{50} from 83.7% to 85.7%, confirming its robust adaptability and practical utility for cross-platform remote sensing tasks.

3) Results on HazyDet: To further evaluate the model's robustness under adverse weather conditions, we conduct experiments on the HazyDet dataset for hazy aerial scenarios. As demonstrated in Table 3, the proposed HiFreq-DETR achieves significant performance gains over the baseline RT-DETR-R18, with the AP improving from 54.9% to 58.8% and the AP_s increasing by 4.9%. These results indicate that by explicitly

preserving high-frequency structural details through the FS-AMFI and S_2 injection modules, our model effectively mitigates feature blurring and contrast degradation caused by atmospheric haze. The superior detection accuracy on HazyDet confirms the structural robustness of HiFreq-DETR and its capability to maintain high-fidelity localization in complex, low-visibility aerial environments.

4.3 Ablation Study

4.3.1 Effectiveness of Core Components

To systematically evaluate the contribution of each core component in HiFreq-DETR, we conducted ablation experiments on the VisDrone-DET2019 dataset, as summarized in Table 4. We employed a progressive evaluation strategy to analyze the impact of each module on both performance metrics and computational efficiency.

Table 4: Ablation study of HiFreq-DETR on the Visdrone-DET2019 dataset.

Exp	ResNeSt-14d	S_2	FS-AMFI	AP	AP ₅₀	AP _S	Params	FLOPs	FPS	Latency (ms)
1 (Baseline)	×	×	×	28.9	48.8	19.9	20.1M	60G	90.6	11.03
2	✓	×	×	29.1	49.2	20.1	19.3M	76G	80.3	12.46
3	✓	×	✓	29.9	51.0	21.4	18.9M	75G	84.8	11.79
4	✓	✓	×	32.3	53.1	23.8	21.5M	157G	65.0	15.38
5 (Ours-base)	✓	✓	✓	32.8	54.1	24.3	21.1M	156G	67.7	14.78
6 (Ours-lite)	✓	✓	✓	32.2	53.7	23.4	19.7M	121G	75.9	13.18

Note: The best results are highlighted in bold.

1) Backbone Optimization (Exps 1 & 2): Replacing the standard ResNet18 backbone with ResNeSt14d resulted in an AP increase from 28.9% to 29.1%. The Split-Attention mechanism and refined downsampling strategies within ResNeSt14d significantly enhance the backbone's ability to discriminate fine-grained features while maintaining a highly efficient architectural footprint. These results demonstrate that leveraging attention-based channel splitting provides higher-quality feature representations with superior parameter efficiency compared to standard deep convolutions. By preserving critical spatial details in the early stages, this optimized backbone establishes a more robust foundation for the subsequent high-resolution injection and multi-scale interaction modules.

2) Encoder Innovation (Exps 2 & 3): Replacing the baseline AIFI encoder with the proposed FS-AMFI module resulted in an AP improvement from 29.1% to 29.9%. This gain validates the superiority of FS-AMFI in feature modeling, which is primarily attributed to its dual-path enhancement: the DMFE component adaptively adjusts receptive fields to capture targets across varying scales, while the Frequency-selective Interaction (FS) explicitly preserves high-frequency structural signals that are typically smoothed in standard transformations.

3) High-Resolution Spatial Anchor Injection (Exps 3 & 4): The integration of the S_2 feature layer represents a critical advancement for HiFreq-DETR. As shown in Table 4, incorporating S_2 features boosted the AP from 29.1% to 32.3% and AP_S by a substantial 3.7%. This confirms the necessity of the S_2 layer as a spatial anchor for tiny object detection. Notably, the S_2 injection path introduces additional convolution modules within the CCFF network, which brings a marginal parameter increase. Even so, this configuration provides a favorable trade-off by significantly enhancing localization robustness. For minuscule targets, these

high-resolution details effectively compensate for the feature degradation caused by downsampling and provide the necessary geometric constraints for precise boundary regression.

