MDPI

Article

HSP-DETR: Hierarchical Sparse Perception Transformer for Real-Time Aerial Object Detection

Qinyu Liu [1], Min Xu [1] and Lei Liao [2,*]

[1] Key Laboratory of Wireless Sensor Networks, Sichuan Normal University, Chengdu 610101, China
[2] School of Physics and Electronic Engineering, Sichuan Normal University, Chengdu 610101, China
* Correspondence: liaolei@sicnu.edu.cn

Abstract

The detection of small objects in unmanned aerial vehicle (UAV) imagery is critical but remains challenging due to severe feature loss during downsampling, interference in multi-scale feature fusion, and the computational constraints of onboard devices. To address these issues, this paper proposes HSP-DETR, a method featuring a hybrid backbone network LSNet for hierarchical feature extraction. A Detail-preserving Local-global Enrichment Pyramid (DLEP) module is designed to mitigate texture loss by employing SPDConv for non-destructive downsampling and triple-path fusion. Furthermore, a Lightweight Hierarchical Channel-wise Gating Fusion (LHCGF) mechanism enables efficient cross-scale feature interaction. During training, a Scale-Adaptive Quality Loss (SAQL) function dynamically adjusts weights to enhance supervision for small objects. Experiments on the VisDrone2019 dataset show that HSP-DETR outperforms RT-DETR, improving APs by 2.9%, mAP50 by 2.8%, and mAP50:95 by 4.1%, while reducing GFLOPs by 29.52%. The model achieves an optimal balance between accuracy and efficiency for deployment on resource-constrained UAV platforms.

Keywords: UAV; small object detection; lightweight model; feature fusion

1. Introduction

In recent years, unmanned aerial vehicle (UAV) technology has advanced rapidly and has been widely applied in critical domains such as traffic monitoring, agricultural mapping, disaster relief, and public safety. Equipped with high-precision sensors, UAVs can capture large-scale, high-resolution images from an aerial perspective. Compared with satellite imagery, UAV imagery is easier to acquire and simpler to operate, thus offering unique advantages. However, due to constraints such as small object sizes, diverse shapes, uneven spatial distributions, and severe occlusions, object detection in UAV aerial images remains challenging.

Traditional aerial object detection algorithms mainly rely on sliding windows [1] or region proposals [2]. Using HOG [3], color histograms [4], local binary patterns [5], and subsequently the Deformable Part Model (DPM) [6], stage-wise progress was made in recognizing vehicles, pedestrians, ships, and other objects. Selective Search, proposed by J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders [2], is a region proposal method that generates initial regions via image segmentation and then merges regions based on features to obtain candidate regions, thereby alleviating the computational redundancy of exhaustive sliding windows. However, traditional aerial object detection

Received:
Revised:
Accepted:
Published:

Citation: Liu, Q.; Xu, M.; Liao, L. Title. *Sensors* **2025**, *1*, 0. https://doi.org/

still struggles to effectively recognize target variations in complex environments, and both accuracy and speed fall short of practical application requirements.

Over the past decade, with the rapid development of deep learning, traditional methods have gradually been supplanted by deep learning. Many mainstream object detection frameworks based on deep convolutional neural network (CNN) architectures are typically categorized into single-stage and two-stage detectors. Single-stage methods such as the YOLO series [7–16] and SSD [17] treat detection as classification and regression on feature maps, significantly reducing latency. Two-stage methods such as Faster R-CNN [18] and Cascade R-CNN [19] first perform background discrimination to generate candidate boxes, followed by category classification and bounding box regression on the candidate regions. Although two-stage detectors achieve higher accuracy than single-stage ones, their complex architectures and substantial computational demands make them unsuitable for the real-time and lightweight requirements of UAV platforms. In contrast, single-stage detectors are constrained by convolutional operations, particularly limited receptive fields, which hinder effective detection in large-scale images from UAV perspectives. RetinaNet [20] mitigates the foreground–background class imbalance by introducing Focal Loss to down-weight easy examples, directing the model's focus toward hard samples and thereby improving detection accuracy.

Given the computational constraints of UAV platforms and the challenges of detecting small objects in aerial imagery, researchers have adopted various lightweight optimization methods. THP-YOLO [21] enhances small-object feature extraction through a pyramid Transformer head, improving detection performance for small objects in complex scenes. FasterNet [22] reduces redundant computation and memory access via Partial Convolution, efficiently extracting spatial features and increasing throughput while maintaining performance.

In recent years, due to its unique advantages of end-to-end design, global contextual understanding, and long-range dependency modeling, researchers have turned their attention to the Transformer architecture. In 2020, Carion et al. proposed End-to-End Object Detection with Transformers (DETR) [23], which first successfully applied the Transformer architecture to object detection, demonstrating its potential in vision tasks. To address the issue of uniformly distributed attention maps, DAB-DETR [24] explicitly modeled object queries using dynamic anchor boxes to replace the original latent vectors. Building upon DAB-DETR, Feng Li, Hao Zhang et al. introduced DN-DETR [25], which incorporated noisy ground-truth bounding boxes during training to accelerate convergence and avoid the instability of traditional bipartite matching. The self-attention mechanism within the DETR encoder requires computation over the entire feature map, leading to significant computational overhead when processing images. To reduce computational cost, Wang Tao, Yuan Li et al. designed the Poll and Pool sampling module in PnP-DETR [26], which abstracted the image feature map into refined object feature vectors and a small number of background context feature vectors, thereby reducing the feature volume processed by the Transformer and mitigating redundant spatial computations. Subsequently, Zhao et al. proposed the Real-Time DETR (RT-DETR) [27], further enhancing the speed and accuracy of DETR-based detectors and achieving a balance between real-time performance and precision.

Despite the progress in Transformer-based object detection, their performance degrades in UAV scenarios due to network-related factors, namely cumbersome backbone design, simplistic feature fusion, and high computational overhead. To enhance the applicability of Transformers to UAV aerial imagery, this paper proposes the Hierarchical Sparse Perception Transformer (HSP-DETR), a lightweight and efficient algorithm based on RT-DETR that integrates channel attention and multi-scale features. It preserves the

end-to-end advantage of DETR while leveraging a lightweight yet powerful network architecture to efficiently extract multi-scale features, maximally retain small-object detail, and achieve stronger cross-scale feature fusion.

The main contributions of this work are summarized as follows:

1. We design a novel small object detection module, Detail-preserving Local-global Enrichment Pyramid (DLEP), which replaces conventional convolution with an SPD-Conv lossless downsampling strategy, embeds Hierarchical Edge-preserving Feature Enhancement (HEFE) into the neck network to concatenate multi-feature maps along channels, and adaptively amplifies target-relevant fine-grained details while performing long-range modeling, thereby laying the foundation for effectively distinguishing targets from background. This addresses the texture blurring and insufficient receptive field caused by downsampling for small objects in aerial scenarios, enhancing the model's representation capability for tiny targets.
2. To overcome the limitations of traditional feature fusion in detail preservation and scale adaptability, we propose a cross-scale feature fusion module, Lightweight Hierarchical Channel-wise Gating Fusion (LHCGF), to replace conventional fusion schemes. LHCGF enables bidirectional interaction between high-frequency detail information and cross-scale features, and before sending features into the decoder, adaptively weights and merges multi-scale features through the LHCGF module. This achieves fine-grained and adaptive feature fusion, further refines feature representations, and effectively preserves edge and texture information in images.
3. To resolve the insensitivity of RT-DETR to small object scales, we introduce Scale-Adaptive Quality Loss (SAQL) into the training process. SAQL dynamically adjusts loss weights based on target size, making the model more sensitive to small targets and capable of suppressing background noise, thereby improving detection accuracy for small-scale objects and environmental robustness.
4. We propose a novel Asymmetric Hierarchical Synergy (ASH) design strategy inspired by biological vision systems. ASH formulates network design as a multi-objective optimization problem and asymmetrically allocates computational resources, dynamically focusing more capacity on high-frequency details and contextual features crucial for small object detection. Redundant information is processed efficiently using lightweight modules, thereby achieving a superior Pareto frontier overall.

2. Related Work

2.1. Transformer-based Object Detection

With the introduction of attention mechanisms, Transformer-based object detection methods have garnered widespread attention. DETR [23] removes reliance on handcrafted components such as non-maximum suppression (NMS), constructing an end-to-end detector. However, its slow training process and inadequate performance on small objects constrain its further development. Subsequently, Deformable DETR [28] employs deformable attention to significantly accelerate convergence while improving performance on small objects and in dense scenes; DN-DETR [25] and DINO [29] enhance efficiency and accuracy through improved query strategies, denoising training, and anchor priors. Unlike prior Transformer-based detectors, RT-DETR [27] is a real-time object detection network that adopts hybrid encoding and IoU-aware query selection to speed up inference. Although these methods have progressed rapidly in accuracy and real-time performance, they primarily focus on spatial and contextual features while comparatively neglecting channel-wise feature refinement, leading to potential detail loss and insufficient cross-scale fusion. This leaves room for improvement in applying Transformers for object detection in aerial imagery.

2.2. Multi-Scale Feature Fusion

High-precision object detection, especially small object detection, faces the challenge of effectively handling the scale diversity of targets in images. The Feature Pyramid Network (FPN) [30] established a cross-scale fusion framework by propagating high-level semantics to lower layers; PANet [31] introduced an additional bottom-up path to enhance low-level details; ASFF [32] employed adaptive fusion to address inter-level feature conflicts. Recent work has further progressed toward attention-based fusion. DetectoRS [33] expanded the effective receptive field through a recursive feature pyramid; Dynamic Head [34] unified feature-level, spatial-level, and channel-level attention within the detection head; InternImage [35] leveraged stacked deformable convolutions and multilayer perceptrons to achieve adaptive spatial aggregation. While effective, most existing multi-scale modules primarily focus on spatial and semantic alignment, with limited discussion on the importance of channel-wise interactions in cross-scale fusion. Moreover, high-frequency details in aerial imagery are prone to being lost during scale aggregation.

