Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 5, No. 3, 2025

DIF-DETR: Dynamic Interactive Fusion Transformer with Adaptive Feature Enhancement for Efficient Aerial Small Object Detection

Jing Wang [1], Hejiang Li [1], Caihong Huangfu [2, *]

[1] The School of Software, Henan Polytechnic University, Jiaozuo 454003, China

[2] The School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo 454003, China

[*] **Corresponding author**: Caihong Huangfu (Email: hfcaihong@hpu.edu.cn)

Abstract: In recent years, object detection models based on Transformers have demonstrated outstanding performance in general scenarios due to their powerful global feature modeling capabilities. However, when directly applied to aerial image detection tasks, their performance often falls short of expectations. The root cause lies in the nature of aerial imagery, which typically contains numerous small objects. These objects occupy an extremely low proportion of pixels, resulting in weak feature representation. They are also susceptible to factors such as complex background noise and mutual interference from densely distributed targets, making it difficult for Transformer models to effectively capture and distinguish small object features. To address these challenges, this paper proposes an enhanced Transformer architecture for aerial small object detection: Dynamic Interactive Fusion DETR (DIF-DETR). Its core innovations comprise two aspects: First, introducing the DIENet backbone feature extraction network embedded with DIEBlocks. These DIEBlocks serve as feature enhancement units within the backbone network, leveraging dynamic Inception multi-branch deep convolutions and adaptive weight allocation mechanisms to efficiently capture multi-scale, long-range contextual information. Second, it introduces Context-Aware Bidirectional Fusion (CABF), which enables adaptive complementary fusion of high-level semantic features and low-level detail features within the FPN-PAN architecture of the neck network, effectively mitigating the issue of small target features being obscured by background interference. Experimental results demonstrate that on the highly challenging VisDrone and HIT-UAV aerial datasets, the proposed DIF-DETR network outperforms existing mainstream models with 30.5% mAP and 82.3% mAPtest, respectively. Simultaneously, it significantly reduces computational cost to 43.6 GFLOPs with only 13.4M parameters, achieving an optimal balance between detection accuracy and computational efficiency. This demonstrates that through the synergistic effects of three core innovations, DIF-DETR significantly enhances detection accuracy and robustness for small objects in aerial images, providing an effective solution for object detection tasks in aerial scenarios.

Keywords: Object Detection, Multi-scale Feature Enhancement, Transformer Architecture, UAV Images.

1. Introduction

Small object detection has always been a critical and challenging task in computer vision, particularly in complex backgrounds and multi-scale scenes. In recent years, Transformer-based detectors like DETR have achieved remarkable results in large object detection, but their performance remains suboptimal for small object detection. Analysis of the VisDrone 2019-DET dataset reveals it contains a vast collection of images and video sequences, each meticulously annotated at the pixel level to identify and classify diverse object categories such as pedestrians, vehicles, bicycles, and more. However, the VisDrone dataset presents challenges including complex backgrounds, diverse object scales, dense object distributions, and unique perspective shifts and motion blur resulting from the drone viewpoint. Further data screening revealed a critical fact: over 60.1% of annotated samples represent small objects, while medium and large objects account for 34.2% and 5.7% respectively, highlighting the importance of small object detection in aerial applications.

Although traditional convolutional neural networks (CNNs) possess certain advantages in extracting local features, they often suffer from insufficient feature representation when handling small objects. To address this issue, numerous researchers have focused on improving models like DETR. For instance, PointDet++ [1] combines human pose estimation techniques, attempting to integrate local and global features to enhance small object detection performance. However, these approaches still fall short in the depth of feature fusion and utilization of contextual information, leaving room for improvement in accuracy and robustness for small object detection. NLFFTNet [2] introduces a Transformer-based convolutional network with non-local feature fusion to capture long-range semantic relationships across feature layers. While effective, its high computational cost limits its applicability in real-time scenarios. DeoT [3] combines an encoder-only Transformer with a feature pyramid fusion module, enhancing fusion effects through a Channel Refinement Module (CRM) and Spatial Refinement Module (SRM). However, it remains underpowered when handling small objects, particularly in aerial scenes with high background noise. HTDet [4] employs a Fine-Grained Feature Pyramid Network (FPN) for feature fusion. While showing improvement, it still faces challenges in small object detection due to insufficient feature representation, particularly in extracting target texture and shape information.