4) Comprehensive Performance and Efficiency Analysis (Exps 4 & 5): The integration of all modules enables HiFreq-DETR to achieve optimal performance, reaching an AP of 32.8% and a significant AP_S of 24.3%. While the inclusion of the S_2 layer increases the total FLOPs, the substantial precision gain for minuscule objects justifies this computational reallocation as a necessary trade-off for high-fidelity remote sensing detection. Crucially, despite the increased computational overhead, HiFreq-DETR maintains high inference efficiency with a latency of 14.78 ms and an FPS of 67.7, which comfortably exceeds the real-time processing requirement (30 FPS) for UAV flight missions. Notably, this performance leap stems from the functional coupling between the FS-AMFI module and the S_2 injection path within a cascaded pipeline. Specifically, FS-AMFI ensures that fine-grained structural anchors are reinforced and preserved at the S_5 stage, preventing the signal annihilation typical of deep semantic branches. These frequency-enriched semantics are subsequently propagated top-down through the Feature Pyramid Network (FPN) to the S_2 scale, which provides the high-resolution grid necessary to manifest these sharp localization cues. Without this synergy, standard upsampling would propagate semantically rich but spatially smoothed features, thereby compromising the regression accuracy for tiny targets. The experimental results confirm that the interdependence of ResNeSt14d, FS-AMFI, and S_2 injection collectively establishes a robust framework for maintaining both semantic discriminability and spatial fidelity in complex aerial environments.

5) Real-Time Optimization with Lite Configuration (Exps 5 & 6): To further optimize the computational efficiency, we evaluate the HiFreq-DETR-lite variant. By substituting dense convolutions in the S_2 stage with depthwise separable operators, the computational load is reduced from 156 to 121G FLOPs, accompanied by a reduction in parameters to 19.7M. Consequently, the FPS improves to 75.9 and the latency drops to 13.18 ms, with only a marginal 0.6% decrease in AP. This demonstrates that our architecture provides a flexible trade-off between high-speed inference and superior detection precision for tiny objects.

4.3.2 Analysis of Downsampling Strategies

To rigorously justify the micro-architectural configurations within the FS-AMFI module, we investigate the influence of various spatial downsampling operators employed for the construction of Key (K) and Value (V) sequences. A comparative analysis was conducted by replacing the default maximum pooling (MaxPool) with average pooling (AvgPool) and a learnable strided convolution (StridedConv).

As summarized in Table 5, the MaxPool-based strategy achieves the optimal detection precision across all metrics, yielding a superior mAP of 32.8% and a small-object accuracy of 24.3%. This advantage stems from the intrinsic sparsity of tiny-object features in UAV-perspective imagery. Since minuscule targets typically occupy minimal spatial extent, MaxPool functions as a saliency-preserving mechanism that effectively captures peak activations within the pooling kernel. In contrast, the AvgPool operator acts as a spatial smoother that dilutes sparse signals by averaging them with surrounding background noise, leading to a degradation in sensitivity for small instances. Notably, while the StridedConv variant introduces additional learnable parameters and computational overhead, it fails to surpass the non-parametric MaxPool in overall precision. This empirical evidence suggests that for extremely small-scale targets, a fixed saliency-based prior is more robust than a learnable sampling paradigm, which may encounter convergence difficulties when modeling highly sparse structural cues. Furthermore, the use of MaxPool ensures maximum computational efficiency for the cross-attention mechanism, as it avoids the weight-update requirements of convolutional layers. Given that our asymmetric design maintains the high-resolution Query (Q) path for precise localization, the results confirm that MaxPool is the most effective operator for providing condensed contextual anchors without compromising the structural integrity of the high-resolution queries.

Table 5: Ablation study on different downsampling strategies.