2.3. Channel Attention Mechanisms

Channel attention mechanisms assign differentiated weights to feature map channels, enabling the network to concentrate capacity on informative channels while suppressing redundancy. SENet [36] introduced the squeeze-and-excitation paradigm: global average pooling first aggregates spatial responses into a channel descriptor (squeeze), followed by a two-layer multilayer perceptron that produces channel-wise gating weights (excitation). CBAM [37] regards channel and spatial attention as complementary and applies them sequentially, enabling the model to aggregate richer contextual information. ECA-Net [38] replaces dimensionality reduction with a lightweight one-dimensional convolution to capture local cross-channel interactions, thereby reducing information loss while keeping model complexity low. Coordinate Attention [39] embeds positional information into channel attention by decomposing global pooling along the horizontal and vertical axes, enhancing directionality and positional sensitivity in the learned attention.

2.4. Small Object Detection in Aerial Imagery

In typical aerial image datasets such as VisDrone2019 [40] and HIT-UAV [41], objects often occupy only a few pixels, exhibiting characteristics such as dense distribution, large scale variations, and strong background noise interference. These factors collectively make object detection in aerial images more challenging than in conventional scenarios. Researchers have developed various approaches to improve model performance on small object detection. For instance, FPN [30] and its variants, such as PANet [31] and BiFPN (employed in EfficientDet) [42], enhance the model's feature representation through multi-scale fusion. To strengthen high-frequency and edge detail preservation, methods like Holistically-Nested Edge Detection (HED) [43] and Richer Convolutional Features (RCF) [44] incorporate explicit edge supervision. Nevertheless, achieving an optimal trade-off between efficiency and practical utility remains a critical challenge for small object detection in aerial imagery. To address this, we integrate a lightweight biomimetic backbone, a cross-scale detail injection mechanism, and an adaptive multi-scale feature fusion strategy to improve robustness. HSP-DETR demonstrates superior robustness on UAV benchmarks, underscoring its practical applicability in complex real-world scenarios.

3. Methodology

3.1. Overall architecture

The proposed HSP-DETR (Figure 1) aims to enable efficient and fine-grained detection of small objects in aerial images, while also ensuring effective deployment on edge devices

and reducing deployment complexity. The original RT-DETR employs ResNet [45] as its backbone network. However, ResNet suffers from significant feature loss after multiple downsampling operations, making small objects indistinguishable and thus difficult to detect. Additionally, its inadequate design for multi-scale features limits effective handling of multi-scale information. Moreover, the substantial computational resources required by ResNet lead to a bloated RT-DETR model. Therefore, we adopt the biologically-inspired LSNet [46] as our feature extraction backbone, which not only substantially reduces computational complexity but also effectively extracts multi-scale features, providing rich information for subsequent feature fusion. To further preserve shallow details critical for small object detection, we feed feature maps from different levels into the HEFE module via the DLEP module for semantic feedback, achieving cross-scale alignment and adaptive fusion through sparse self-attention. We design the LHCGF module, which is deployed at various stages of the network to adaptively fuse multi-scale features from diverse sources. While retaining the RT-DETR decoder, we introduce the SAQL loss into the training objective. This loss dynamically adjusts its weight based on the matching quality between predicted and ground-truth boxes, thereby improving the matching quality for small objects and accelerating model convergence.

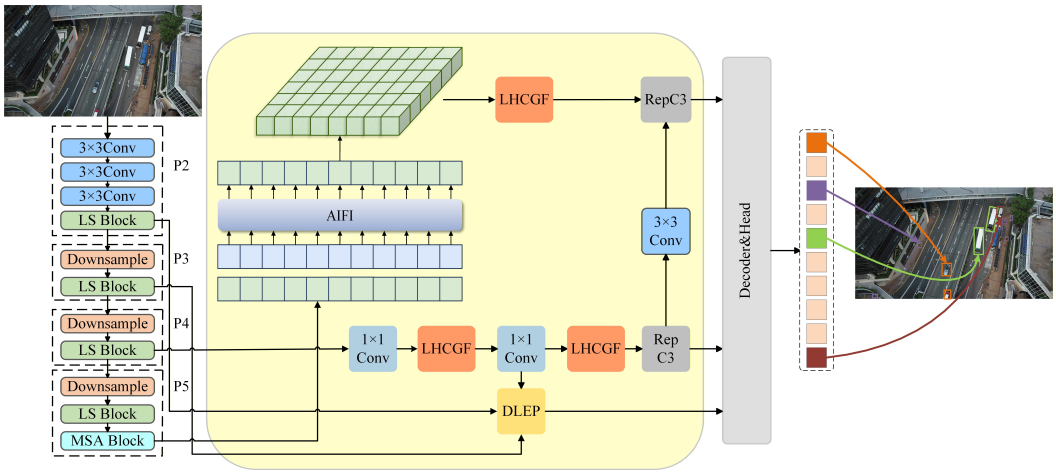


Figure 1. Overall architecture of HSP-DETR.

3.2. DLEP

In UAV aerial imaging scenarios, small-sized objects are not only densely distributed but also highly sensitive to high-frequency edge and texture features. Standard convolution and the original CNN-based cross-scale feature fusion (CCFF) operations can degrade these edge and texture details during downsampling, causing small objects to vanish during feature propagation. Motivated by this observation, we improve upon CCFF and design the DLEP module, which integrates feature enhancement and dynamic fusion into a unified block, as illustrated in Figure 2. Specifically, the P2 feature map from the backbone network is processed through SPDConv to obtain features enriched with small-object information. The resulting features are then concatenated with the P3 features from the backbone and the context-enhanced features derived from the deeper P4 layer (denoted as Feature(P4)) after a series of transformations. This three-branch concatenated fusion is subsequently processed by the HEFE module to achieve effective multi-scale feature integration. The proposed design reduces both the number of parameters and FLOPs, improves the preservation of shallow details, and alleviates the problem of edge and texture feature loss during downsampling and cross-layer propagation.

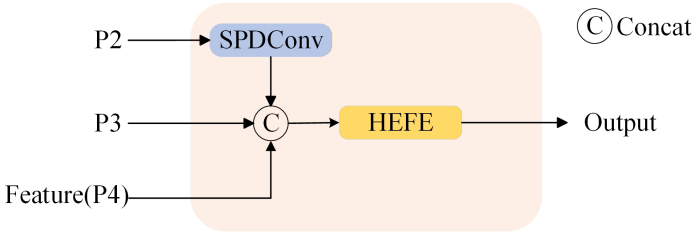


Figure 2. Architecture diagram of the DLEP module.

3.2.1. SPDConv

Standard convolution inevitably discards a large amount of high-frequency information during feature sampling; the loss of these critical details impairs the correctness of small object detection and degrades model performance. To address this issue, we introduce SPDConv to replace standard convolution. SPDConv, namely space-to-depth convolution, can effectively preserve the high-frequency details of the feature map while enhancing the feature representation capability for small objects. It consists of a space-to-depth transformation layer followed sequentially by a stride-1 convolution layer. The structure of SPDConv is shown in Figure 3.

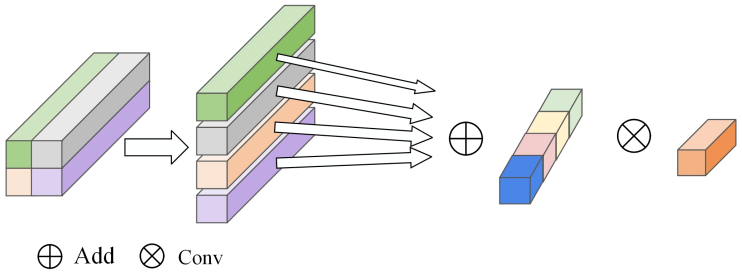


Figure 3. Illustration of SPD-Conv.

Specifically, SPDConv extracts information from four distinct spatial partitions to construct a spatial pyramid. Let the input feature map $\mathbf{X}$ have a shape of (B, H, W, C), where B is the batch size, H and W are the height and width, C is the channel number, and `scale` is the downsampling factor.

The operation first samples features with a stride of `scale`, generating `scale`^{2} sub-images. For any indices (i, j) where $i, j \in \{0, 1, \ldots, \text{scale} - 1\}$, the corresponding sub-image $\mathbf{X}_{i,j}$ is formulated as:

$$f_{i,j}[b, h, w, c] = X[b, \text{scale} \cdot h + i, \text{scale} \cdot w + j, c] \tag{1}$$

where $b \in \{0, 1, ..., B - 1\}$ is the batch index, $h \in \{0, 1, ..., \frac{H}{\text{scale}} - 1\}$ is the height index, $w \in \{0, 1, ..., \frac{W}{\text{scale}} - 1\}$ is the width index, and $c \in \{0, 1, ..., C - 1\}$ is the channel index.

All sub-images are then concatenated along the feature channel dimension to obtain the intermediate feature map X', a process that ensures that spatial information is more completely preserved in the channel dimension. This is expressed as:

$$X' = \text{concat}\Big(\{f_{i,j} \mid i, j \in \{0, 1, \ldots, \text{scale} - 1\}\}\Big) \tag{2}$$

Finally, a standard convolutional layer with stride 1 is applied to the concatenated feature map to adjust the output size and ensure no information is lost.