Furthermore, applying DETR-like models (e.g., Deformable DETR [5] and Sparse DETR [6]) directly on high-resolution feature maps of aerial images poses efficiency challenges. These models process significantly more pixels when handling high-resolution images, leading to

substantially increased computational complexity and memory requirements. Although the deformable attention mechanism in Deformable DETR is more efficient than traditional global attention, its computational cost remains high on high-resolution aerial images. This is particularly pronounced in real-time demanding aerial scenarios, limiting its widespread deployment in practical applications. To achieve synergistic optimization of high accuracy and high efficiency, RT-DETR [7] innovatively proposes an efficient hybrid encoder architecture. By decoupling multi-scale feature interactions into two subtasks—intra-scale interactions (AIFI) and cross-scale fusion (CCFF)—it substantially reduces redundant computational overhead. However, this approach faces a trade-off challenge in backbone network selection: using deeper ResNet50/101 as the backbone directly compromises inference real-time performance due to high computational demands; conversely, opting for the lightweight ResNet18 to enhance speed inevitably sacrifices detection accuracy.

Based on this analysis, we propose DIF-DETR, a lightweight aerial small object detection model based on RT-DETR. We innovatively employ a backbone network incorporating the DynamicIncep-EnhanceBlock as the feature extraction foundation. This backbone is specifically tailored to the characteristics of small aerial targets. Through dynamic Inception multi-branch deep convolutions and an adaptive weight allocation mechanism, it efficiently captures multi-scale and long-range contextual information while reducing computational redundancy. This enhances the initial feature representation of small targets, improving feature extraction performance in complex aerial scenes. Subsequently, the extracted features are fed into our proposed Context-Aware Bidirectional Fusion module. This module employs a bidirectional attention fusion mechanism to facilitate effective information transfer between high-resolution low-level detail features and low-resolution high-level semantic features, further enhancing multi-scale feature fusion. This fusion strategy better preserves small object detail information, which is crucial for improving detection accuracy of small aerial targets.

The main contributions of this paper are as follows:

Proposing the DIENet backbone feature extraction network embedded with DIEBlock. By utilizing dynamic Inception multi-branch deep convolutions and an adaptive weight allocation mechanism, it efficiently captures multi-scale, long-range contextual information. This reduces computational redundancy while strengthening the initial feature representation of small aerial targets, thereby improving feature extraction performance in complex scenarios.

Designs the Context-Aware Bidirectional Fusion (CABF) module, which employs a bidirectional attention fusion mechanism to facilitate comprehensive information exchange between high- and low-resolution features. This enhances multi-scale feature fusion effectiveness, significantly optimizing small target detection performance in aerial imagery.

Our proposed DIF-DETR underwent comprehensive evaluation on the VisDrone2019-Det and HIT-UAV datasets. Results demonstrate that our model outperforms existing DETR-based and non-DETR detectors in both detection accuracy and computational speed. Notably, our model achieves significant improvements in accuracy for small object detection.

2. Related Work

2.1. Small Object Detection in Aerial Images

In recent years, research on aerial object detection has made progress in multiple aspects. Some studies suppress certain features to help models focus on small foreground objects. SDP [8] proposed a scale-decoupling module to mask large object features and concentrate on small objects. Additionally, it introduced sparse non-local attention mechanisms and adaptive anchor matching strategies to enhance small object detection performance. FSDet [9] designed a class-aware context aggregation module that integrates intra-class contextual information while suppressing background interference, thereby addressing the foreground-background imbalance issue. DEA-Net [10] proposed a novel training generator that interacts between anchor-based regions of interest (RoIs) and anchor-free units, improving anchor quality. To tackle background redundancy, PENet [11] employs coarse anchor-free detectors to efficiently predict centroid points of small object clusters, while using fine-grained anchor-free detectors to pinpoint precise locations within these clusters. UFPMP-Det [12] merges subregions generated by coarse detectors into a unified image and designs a multi-agent detection network to enhance small object detection accuracy. Although these methods achieve high accuracy, they require multiple inferences on the same image, resulting in low efficiency. Furthermore, all the aforementioned approaches first predict subregions of the input image and then detect objects in a second stage, hindering the realization of an end-to-end workflow.