Strategy	AP	AP ₅₀	AP _S	AP _M	AP _L
AvgPool	32.6	53.8	24.0	43.7	52.8
StridedConv	32.6	53.9	24.1	43.8	53.1
MaxPool (Ours)	32.8	54.1	24.3	44.0	53.8

4.3.3 Analysis of Boundary Localization Precision

To evaluate localization quality, we analyze model performance across two scale subsets: Small (area $< 32^2$ px²) and All (full distribution). Precision is quantified using Mean Center Point Error (CPE) and IoU Variance, which respectively characterize spatial alignment accuracy and regression stability.

As summarized in Table 6, HiFreq-DETR exhibits superior localization robustness. Specifically, for the Small category, the Mean CPE is reduced from 1.88 to 1.69 px, representing a 10.1% improvement in spatial alignment. This reduction confirms that the high-resolution spatial anchors from the S_2 layer effectively mitigate the quantization errors and information loss inherent in deep feature downsampling. Furthermore, the decrease in IoU Variance from 0.0128 to 0.0122 and the consistent gain in Mean IoU indicate a more stable and precise regression process, effectively suppressing the prediction drifting common in tiny-object detection. Notably, HiFreq-DETR captures 1144 additional small instances while maintaining high boundary fidelity. These quantitative optimizations provide rigorous empirical support for the visual analysis presented in Fig. 3, collectively demonstrating that preserving high-frequency spatial signals is critical for precise boundary regression in complex aerial scenes.

Table 6: Quantitative localization analysis on the VisDrone validation set.

Model	Scale	Nmatch $\uparrow$	Mean CPE (px) $\downarrow$	Mean IoU (%) $\uparrow$	IoU Var $\downarrow$
RT-DETR	Small	19,366	1.88	73.12	0.0128
	All	30,800	2.16	77.66	0.0154
HiFreq-DETR	Small	20,510	1.69	74.41	0.0122
	All	31,988	1.95	78.66	0.0144

Note: CPE refers to the L_2 distance between matched centers; Mean IoU and IoU Var are calculated based on matched prediction-target pairs. Greedy matching is applied with an IoU threshold of 0.5 and confidence threshold of 0.3. Scale definitions follow the COCO standard.

5 Conclusion

This study introduces HiFreq-DETR, a specialized architecture designed to address the critical bottleneck of tiny object detection in UAV-perspective imagery. By synergizing ResNeSt14d with a high-resolution S_2 spatial anchor injection mechanism, we effectively mitigate the signal dissipation inherent in deep downsampling, preserving the structural integrity of targets at the source. Central to this framework, the FS-AMFI module decouples object saliency from low-frequency background noise through frequency-aware interaction, overcoming the contextual smoothing typical of standard Transformers. While the base HiFreq-DETR establishes a new performance benchmark, its lightweight variant significantly reduces computational complexity while maintaining robust detection precision. Experimental results on the VisDrone, SIMD, and HazyDet benchmarks demonstrate that our framework achieves state-of-the-art performance while maintaining favorable inference efficiency. Looking ahead, our research will focus on enhancing model robustness under extreme meteorological conditions via domain-adaptive fusion. Furthermore, we aim to

extend the frequency-selective paradigm to event-based data and large-scale multi-modal tasks, exploring its potential for high-fidelity sensing in high-speed, dynamic UAV applications.

Acknowledgement: Not applicable.

Funding Statement: This work was supported in part by the Funds for Central-Guided Local Science & Technology Development (Grant No. 202407AC110005) Key Technologies for the Construction of a Whole-Process Intelligent Service System for Neuroendocrine Neoplasm; in part by the Key Technology Research and Application of Cross-domain Automatic Business Collaboration in Smart Tourism (No. XYCY-CYCX-2022-0005); and in part by the Yunnan Provincial Department of Education's Enterprise-Proposed Problem-Solving Project, "Research and Application Demonstration of Urban Low-Altitude IoT Intelligent Service System" (Project No. FWCY-QYCT2025001).

Author Contributions: Linyu Dong and Hao Li are responsible for the thesis manuscript and experiments, while Tao Li is mainly in charge of the data. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets utilized in this study are publicly available benchmarks for object detection, and their official access links are provided below:

- VisDrone-DET2019: <https://github.com/VisDrone/VisDrone-Dataset>
- SIMD: <https://github.com/ihians/simd>
- HazyDet: <https://github.com/GrokCV/HazyDet>.