3.2.2. HEFE

Although SPDConv preserves the texture and edge details of features, the inherent receptive field limitation of the P3 layer persists. While the P3 layer is rich in high-resolution spatial details, it lacks the semantic context necessary for object discrimination; other layers possess strong semantic expressiveness but are deficient in fine-grained information. To overcome this inherent limitation, we design the HEFE module, which performs three-way feature fusion through a linear attention mechanism and local detail enhancement. HEFE adopts a density-aware dynamic region selection mechanism to adaptively fuse features according to the object density of the input image, achieving efficient cross-scale information injection without sacrificing robustness or incurring additional computational cost.

Specifically, we first denote the low-level high-resolution feature as $X \in \mathbb{R}^{C\times H\times W}$, and the high-level low-resolution feature as $U \in \mathbb{R}^{\frac{C\times H\times W}{s}}$ (where s is the downsampling factor). We then perform block average pooling on the high-level low-resolution feature to obtain the sparse context feature $C = \Phi(U)$, apply a 1×1 convolution to map X and U to the query Q and key-value K, V, and finally employ linear attention to achieve cross-scale interaction and semantic enhancement, as follows:

$$Y_{\text{att}} = \Phi\left(\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V\right), \quad Q = \Psi_q(W_qX),\ K = \Psi_k(W_kC),\ V = \Psi_v(W_vC) \tag{3}$$

where Φ denotes multi-head rearrangement, $X \in \mathbb{R}^{C\times H\times W}$, W_q, W_k, W_v are 1×1 convolutions, and Ψ_q, Ψ_k, Ψ_v denote flattening and block partitioning. To enable HEFE to better recognize and detect fine-grained structures and to enhance its detection capability in target-dense regions, we employ channel gating to generate adaptive fusion weights g, which weight the importance of texture information. The computation of g is as follows:

$$g = \sigma(W_z\phi(W_1\text{GAP}([L(X), Y_{\text{att}}]))) \tag{4}$$

where σ is the Sigmoid activation function, ϕ is the ReLU activation function, W_1, W_2 are 1×1 convolutions, $L(X)$ applies depthwise separable convolutions with different kernel sizes to the original low-level features, and GAP denotes global average pooling. The gating employs two 1×1 convolutions to avoid excessive computational overhead while preserving expressive power, and compensates for the lack of fine details in high-level features and insufficient localization in low-level features through global–local collaboration.

Subsequently, we quantitatively inject cross-scale semantics into the channels and spatial regions that require semantic reinforcement, preserving structural responses locally, reducing edge blurring caused by semantic diffusion, and preventing edge information from being overridden by semantics. Specifically, after obtaining the gating weights g, we adaptively distribute them between the semantic enhancement attention branch Y_{att} and the local detail attention branch $L(X)$ according to the proportion of channel semantics, and selectively inject cross-scale semantics based on the different weights, suppressing low-level noise interference and enhancing the model's robustness in densely crowded small-object scenarios. The formulation is as follows:

$$Y = X + g \odot Y_{\text{att}} + (1 - g) \odot L(X) \tag{5}$$

3.3. LHCGF

Most network architectures fuse the multi-scale feature maps generated by the backbone simply by concatenation; although this does not increase the number of parameters, it easily introduces irrelevant gradient noise and causes channel redundancy, ultimately

making it difficult to capture key information and local features in complex scenes. We observe that in aerial images target scales often vary dramatically while semantic correlations across channels remain stable. Existing spatial-attention-based multi-scale fusion modules can efficiently model local–global relationships, but their computational complexity is high, making real-time inference on edge devices difficult. In addition, redundant spatial attention at low-resolution layers expends substantial computation for coarse-grained feature fusion without a corresponding performance gain. Therefore, we design a lightweight hierarchical channel-gated fusion module, whose structure is shown in Figure 4. This computationally efficient and structurally concise adaptive feature fusion architecture better handles channel information and suppresses redundant features. Abandoning spatial attention modeling, we instead model channel semantic attention: global pooling is used to obtain channel statistics, a nonlinear transformation produces gating weights, preserving cross-scale feature alignment, avoiding background interference from spatial attention, and capturing critical channel information.

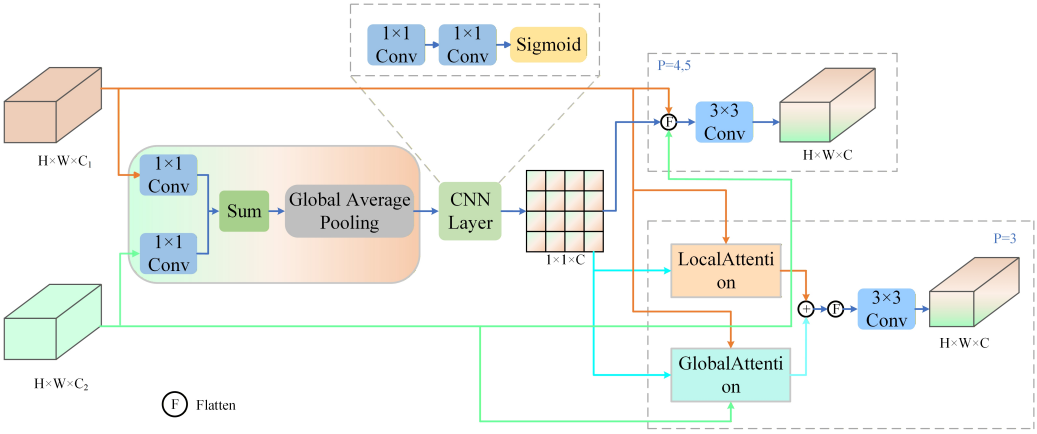


Figure 4. Architecture diagram of the LHCGF module.

As shown in Figure 4, LHCGF first aligns the channels of the two input feature maps at different scales with a 1×1 convolution layer to facilitate subsequent operations. It then applies an AND-based unified selection gate to avoid computational redundancy of the two gating paths and reduce noise. Next, global average pooling is used to compress spatial information, providing channel statistics and a global contextual summary. The first 1×1 convolution layer can dynamically set the compression ratio according to the feature map level in which the target resides, based on the variance of the input feature statistics, thereby reducing computational complexity while preserving fine details of small objects. A second 1×1 convolution layer then restores the dimensionality and produces the attention weights. The generated attention is normalized and fed into the gate to enable subsequent channel selection. Finally, the gating is adjusted according to the feature map hierarchy and routed into different branches. The formulas are as follows:

$$F_{\text{sel}}(P) = \begin{cases} \hat{x} \odot (\gamma_L + \gamma_G) + \hat{y} \odot (\gamma_L + \gamma_G), & P = 3, \\ \hat{x} \odot \gamma + \hat{y} \odot (1 - \gamma), & P \in \{4, 5\}. \end{cases} \tag{6}$$

$$\gamma = \sigma(\mathcal{W}_2 \delta(\mathcal{W}_1 \text{GAP}(\hat{x} + \hat{y}))), \quad [\gamma_L, \gamma_G] = \text{split}(\gamma, 2) \tag{7}$$

Among them, $\hat{x}$ and $\hat{y}$ denote the aligned features, $\odot$ represents element-wise multiplication; GAP stands for global average pooling. The symbols σ and δ represent the Sigmoid and ReLU activation functions, respectively. P denotes the number of feature levels, and γ signifies the gating weight. The operations $\mathcal{W}_1$ and $\mathcal{W}_2$ correspond to two distinct 1×1

convolutional layers, with the subscripts L and G denoting Local Attention and Global Attention, respectively.

In the case of $P = 3$, the features are decoupled into two attention branches via a Split operation. Specifically, Global Attention efficiently models global context and establishes long-range dependencies with negligible computational overhead, thereby alleviating the risk of local optima entrapment. Complementarily, Local Attention targets local discrimination by capturing edge textures and shapes to retrieve details omitted by the global component; by utilizing a diverse set of convolutions, it ensures the receptive field remains localized even after dilation rate expansion. This synergy sensitizes the model to dynamic contextual shifts and facilitates rapid, adaptive feature fusion within complex visual environments.

Conversely, for stages $P = 4$ and 5, the reduced resolution of feature maps renders additional attention branches computationally prohibitive. To address this, a gated complementary selection mechanism is introduced to enforce weight complementarity with a capped maximum value. This mutual constraint not only minimizes redundancy but also enhances feature representation capabilities. Furthermore, the mechanism enables real-time scheduling based on input relevance; in extreme scenarios, the gate automatically suppresses the contribution of specific information, demonstrating robust adaptability.

3.4. Scale Adaptive Quality Loss

In object detection tasks, especially in scenarios such as aerial image detection with significant scale variation, the design of the loss function is critical to model performance. Traditional quality-aware loss functions such as Focal Loss [20], Varifocal Loss, and GIoU improve detection accuracy to some extent by emphasizing hard samples and focusing on prediction quality. However, they generally lack explicit modeling of object scale, and this scale-insensitive property is particularly detrimental for small object detection. Specifically, small objects, due to their minute size, usually yield lower IoU values for their bounding boxes than larger objects. Under existing loss formulations, this low IoU is directly used as classification confidence and regression weight, leading to insufficient supervision for positive samples. Moreover, small objects are often embedded in large amounts of irrelevant background information, which interferes with their feature representation. If a uniform suppression strategy is applied, these already ambiguous small objects tend to be excessively suppressed together with the background, exacerbating missed and false detections.

To address these issues, we design SAQL. SAQL is a new loss function derived from Varifocal Loss. Its core idea is to introduce object scale as a dynamic modulation factor, assigning different loss weights to samples of different sizes. For small objects, SAQL enlarges the loss to compensate for insufficient supervision, while imposing stronger suppression on the surrounding background information, thereby enhancing the model's detection capability under extreme conditions.