2.2. Vision Transformers (ViTs) for Detection

In recent years, Transformer models have demonstrated formidable potential in visual tasks. ViT [13] pioneered the successful application of a pure Transformer architecture to image classification, establishing a new paradigm in computer vision. Subsequently, DETR [14] modeled object detection as a “collective prediction” problem, eliminating the non-maximum suppression step from traditional detection pipelines. However, its training convergence was slow, demanding substantial computational time. To accelerate convergence, subsequent studies like Conditional-DETR [15] enhanced the model's ability to capture spatial relationships by explicitly introducing positional information into input queries. While these methods perform well in natural scene object detection, they often fail to adequately address the unique challenges of aerial imagery, such as large object scale variations, dense object distributions, complex backgrounds, and numerous distractions. Models like DINO [16] enhance robustness on difficult samples by optimizing label noise handling and anchor box design. However, their complex architectures yield high accuracy on aerial datasets at the cost of significant computational overhead and prolonged inference latency, failing to meet the real-time detection demands of drone platforms.

2.3. Feature Pyramid Network

Furthermore, multi-scale feature fusion represents another crucial technical approach to enhancing detection accuracy. For instance, FENet [17] leverages multi-granularity deformable convolutions and high-resolution feature pyramids to extract rich spatial contextual information, thereby enhancing small object features; EfficientDet [18]

proposes the Bidirectional Feature Pyramid Network (BiFPN), which achieves weighted fusion of cross-layer features through residual connections and adaptive weighting mechanisms; CAD-Net [19] designed a dual-sensing attention module for both spatial and scale dimensions, guiding different pyramid levels to focus on target regions at corresponding scales; CE-FPN [20] introduced PixelShuffle operations to mitigate channel information loss during upscaling, yet channel compression issues persist in deep low-resolution features. Although these approaches optimize multi-scale representation capabilities to varying degrees, they generally fail to explicitly model feature redundancy. Redundant information significantly interferes with the already sparse semantic representation of small objects, limiting further improvements in detection performance.

The aforementioned solutions optimize various detectors for object detection scenarios in aerial images. We propose a novel detection model specifically tailored for small object detection in aerial photography scenarios.

3. Method

In this section, we will detail the architecture and component modules of DIF-DETR. The overall network structure is introduced in Section 3.1, the backbone network DIENet in Section 3.2, the CABF module in Section 3.3, and the DGL-Block module in Section 3.4.

3.1. Overall Structure

The overall network architecture of DIF-DETR is shown in Figure 1. RT-DETR is adopted as the baseline, consisting of a backbone, a multi-scale fusion network, and a Transformer decoder with an auxiliary prediction head. First, we employ the newly designed DIENet as the backbone network to extract hierarchical features. DIENet comprises four stages, each built from DIEBlocks. Each stage applies downsampling with a factor of 2, yielding four feature maps at distinct resolutions: H/4×W/4, H/8×W/8, H/16×W/16, and H/32×W/32. The output features {P3, P4, P5} from the last three stages of the backbone network serve as input to the CABF fusion network. After intra-scale interactions and inter-scale fusion, these multi-scale features are transformed into a sequence of image features. Subsequently, an IOU-aware query selection method selects a fixed number of image features from the encoder's output sequence as the initial object queries for the decoder. Finally, the decoder incorporates an auxiliary prediction head to iteratively refine these object queries, generating bounding boxes and confidence scores.