Interested researchers can access the raw images and official annotations from the above official repositories in accordance with the terms and conditions specified by the dataset providers.

Ethics Approval: Not applicable.

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

References

1. Cheng G, Xie X, Han J, Guo L, Xia GS. Remote sensing image scene classification meets deep learning: challenges, methods, benchmarks, and opportunities. *IEEE J Selc Top Appl Earth Observat Remote Sens.* 2020;13:3735–56. doi:10.1109/jstars.2020.3005403.
2. Li K, Wan G, Cheng G, Meng L, Han J. Object detection in optical remote sensing images: a survey and a new benchmark. *ISPRS J Photogramm Remote Sens.* 2020;159:296–307.
3. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: *Computer vision—ECCV 2020*. Cham, Switzerland: Springer; 2020. p. 213–29.
4. Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, et al. DETRs beat YOLOs on real-time object detection. *arXiv:2304.08069*. 2024.
5. Du D, Zhu P, Wen L, Bian X, Lin H, Hu Q, et al. VisDrone-DET2019: the vision meets drone object detection in image challenge results. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. Piscataway, NJ, USA: IEEE; 2019. p. 213–26.
6. Wang J, Yang W, Guo H, Zhang R, Xia G. Tiny object detection in aerial images. In: *2020 25th International Conference on Pattern Recognition*. Piscataway, NJ, USA: IEEE; 2021. p. 3791–8.
7. Sun Z, Zhang G, Xing Y, Liu Y. A scale-adaptive aggregation and multi-domain feature fusion architecture for small-target detection in UAV aerial imagery. *Sensors.* 2026;26(5):1610. doi:10.3390/s26051610.
8. Sunkara R, Luo T. No more strided convolutions or pooling: a New CNN building block for low-resolution images and small objects. *arXiv:2208.03641*. 2022.
9. Wang A, Xu Y, Wang H, Wu Z, Wei Z. CDE-DETR: a real-time end-to-end high-resolution remote sensing object detection method based on RT-DETR. In: *2024 IEEE International Geoscience and Remote Sensing Symposium*. Piscataway, NJ, USA: IEEE; 2024. p. 8090–4.