SAQL enhances the standard binary cross-entropy (BCE) loss with two dynamic, scale-aware modulation components. For positive samples, a scale amplification factor s_i is introduced to assign higher loss weights to smaller objects. This is achieved through an intermediate normalization factor α_i based on the normalized square root of the target area a_i. The factor s_i is computed as:

$$s_i = \text{clip}(\exp(\beta \cdot (1 - \alpha_i)), s_{\min}, s_{\max}) \tag{8}$$

The hyperparameter β governs sensitivity to scale variations and algorithmic characteristics. Clipping s_i within the bounded interval $[s_{\min}, s_{\max}]$ maintains training stability while preventing numerical overflow.

An adaptive concentration parameter γ_i is incorporated for negative samples. Coupled with s_i, this parameter enhances background suppression capabilities, particularly for small-scale objects. The focusing range boundaries are established by $\gamma_{\min}$ and $\gamma_{\max}$. Specifically, for negative samples, s_i is fixed at unity, yielding $\gamma_i = \gamma_{\min}$. For positive samples, dimensionality reduction results in higher s_i values, which in turn lowers γ_i as defined by the following equation:

$$\gamma_i = \gamma_{\max} - (\gamma_{\max} - \gamma_{\min}) \cdot s_i \tag{9}$$

The final SAQL is defined as:

$$\mathcal{L}_{SAQL}(p, q) = \begin{cases} -s_i \cdot q \log(p) & \text{if positive} \\ -(1-p)^{\gamma} \log(1-p) & \text{if negative} \end{cases} \tag{10}$$

SAQL introduces scale-related dynamic weight computation only during training, explicitly models scale-aware supervisory signals without adding any inference cost, and adaptively strengthens supervision for small objects, thereby enhancing foreground discrimination in dense scenes and improving small-object recall and overall detection robustness.

3.5. Asymmetric Hierarchical Synergy

When constructing a lightweight small-object detector, existing approaches either introduce excessive modules that lead to computational redundancy, or adopt symmetric and homogeneous architectures that uniformly allocate model resources. These issues prevent a satisfactory balance between performance and real-time capability. Such egalitarian resource allocation is inconsistent with the efficient information processing mechanism of the biological visual system. The biological visual system achieves low-power, rapid perception and recognition of complex scenes through the cooperative division of labor between the fovea and peripheral vision. Inspired by bionics, we adopt a novel design paradigm, ASH, to construct our entire network. The core idea is to emulate the biological visual system by dynamically allocating limited computational resources according to task priority: employing peripheral vision processing modules for ordinary, less critical information to reduce computational overhead; and using a somewhat more computation-intensive foveal processing module for high-value information directly related to the core of small-object detection. Ultimately, the components cooperate to realize an optimal Pareto frontier. Specifically, we define the problem addressed by ASH as a multi-objective optimization problem, so that our goal becomes finding the optimal set of model configuration parameters. The problem is formulated as follows:

$$\min_{\theta \in \Theta} \mathbf{F}(\theta) = \left(\mathcal{L}_{\text{per}}(\theta), C_{\text{comp}}(\theta) \right) \tag{11}$$

Let θ parameterize the model architecture. We aim to optimize a multi-objective function $\mathbf{F}(\theta)$ comprising a performance loss $\mathcal{L}_{\text{var}}(\theta)$, which measures detection accuracy, and a computational cost term $C_{\text{comp}}(\theta)$, which quantifies model complexity. ASH method introduces a novel formulation for $C_{\text{comp}}(\theta)$. This asymmetric, hierarchical design effectively shifts the Pareto front towards a superior trade-off region, yielding a more optimal Pareto frontier.

At the outset of our design, we mapped the biomimetic division of labor between the fovea and peripheral visual field onto the network hierarchy and the allocation of computational resources. Instead of adopting the computationally heavy, aggressively downsampling ResNet used in the original RT-DETR, we selected LSNet [46]. LSNet itself is a biomimetic architecture for local–global collaborative processing, providing an efficient

foundational partitioning framework at the early stage of feature extraction. Building on this, we feed P5 directly into AIFI, and selectively connect two LHCGF modules via channel gating specifically for peripheral visual field processing, avoiding unnecessary computational expenditure on low-resolution representations. The foveal processing module further emulates the biological visual system to achieve fine-grained information refinement. High-value signals are handled by the DLEP module, which aggregates the downsampled fine details from the SPDConv in P2, the high-resolution local edges from P3, and the contextual expressions from P4, preserving fine-grained spatial information to the greatest extent.

However, an efficient feature fusion module alone cannot resolve the inherent semantic insufficiency of shallow features. Therefore, we employ the HEFE module to perform cross-layer interaction internally, injecting global context back into high-resolution features to compensate for semantic deficiencies and provide precise semantic guidance without increasing computational burden. All these features are fused along a global path dominated by LHCGF, using lower computational overhead in place of conventional fusion, ensuring that local–global information is not lost during integration. The foveal processing module enhances the treatment of salient information by ingesting high-frequency details from P2 and local structures from P3, while absorbing semantics fed back from P4, and allocating greater computational resources accordingly.

Finally, small objects, due to their limited size and ambiguous features, are more prone to low-IoU matches under DETR queries and are thus ignored by the loss function. To enable the network to fully exploit and learn from these feature signals during training, we introduce SAQL as the supervisory loss; SAQL assigns higher weights to low-quality matches, producing a stronger supervision signal. This directs the model to focus on small-object samples and improves the detection accuracy for small targets.

4. ExperimentsandResultsAnalysis

4.1. Dataset

In this paper, we employ the VisDrone2019 dataset for ablation studies, comparative experiments, and visualization analysis. The VisDrone2019 dataset, constructed by the AISKYEYE team of the Machine Learning and Data Mining Laboratory at Tianjin University, provides detailed annotations for object detection. It supplies more than 2.6 million bounding box annotations across 14 different cities and diverse geographic environments (from urban centers to rural suburbs), covering various scene types (e.g., intersections, parking lots), weather conditions (e.g., sunny, overcast), and illumination conditions (e.g., daytime, nighttime), thereby enhancing its utility.

The dataset contains 10,209 static high-resolution images, divided into three subsets for object detection: 6,471 images for training, 548 images for validation, and 1,610 images for testing. It comprises 10 categories: pedestrian, person, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motorcycle. Notably, more than 75% of annotated objects occupy less than 0.1% of the image area, underscoring the central challenge of small object detection in this dataset.

A scatter plot of object size distribution is shown in Figure 5, and a histogram of the number of objects per image is presented in Figure 6. Figure 5 demonstrates that small objects dominate the VisDrone2019 dataset, while large objects are rare; nevertheless, the wide variation in object scales poses significant challenges for detection models. Figure 6 illustrates the high object density within individual images, with most images containing dozens of targets, which imposes stricter requirements on accurate detection and robustness under complex conditions.

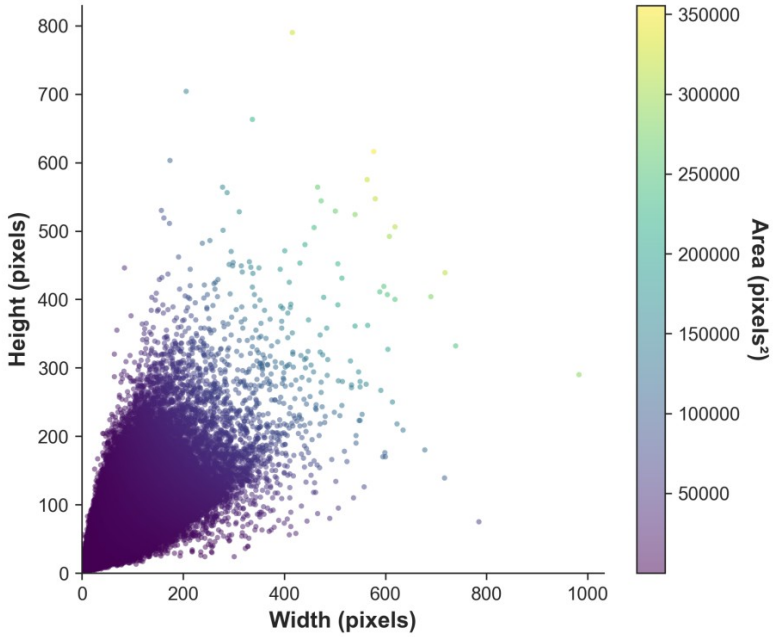


Figure 5. Scatter plot of object size distribution on the VisDrone2019 dataset.

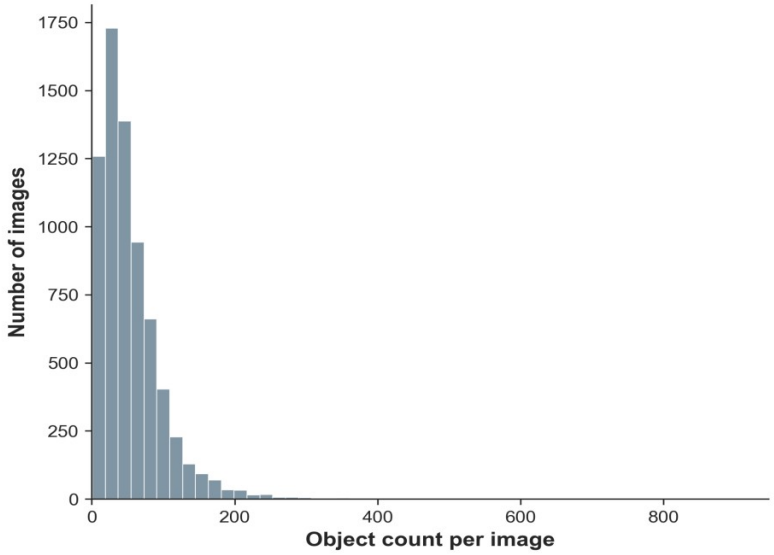


Figure 6. Histogram of object count per image in the VisDrone2019 dataset.

4.2. Experimental environment

Experiments were conducted on a system with an Intel Core i9-14900HX processor and an NVIDIA GeForce RTX 5070 Ti 12 GB GPU. The training environment was configured with Python 3.11, CUDA 12.8, PyTorch 2.7, and Torchvision 0.22.1. During training, we used the AdamW optimizer with an initial learning rate of 0.0001, a warm-up of 5 epochs, and a weight decay of 0.0005. All models were trained from scratch for 150 epochs without pretrained weights. The input image size was 640×640 and the batch size was set to 4. We did not employ early stopping, exponential moving average (EMA), or layer-freezing strategies.

Table 1. Experimental Details.

Type	Value
Python	3.11
CUDA	12.8
Pytorch	2.7.1
Torchvision	0.22.1
Epoch	150
Patience	0
Initial Learning rate	0.0001
Image size	640
Optimizer	AdamW
Weight Decay	0.0001
Mosaic	0.5
MixUp	0.1
Fliplr	0.5

4.3. Evaluation Metrics

Our core evaluation metrics follow standard object detection practice, including mAP50, mAP50:95, APs, APm, APl, GFLOPs, and FPS, to comprehensively assess the detection performance of the model. mAP is the mean of AP over all categories; unlike AP (per class), mAP effectively measures the model's detection performance across the entire category set, defined as:

$$\text{mAP} = \frac{1}{C}\sum_{i=1}^{C}\text{AP}_i \tag{12}$$

where C denotes the number of categories and AP_i is the Average Precision of the i-th category. mAP50–95 is the average AP over multiple IoU thresholds from 0.50 to 0.95 (step size 0.05), providing a more stringent comparison.

Since mAP alone cannot adequately reflect detection capability across object scales, we further employ APs, APm, and APl to evaluate performance at different scales and localization granularities: APs measures performance on small objects (area $< 32 \times 32$ pixels), which is crucial for assessing our small-object detection improvements; APm measures performance on medium objects ($32 \times 32 \leq$ area $< 96 \times 96$ pixels); APl measures performance on large objects (area $\geq 96 \times 96$ pixels).

In addition, we report GFLOPs to quantify computational complexity (in billions of floating-point operations) and frames per second (FPS) to evaluate real-time capability. These two metrics indicate whether the model reduces deployment barriers while achieving real-time detection requirements. During evaluation, we uniformly set the batch size to 1 to ensure consistency of experimental results.

4.4. Ablation Experiments

To thoroughly validate the effectiveness of the proposed improvement modules, RT-DETR was selected as the baseline model, with ResNet18 as its backbone. The base model is denoted as Model A. Ablation studies were conducted on LSNet (B), DLEP (C), LHCGF (D), and SAQL (E) to assess the impact of each configuration on model performance. To ensure result consistency, identical hyperparameter settings were applied across all experiments.

First, the standard ResNet18 backbone in RT-DETR was replaced with LSNet; then DLEP was added to extract multi-level information and perform cross-scale feature interaction; next, LHCGF was used to process channel information and suppress redundant features; finally, the loss function was replaced with SAQL. Each improvement module was evaluated sequentially through ablation experiments. The results are shown in Table 2.

Table 2. Ablation study comparing various metrics on the dataset.

Method	mAP50:95 (%)	mAP50 (%)	APs (%)	APm (%)	APl (%)	FLOPs (G)	FPS
A (base)	18.5	33.3	10.7	29.1	35.6	58.6	32.5
A+B	18.7	32.1	10.2	28.4	39.7	42.6	41.1
A+B+C	19.6	34.4	11.1	29.6	39.8	40.0	43.3
A+B+C+D	20.8	35.8	12.0	30.7	39.4	41.3	40.8
A+B+C+D+E (ours)	21.3	37.4	13.6	32.0	42.7	41.3	40.9

The ablation experiments provide a detailed analysis of how different model configurations affect object detection performance metrics. Using RT-DETR-R18 as the baseline (Model A), the model has 19.88 MB of parameters, a computational complexity of 58.6 GFLOPs, an mAP50 of 33.3%, and runs at 32.5 FPS. The experimental results show that replacing the ResNet18 backbone with LSNet (Model A+B) reduces computational complexity by 16.0 GFLOPs, increases FPS by 8.6, and improves APl by 4.1%, while mAP50 and APs decrease by 1.2% and 0.5%, respectively. This indicates that the model achieves lightweight design while effectively maintaining real-time performance and preserving multi-scale feature extraction capability.

From Experiment 3 (Model A+B+C), it can be observed that using DLEP for cross-scale feature fusion significantly improves model accuracy, especially for small and medium objects, while further reducing computational complexity. On top of LSNet, complexity is reduced by 2.6 GFLOPs. mAP50:95 and mAP50 increase by 1.1% and 2.3%, respectively; APs and APm increase by 0.9% and 1.2%; and FPS rises to 43.3. DLEP not only effectively fuses features at different scales but also enhances real-time detection performance.

Subsequent Experiment 5 (Model A+B+C+D+E) shows that, at a minimal cost of 1.3 GFLOPs and 0.64 M additional parameters, mAP50:95 and mAP50 further increase by 0.9% and 1.4%, while APs and APm improve by 0.9% and 1.1%. Experiments 4 and 5 demonstrate that DLEP and LHCGF effectively fuse and select the multi-scale features provided by LSNet, improving small and medium object detection by 1.6% and 1.3% over the baseline, while further reducing computational complexity and easing deployment constraints.

Experiment 5 also shows that replacing the loss function with SAQL applies differentiated supervision signals to targets of different scales. mAP50:95 and mAP50 increase by 1.2% and 1.4%, and APs, APm, and APl improve by 1.6%, 1.3%, and 3.3%, respectively.

Considering all improvements, the overall network exhibits only a slight increase in parameter count compared with the baseline, while computational complexity decreases by 29.52%, lowering the computational requirements for edge devices and making deployment easier. mAP50 increases by 4.1%, and detection performance across scales—particularly for small objects—improves markedly, with APs rising by 3.1%, meeting the demands of small aerial object detection. The frame rate increases by 8.4 FPS, satisfying real-time detection requirements. These changes directly demonstrate the effectiveness of the proposed architecture for small aerial object detection tasks. The method reduces computational complexity while accelerating convergence and improving detection accuracy.

4.5. Comparative Study

After validating the effectiveness of the proposed method, we conducted comparative experiments between HSP-DETR and other mainstream advanced algorithms, including Faster R-CNN [18], Cascade R-CNN [19], ATSS-R50 [47], TOOD-R50 [48], GFL [49], RetinaNet-R50-FPN [20], the YOLO series and its derivatives on the VisDrone2019 dataset.

The experiments considered the influence of model parameter count, computational complexity, mAP, mAP50–95, and AP at different scales. All experiments were performed in the same environment as the ablation studies.

As shown in Table 3, with our improvements HSP-DETR achieves outstanding performance, reaching an AP of 21.3%, which represents the current best detection accuracy and surpasses other mainstream methods. Even compared with the efficiency-oriented YOLO series, HSP-DETR attains a favorable overall balance, exhibiting clear improvements in mAP50:95 and mAP50 and exceeding the highest-performing YOLO10m and YOLO11m variants, fully reflecting the trade-off between accuracy and efficiency. This demonstrates that when confronting the common multi-scale and multi-orientation targets in aerial imagery, HSP-DETR's global contextual modeling capability and end-to-end detection pipeline offer prominent advantages, avoiding the excessive post-processing that can lead to occlusion issues for dense targets and inflated computational cost.

Table 3. Comparison of HSP-DETR with state-of-the-art methods on the VisDrone2019 dataset.

Methods	mAP50:95 (%)	mAP50 (%)	APs (%)	APm (%)	APl (%)	FLOPs (G)	Params (M)
Faster-RCNN [18]	19.4	39.2	9.5	30.9	42.9	208	41.39
Cascade-RCNN [19]	19.7	32.6	9.9	30.9	40.6	236	69.29
ATSS-R50 [47]	20.4	33.8	10.1	31.7	48.5	110	38.91
TOOD-R50 [48]	20.4	33.9	10.2	31.7	40.3	199	32.04
GFL [49]	19.3	32.1	9.4	30.0	40.9	206	32.28
RetinaNet-R50-FPN [20]	16.4	27.6	6.3	27.4	42.7	210	36.52
YOLOv8m [9]	19.1	33.2	9.0	29.4	41.7	78.7	25.85
YOLOv10m [11]	19.5	34.5	9.7	30.0	41.4	58.6	15.32
YOLOv11m [12]	20.3	35.3	9.8	31.2	41.3	67.7	20.04
YOLOv12m [13]	19.2	33.6	9.4	29.8	38.6	67.2	20.11
YOLOv13s [14]	16.7	29.7	7.7	25.8	31.7	6.2	9.0
FBRT-YOLO-M [15]	19.6	34.4	9.4	30.9	42.1	58.7	7.36
RT-DETR [27]	18.5	33.3	10.7	29.1	35.6	58.6	19.88
HSP-DETR	21.3	37.4	13.6	32.1	42.7	41.3	20.21

Table 3 also shows that, through our enhancements, HSP-DETR delivers excellent performance in small object detection, significantly outperforming other mature models. It achieves an APs of 13.6%, which is 2.9% higher than the baseline RT-DETR, validating the effectiveness of HSP-DETR in cross-scale information extraction and fusion. Its advantages over the established CNN detectors ATSS-R50 and TOOD-R50 are even more pronounced, with gains of 3.2% and 3.7%, respectively. It remains competitive against the currently popular YOLO series, maintaining a leading edge in small object detection, exceeding the best YOLO11m APs by 3.6%. Since targets in aerial images predominantly appear at small sizes with large scale variation and uneven spatial distribution, improvement in this metric is of substantial practical significance for real-world aerial object detection. This gain is attributed to HSP-DETR's refined multi-scale feature fusion and global–local contextual modeling mechanism, which effectively enhances its ability to detect low-resolution, small-sized objects. The advantages of HSP-DETR are not limited to small object detection; it also remains competitive for medium-scale and large-scale objects. As shown in Table 3, although its performance on large objects (APl) still lags behind certain models, its computational complexity is only approximately one-fifth that of GFL and TOOD-R50—which perform well on large object detection—while achieving higher overall accuracy (mAP50:95) and superior small-object performance (APs). This demonstrates that HSP-DETR avoids the common defect of favoring large objects while neglecting small ones. With lower computational cost, HSP-DETR attains stronger comprehensive performance,

and the performance gains are achieved through enhanced feature representation and contextual interaction.