10. Zhang H, Wu C, Zhang Z, Zhu Y, Lin H, Zhang Z, et al. ResNeSt: split-attention networks. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway, NJ, USA: IEEE; 2022. p. 2735–45.
11. Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans Pattern Analy Mach Intell.* 2021;43(10):3349–64. doi:10.1109/tpami.2020.2983686.
12. Alazeb A, Chughtai BR, Al Mudawi N, AlQahtani Y, Alonazi M, Aljuaid H, et al. Remote intelligent perception system for multi-object detection. *Front Neurorobot.* 2024;18:1398703. doi:10.3389/fnbot.2024.1398703.
13. Alshehri M, Xue T, Mujtaba G, AlQahtani Y, Almujaally NA, Jalal A, et al. Integrated neural network framework for multi-object detection and recognition using UAV imagery. *Front Neurorobot.* 2025;19:1643011. doi:10.3389/fnbot.2025.1643011.
14. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Analy Mach Intell.* 2017;39(6):1137–49. doi:10.1109/TPAMI.2016.2577031.
15. Wang J, Yu J, He Z. ARFP: a novel adaptive recursive feature pyramid for object detection in aerial images. *Appl Intell.* 2022;52(11):12844–59. doi:10.1007/s10489-021-03147-.
16. Liu J, Wu Z, Xiao L. A spectral diffusion prior for unsupervised hyperspectral image super-resolution. *IEEE Trans Geosci Remote Sens.* 2024;62(1):5528613. doi:10.1109/tgrs.2024.3449073.
17. Tang S, Zhang S, Fang Y. HIC-YOLOv5: improved YOLOv5 for small object detection. In: 2024 IEEE International Conference on Robotics and Automation. Piscataway, NJ, USA: IEEE; 2024. p. 6614–9.
18. Qi S, Song X, Shang T, Hu X, Han K. MSFE-YOLO: an improved YOLOv8 network for object detection on drone view. *IEEE Geosci Remote Sens Lett.* 2024;21:6013605. doi:10.1109/LGRS.2024.3432536.
19. Zhang Z. Drone-YOLO: an efficient neural network method for target detection in drone images. *Drones.* 2023;7(8):526.
20. Tian Y, Ye Q, Doermann D. YOLOv12: attention-centric real-time object detectors. arXiv:2502.12524. 2025.
21. Zhu X, Su W, Lu L, Li B, Wang X, Dai J, et al. Deformable DETR: deformable transformers for end-to-end object detection. arXiv:2010.04159. 2021.
22. Peng Y, Li H, Wu P, Zhang Y, Sun X, Wu F. D-FINE: redefine regression task in DETRs as fine-grained distribution refinement. arXiv:2410.13842. 2024.
23. Wang Z, Li C, Xu H, Zhu X, Li H. Mamba YOLO: a simple baseline for object detection with state space model. arXiv:2406.05835. 2024.
24. Sun F, Cheng D, Zheng P, Song T, Chen L, Kou Q. TGCADNet: text-guided context-aware detection via CLIP for small objects in UAV scenes. *IEEE Trans Circuits Syst Video Technol.* 2026;1. doi:10.1109/TCSVT.2026.3662475.
25. Deng S, Li S, Xie K, Song W, Liao X, Hao A, et al. A global-local self-adaptive network for drone-view object detection. *IEEE Trans Image Process.* 2021;30:1556–69. doi:10.1109/tip.2020.3045636.
26. Wang J, Xu C, Yang W, Yu L. A normalized gaussian wasserstein distance for tiny object detection. arXiv:2110.13389. 2022.
27. Zhou W, Zhu Y, Lei J, Wan J, Yu L. CCAFNet: crossflow and cross-scale adaptive fusion network for detecting salient objects in RGB-D images. *IEEE Trans Multim.* 2022;24:2192–204.
28. Min L, Fan Z, Lv Q, Reda M, Shen L, Wang B. YOLO-DCTI: small object detection in remote sensing base on contextual transformer enhancement. *Remote Sens.* 2023;15(16):3970. doi:10.3390/rs15163970.
29. Zhou W, Sun F, Jiang Q, Cong R, Hwang JN. WaveNet: wavelet network with knowledge distillation for RGB-T salient object detection. *IEEE Trans Image Process.* 2023;32:3027–39. doi:10.1109/TIP.2023.3275538.
30. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2016. p. 770–8.
31. Wightman R. PyTorch image models. GitHub. 2019 [cited 2026 May 10]. Available from: <https://github.com/rwightman/pytorch-image-models>.
32. He T, Zhang Z, Zhang H, Zhang Z, Xie J, Li M. Bag of tricks for image classification with convolutional neural networks. arXiv:1812.01187. 2018.
33. Haroon M, Shahzad M, Fraz MM. Multisized object detection using spaceborne optical imagery. *IEEE J Selec Top Appl Earth Observat Remote Sens.* 2020;13:3032–46. doi:10.1109/jstars.2020.3000317.

34. Feng C, Chen Z, Li X, Wang C, Yang J, Cheng MM, et al. HazyDet: open-source benchmark for drone-view object detection with depth-cues in hazy scenes. arXiv:2409.19833. 2025.
35. Lin TY, Maire M, Belongie S, Bourdev LD, Girshick R, Hays J, et al. Microsoft COCO: common objects in context. arXiv:1405.0312. 2015.
36. Ouyang D, He S, Zhang G, Luo M, Guo H, Zhan J, et al. Efficient multi-scale attention module with cross-spatial learning. In: 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ, USA: IEEE; 2023. p. 1–5.