The above experiments reveal the limitations faced by existing mainstream methods in real-time monitoring of aerial imagery, while highlighting the strong performance of HSP-DETR. The proposed model exhibits stable behavior in real-world scenarios characterized by dense targets and drastic scale variations. It strengthens small object detection capability while preserving inference efficiency, providing an effective solution for practical deployment.

4.6. Extended experiments

To assess the model's generalization capability and robustness, we further conducted generalization experiments on the HIT-UAV dataset. The HIT-UAV [41] dataset, provided by Harbin Institute of Technology, is a high-altitude infrared thermal imaging dataset for object detection on UAV platforms. It contains 2,898 infrared thermal images with a resolution of 640×512, and a total of 24,899 annotated instances. The scenes cover diverse real-world environments such as campuses, parking lots, roads, and playgrounds, and include both daytime and nighttime imagery. The dataset features densely distributed small-sized objects, categorized into five classes: person, bicycle, car, other vehicles, and DontCare, making it suitable for effectively validating the model's generalization performance on small object detection.

Table 4. Comparison results on HIT-UAV dataset.

Method	Person (%)	Car (%)	Bicycle (%)	Other Vehicle (%)	Dont Care (%)	mAP50 (%)	FLOPs (G)
YOLOv5m [7]	91.6	95.7	89.9	71.2	72.2	84.1	48.0
YOLOv6m [8]	90.3	97.5	90.8	54.1	48.4	76.2	85.6
YOLOv8m [9]	91.9	96.8	91.4	66.9	63.3	82.0	78.7
YOLOv9m [10]	92.0	96.6	92.9	66.6	67.3	83.1	76.0
YOLOv10m [11]	91.1	96.7	90.9	67.3	50.1	79.2	58.6
YOLOv11m [12]	92.5	96.9	91.2	66.3	69.2	83.2	67.7
YOLOv12m [13]	93.2	97.4	92.7	67.6	62.4	82.7	67.2
MHAF-YOLO [16]	89.8	95.8	89.8	64.9	47.4	77.5	67.6
AMFEF-DETR [50]	94.1	96.1	91.0	58.5	67.5	81.5	142
RT-DETR [27]	93.6	97.4	90.1	59.6	53.2	78.8	58.6
Freq-DETR [51]	93.8	99.1	90.1	77.1	45.2	81.6	80
HSP-DETR	93.3	96.4	89.1	74.5	63.4	83.4	41.3

As shown in Table 4, we compare the proposed HSP-DETR model with current mainstream advanced models YOLOv5, YOLOv6, YOLOv8, YOLOv9, YOLOv10, YOLOv11, YOLOv12, MHAF-YOLO, and Transformer-based improved models AMFEF-DETR and Freq-DETR, to validate that HSP-DETR still delivers superior performance on another dataset. HSP-DETR achieves a significant improvement over many advanced small object detection models on the HIT-UAV dataset.

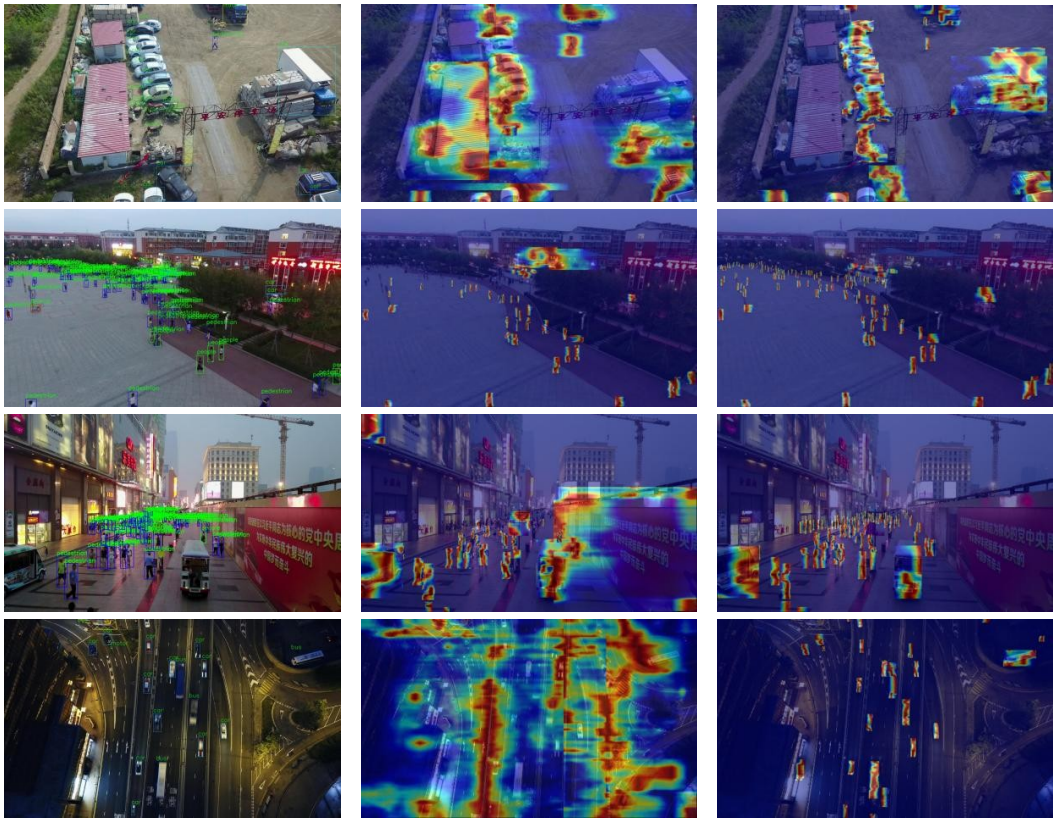
Our model improves mAP50 by 1.4% and 4.2% relative to the current mainstream YOLOv8m and YOLOv10m models, respectively, and surpasses the baseline RT-DETR by 4.6%. Even compared with the state-of-the-art YOLOv11m and YOLOv12m, HSP-DETR exceeds them by 0.2% and 0.7% with lower GFLOPs. However, our model is weaker than YOLOv5m in overall accuracy; as seen in the table, this gap mainly arises because HSP-DETR underperforms YOLOv5m by 8.8% on the DontCare category. Upon further analysis, we attribute this to the blurred thermal signatures of DontCare targets in infrared images, which require stronger edge extraction capability. HSP-DETR is constrained by the

feature extraction capacity of LSNet and thus cannot further capture faint edge features in infrared imagery. In contrast, for deterministic small objects such as pedestrians, cars, and bicycles, our model maintains excellent detection performance.

Collectively, these results indicate that our approach not only performs well on conventional UAV imagery, but also sustains superior performance in more challenging infrared scenarios. The generalization experiment further demonstrates strong adaptability in infrared scenes, highlighting the potential of the method for widespread deployment in complex environments.

4.7. Visualization Analysis

Although experiments have already demonstrated the effectiveness of the model, to further highlight its key characteristics and enhance interpretability, this section conducts an in-depth analysis using visualization techniques. We employ the LayerCAM method to generate heatmaps of the model during detection, enabling us to assess whether, as anticipated, it produces strong detection responses for densely distributed small objects, and to examine its behavior under different environmental conditions. Four representative images under distinct scenarios—daytime, nighttime, dense pedestrian flow, and vehicular traffic scenes—are selected for this analysis.



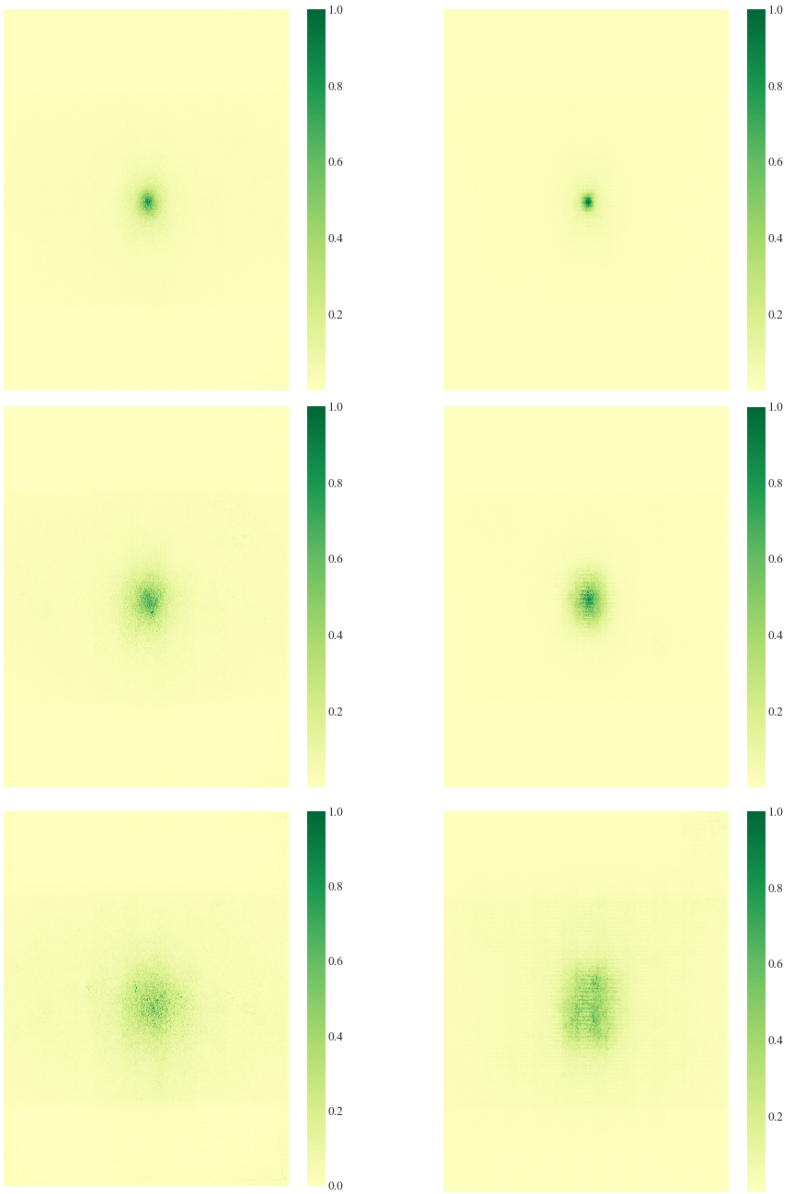
(a) GroundTruth (b) RT-DETR (c) HSP-DETR

Figure 7. Heatmap analysis across different scenes.

As shown clearly in Figure 7, RT-DETR is easily distracted by other objects, whether by the red canopy in the parking lot or the display boards on the pedestrian street, and performs poorly under such circumstances. In contrast, HSP-DETR, whether for detecting medium-sized cars, dense pedestrian flows in open areas, or even dense crowds in the challenging pedestrian street scenario, is able to focus on small objects while also recognizing occluded ones—capabilities that RT-DETR does not possess. It is worth noting that in the last image we use late-night aerial imagery to demonstrate the limit of small-object

detection. In nighttime high-angle UAV scenes, the elevated viewpoint makes targets even smaller than usual. Moreover, under nighttime lighting conditions targets may be lost, posing a significant challenge to our detection model.

By analyzing the Ground Truth and the performance of RT-DETR-18 in the last four images, we observe that the attention distribution of RT-DETR is relatively diffuse, covering not only the target vehicles but also responding strongly to background regions such as lane markings and ground textures. This suggests that in small object detection it may overly rely on scene context while insufficiently modeling the local features of the objects themselves. In contrast, HSP-DETR is still able to perform detection in nighttime scenes, even recognizing the bus in the upper right and the motorcycle on the left that is an even smaller target. These results validate that the proposed HSP-DETR model, through its global context interaction and cross-scale feature fusion mechanisms, enhances the robustness of object detection under extreme conditions.



(a) RT-DETR (b) HSP-DETR

Figure 8. Effective receptive field comparison of the model at different stages.

To further investigate model performance, we plot the effective receptive field maps of RT-DETR and HSP-DETR at P3, P4, and P5, as shown in Fig. 8. It can be observed

that at every stage the receptive fields of HSP-DETR are more concentrated than those of RT-DETR, and they exhibit higher response weights in the central region. In aerial imagery, which requires both broad coverage and the capture of local details, maintaining high and focused response weights at higher layers enables better detection of small objects. Although RT-DETR presents a larger receptive field at higher stages, its response weights are generally diffuse and low, which is unfavorable for small-object detection. Small objects occupy few pixels and are commonly accompanied by occlusion and blurred boundaries, making them highly susceptible to background interference; hence, they impose stricter requirements on the model's receptive field. The narrow and sharply peaked effective receptive field of HSP-DETR at the P3 stage helps the model precisely focus on the small object itself in high-resolution images, enhancing local feature sensitivity, reducing background noise interference, and preventing missed and false detections within small-object regions, thereby improving accuracy in small-object detection. In subsequent stages, the receptive field continues to demonstrate high-quality characteristics, progressively expanding its coverage while retaining concentrated response weights. Under the synergistic effect of the global context interaction mechanism and cross-scale feature fusion, the model enlarges the captured contextual information while maintaining high precision.

Figure 9 presents a comparison of RT-DETR and HSP-DETR for object detection from a UAV perspective. Panel (a) is the original input image, panel (b) shows the RT-DETR detection results, and panel (c) shows the HSP-DETR detection results. The image set comprises three complex scenarios selected from VisDrone2019: a daytime mixed-traffic intersection scene, a bird's-eye view of a traffic crossroads, and a nighttime low-light crowd-dense scene. Figure 9 presents a comparison of RT-DETR and HSP-DETR for object detection from a UAV perspective. Panel (a) is the original input image, panel (b) shows the RT-DETR detection results, and panel (c) shows the HSP-DETR detection results. The image set comprises three complex scenarios selected from VisDrone2019: a daytime mixed-traffic intersection scene, a bird's-eye view of a traffic crossroads, and a nighttime low-light crowd-dense scene.

(a) Original image (b) RT-DETR (c) HSP-DETR

Figure 9. Comparison of detection results.

In the first daytime parking lot image, confronted with a complex environment containing clustered multiple targets, HSP-DETR (Fig. 9(c)) not only achieves higher accuracy

than RT-DETR (Fig. 9(b)), but also produces stable high-confidence bounding boxes when targets are occluded. In densely populated target regions (such as the junction between the parking lot and the main arterial road), the baseline RT-DETR model exhibits obvious missed detections, whereas HSP-DETR shows stronger bounding-box localization precision and category discrimination, effectively recognizing overlapping vehicles and small vehicles.

In the second bird's-eye view of a traffic crossroads, large scale variations and complex texture interference are present. RT-DETR loses numerous targets in distant traffic flow regions and, when faced with closely overlapping targets such as bicycle-riding pedestrians, can recognize only one of them. Leveraging refined cross-scale feature processing, HSP-DETR not only maintains detection capability in distant traffic areas, but also accurately identifies targets near building edges. Precise bounding-box localization effectively prevents missed and false detections when detection boxes are nearly overlapping, demonstrating the model's robustness under extreme scale variation and complex backgrounds.

The nighttime low-light crowd-dense scene poses significant challenges, primarily due to insufficient illumination, building glare, densely packed crowds, and accompanying shadows. RT-DETR can recognize only a small portion of targets in such extreme conditions, with a large number of missed detections. Our model, through multi-scale feature extraction, cross-scale feature fusion, and the strong signal supervision for small objects provided by SAQL, not only detects targets under shadow but also exhibits stronger resistance to glare interference, producing bounding boxes tightly adhering to object contours. It shows greater adaptability and stability under illumination changes in complex nighttime scenes.

In summary, comparative analysis of detection outputs shows that HSP-DETR outperforms RT-DETR across diverse environments (daytime, complex intersections, nighttime low-light conditions, dense targets, and occlusions) in terms of accuracy, missed detections, and false detections. This advantage stems from HSP-DETR's fine-grained feature extraction and fusion, which enhance the model's capability for detail capture and context interaction, yielding superior robustness and effectiveness in UAV detection scenarios.

5. Conclusions and Future Work

This paper presents HSP-DETR, a lightweight small object detection model designed for real-time aerial imagery, developed through enhancements to the RT-DETR framework. HSP-DETR is a real-time Transformer-based detection architecture that incorporates an asymmetric hierarchical collaboration mechanism. Drawing inspiration from biomimetic principles, the model employs asymmetric hierarchical cooperation to achieve a lightweight design without compromising efficiency, thereby reaching a favorable Pareto optimum. By jointly optimizing cross-scale feature extraction and global contextual interaction modeling, our approach not only improves the detection performance for small objects but also substantially reduces computational overhead. Experimental results demonstrate that, compared to the original RT-DETR, HSP-DETR achieves improvements of 2.8% in mAP50:95 and 4.1% in mAP50, with a 2.9% increase in APs. Furthermore, the computational complexity is reduced by 29.52% relative to the baseline, indicating enhanced robustness and adaptability for aerial image detection, particularly in challenging scenarios. Nevertheless, the current work has certain limitations. First, although the adoption of LSNet as the backbone improves efficiency, visualizations of the effective receptive field suggest that its performance may plateau on extremely complex and rare hard samples. Second, the architectures and hyperparameters of modules such as PST and LHCGF remain static after training; such a fixed design cannot adapt flexibly to all input images. Finally, although the model reduces computational complexity through careful redesign of RT-DETR, its deploy-

ment demands on edge devices remain higher than those of the YOLO family, indicating potential for further optimization. Future research will focus on several directions. First, we will redesign the LSNet building blocks to enhance performance under highly complex conditions. Second, we plan to develop an adaptive inference mechanism that dynamically adjusts computational pathways based on input image content, enabling more granular allocation of computational resources while preserving accuracy and accelerating inference. Additionally, techniques such as knowledge distillation, quantization, and pruning will be explored to further reduce computational cost and parameter count, thereby easing deployment requirements on edge devices.

Author Contributions: Conceptualization, Qinyu Liu, Min Xu and Lei Liao; Methodology, Qinyu Liu; Software, Qinyu Liu; Validation, Min Xu and Lei Liao; Formal analysis, Qinyu Liu; Investigation, Qinyu Liu and Min Xu; Resources, Lei Liao; Data curation, Qinyu Liu; Writing—original draft preparation, Qinyu Liu; Writing—review and editing, Min Xu and Lei Liao; Visualization, Qinyu Liu; Supervision, Lei Liao; Project administration, Lei Liao; Funding acquisition, Lei Liao. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed in this study are publicly available. The VisDrone2019 dataset [40] can be accessed at the official website: http://aiskyeye.com/. The HIT-UAV dataset [41] is available through the GitHub repository: https://github.com/suojiashun/HIT-UAV.

Conflicts of Interest: The authors declare that there are no known competing financial interests or personal relationships that could have influenced the research presented in this study.

References

1. Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)* (Vol. 1, p. I). https://doi.org/10.1109/CVPR.2001.990517
2. Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., & Smeulders, A. W. M. (2013). Selective search for object recognition. *International Journal of Computer Vision*, *104*(2), 154–171. https://doi.org/10.1007/s11263-013-0620-5
3. Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 886–893). https://doi.org/10.1109/CVPR.2005.177
4. Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, *7*(1), 11–32. https://doi.org/10.1007/BF00130487
5. Ojala, T., Pietikäinen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971–987. https://doi.org/10.1109/TPAMI.2002.1017623
6. Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008)* (pp. 1–8). https://doi.org/10.1109/CVPR.2008.4587597
7. Jocher, G., & others. (2020). *YOLOv5* [Computer software]. GitHub. https://github.com/ultralytics/yolov5
8. Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., Li, Y., Zhang, B., Liang, Y., Zhou, L., Xu, X., Chu, X., Wei, X., & Wei, X. (2022). YOLOv6: A single-stage object detection framework for industrial applications. *arXiv preprint*. arXiv:2209.02976. https://arxiv.org/abs/2209.02976
9. Jocher, G., & others. (2023). *Ultralytics YOLOv8* [Computer software]. GitHub. https://github.com/ultralytics/ultralytics
10. Wang, C.-Y., Yeh, I.-H., & Liao, H. (2024). YOLOv9: Learning what you want to learn using programmable gradient information. *arXiv preprint*. arXiv:2402.13616. https://arxiv.org/abs/2402.13616
11. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., & Ding, G. (2024). YOLOv10: Real-time end-to-end object detection. *arXiv preprint*. arXiv:2405.14458. https://arxiv.org/abs/2405.14458
12. Ultralytics. (2024). *YOLOv11* [Computer software]. GitHub. https://github.com/ultralytics/ultralytics

13. Tian, Y., Ye, Q., & Doermann, D. S. (2025). YOLOv12: Attention-centric real-time object detectors. *arXiv preprint*. arXiv:2502.12524. https://arxiv.org/abs/2502.12524
14. Lei, M., Li, S., Wu, Y., Hu, H., Zhou, Y., Zheng, X., Ding, G., Du, S., Wu, Z., & Gao, Y. (2025). YOLOv13: Real-time object detection with hypergraph-enhanced adaptive visual perception. *arXiv preprint*. arXiv:2506.17733. https://arxiv.org/abs/2506.17733
15. Xiao, Y., Xu, T., Xin, Y., & Li, J. (2025). FBRT-YOLO: Faster and better for real-time aerial image detection. *arXiv preprint*. arXiv:2504.20670. https://arxiv.org/abs/2504.20670
16. Yang, Z., Guan, Q., Yu, Z., Xu, X., Long, H., Lian, S., Hu, H., & Tang, Y. (2025). MHAF-YOLO: Multi-branch heterogeneous auxiliary fusion YOLO for accurate object detection. *arXiv preprint*. arXiv:2502.04656. https://arxiv.org/abs/2502.04656
17. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single shot MultiBox detector. In *Computer Vision – ECCV 2016* (Vol. 9905, pp. 21–37). Springer. https://doi.org/10.1007/978-3-319-46448-0_2
18. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28* (pp. 91–99). https://doi.org/10.48550/arXiv.1506.01497
19. Cai, Z., & Vasconcelos, N. (2018). Cascade R-CNN: Delving into high quality object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6154–6162). https://doi.org/10.1109/CVPR.2018.00644
20. Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2020). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(2), 318–327. https://doi.org/10.1109/TPAMI.2018.2858826
21. Zhu, X., Lyu, S., Wang, X., & Zhao, Q. (2021). TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (pp. 2778–2788). https://doi.org/10.48550/arXiv.2108.11539
22. Chen, J., Kao, S., He, H., Zhuo, W., Wen, S., Lee, C., & Chan, S. (2023). Run, don't walk: Chasing higher FLOPS for faster neural networks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12021–12031). https://doi.org/10.1109/CVPR52729.2023.01157
23. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. *arXiv preprint*. arXiv:2005.12872. https://arxiv.org/abs/2005.12872
24. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., & Zhang, L. (2022). DAB-DETR: Dynamic anchor boxes are better queries for DETR. *arXiv preprint*. arXiv:2201.12329. https://arxiv.org/abs/2201.12329
25. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L. M., & Zhang, L. (2024). DN-DETR: Accelerate DETR training by introducing query denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(4), 2239–2251. https://doi.org/10.1109/TPAMI.2023.3335410
26. Wang, T., Yuan, L., Chen, Y., Feng, J., & Yan, S. (2021). PnP-DETR: Towards efficient visual analysis with transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 4641–4650). https://doi.org/10.1109/ICCV48922.2021.00462
27. Lv, W., Xu, S., Zhao, Y., Wang, G., Wei, J., Cui, C., Du, Y., Dang, Q., & Liu, Y. (2024). DETRs beat YOLOs on real-time object detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 16965–16974). https://doi.org/10.1109/CVPR52733.2024.01605
28. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv preprint*. arXiv:2010.04159. https://arxiv.org/abs/2010.04159
29. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., & Shum, H.-Y. (2022). DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint*. arXiv:2203.03605. https://arxiv.org/abs/2203.03605
30. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 936–944). https://doi.org/10.1109/CVPR.2017.106
31. Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 8759–8768). https://doi.org/10.1109/CVPR.2018.00913
32. Liu, S., Huang, D., & Wang, Y. (2019). Learning spatial fusion for single-shot object detection. *arXiv preprint*. arXiv:1911.09516. https://arxiv.org/abs/1911.09516
33. Qiao, S., Chen, L.-C., & Yuille, A. L. (2020). DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10208–10219). https://doi.org/10.48550/arXiv.2006.02334
34. Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., & Zhang, L. (2021). Dynamic head: Unifying object detection heads with attentions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7369–7378). https://doi.org/10.1109/CVPR46437.2021.00729
35. Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., Wang, X., & Qiao, Y. (2023). InternImage: Exploring large-scale vision foundation models with deformable convolutions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 14408–14419). https://doi.org/10.1109/CVPR52729.2023.01385
36. Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7132–7141). https://doi.org/10.1109/CVPR.2018.00745

37. Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Computer Vision – ECCV 2018* (Vol. 11211, pp. 3–19). Springer. https://doi.org/10.1007/978-3-030-01234-2_1
38. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11531–11539). https://doi.org/10.1109/CVPR42600.2020.01155
39. Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 13708–13717). https://doi.org/10.1109/CVPR46437.2021.01350
40. Du, D., & others. (2019). VisDrone-DET2019: The vision meets drone object detection in image challenge results. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (pp. 213–226). https://doi.org/10.1109/ICCVW.2019.00030
41. Suo, J., Wang, T.-M., Zhang, X., Chen, H.-m., Zhou, W., & Shi, W. (2022). HIT-UAV: A high-altitude infrared thermal dataset for unmanned aerial vehicle-based object detection. *Scientific Data, 10*(1), Article 227. https://doi.org/10.1038/s41597-023-02066-6
42. Tan, M., Pang, R., & Le, Q. V. (2020). EfficientDet: Scalable and efficient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10778–10787). https://doi.org/10.1109/CVPR42600.2020.01079
43. Xie, S., & Tu, Z. (2017). Holistically-nested edge detection. *International Journal of Computer Vision, 125*(1-3), 3–18. https://doi.org/10.1007/s11263-017-1004-z
44. Liu, Y., Cheng, M.-M., Hu, X., Wang, K., & Bai, X. (2019). Richer convolutional features for edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(8), 1939–1946. https://doi.org/10.1109/TPAMI.2018.2878849
45. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90
46. Wang, A., Chen, H., Lin, Z., Han, J., & Ding, G. (2025). LSNet: See large, focus small. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9718–9729). https://arxiv.org/abs/2503.23135
47. Zhang, S., Chi, C., Yao, Y., Lei, Z., & Li, S. Z. (2020). Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9756–9765). https://doi.org/10.1109/CVPR42600.2020.00978
48. Feng, C., Zhong, Y., Gao, Y., Scott, M. R., & Huang, W. (2021). TOOD: Task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 3490–3499). https://doi.org/10.1109/ICCV48922.2021.00349
49. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., & Yang, J. (2020). Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *arXiv preprint*. arXiv:2006.04388. https://arxiv.org/abs/2006.04388
50. Wang, S., Jiang, H., Yang, J., Ma, X., & Chen, J. (2024). AMFEF-DETR: An end-to-end adaptive multi-scale feature extraction and fusion object detection network based on UAV aerial images. *Drones, 8*(10), Article 523. https://doi.org/10.3390/drones8100523
51. Chen, J., Liu, N., Sun, H., & Wang, Y. (2026). Freq-DETR: Frequency-aware transformer for real-time small object detection in unmanned aerial vehicle imagery. *Expert Systems with Applications, 298*, Article 129710. https://doi.org/10.1016/j.eswa.2025.129710