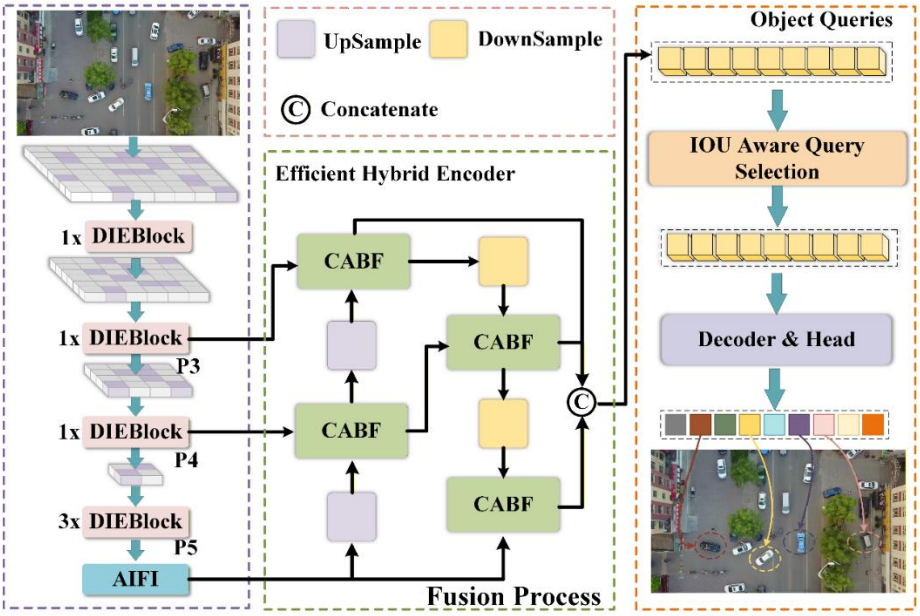

Figure 1. DIF-DETR Overall Structure Diagram

3.2. Additive Convolution Gate Network

In intensive visual prediction tasks such as object detection and image segmentation, the performance of feature extraction and fusion modules directly determines the model's final accuracy and inference efficiency. Traditional mainstream modules (e.g., C2f, C3) rely on static convolutional architectures, exhibiting significant limitations when processing complex, variable multi-scale inputs. Specifically, these modules employ fixed-size convolutional kernels, leading to limited kernel selection that struggles to balance fine-grained details with global semantic information, resulting in inadequate multi-scale feature capture capabilities. simultaneously, their shortcut branches merely perform simple feature concatenation, failing to achieve deep integration between low-level details and high-level semantics. Channel information exchange relies on ordinary 1×1 convolutions, lacking targeted feature selection mechanisms. This results in severe loss of cross-layer and cross-channel information, with prominent channel redundancy issues; Furthermore, the fixed weight-to-stride ratio of convolutional kernels after training prevents dynamic adaptation of feature extraction methods to variations in image texture, lighting, and object morphology. This results in limited architectural flexibility and weak adaptability. Finally, the extensive use of standard convolutions and repetitive feature processing flows cause severe computational and parameter redundancy. This hinders real-time inference on resource-constrained devices like edge computing systems, limiting the model's practical deployment value.

To address these challenges, this paper proposes DIENet, a backbone network based on Dynamic Inception Convolution. Its architecture, illustrated in Figure 2, is built upon the

DIEBlock core module. This module focuses on three key design principles: dynamic adaptive convolution, multidimensional efficient fusion, and lightweight architecture. It enhances feature expression capabilities while maintaining real-time inference performance. First, DIEBlock adopts the CSP design philosophy, employing a "basic branching-dynamic enhancement-feature aggregation" architecture. It replaces the standard convolutional blocks of traditional CSP with stacked InceptionMixerBlocks, upgrading static convolutions to dynamic adaptive convolutions while enhancing cross-layer, cross-scale, and cross-channel feature fusion capabilities.

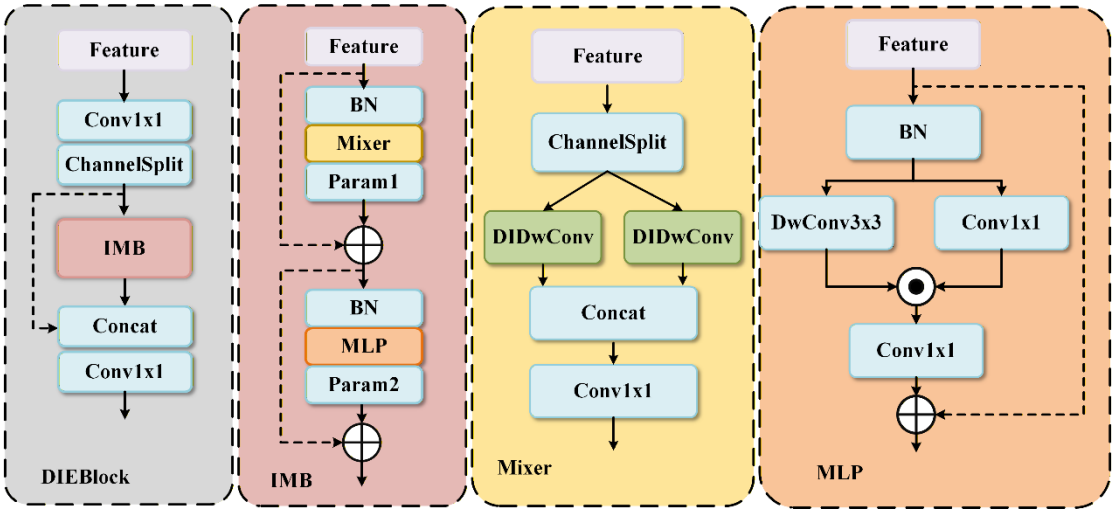

Figure 2. DIEBlock Structure Diagram

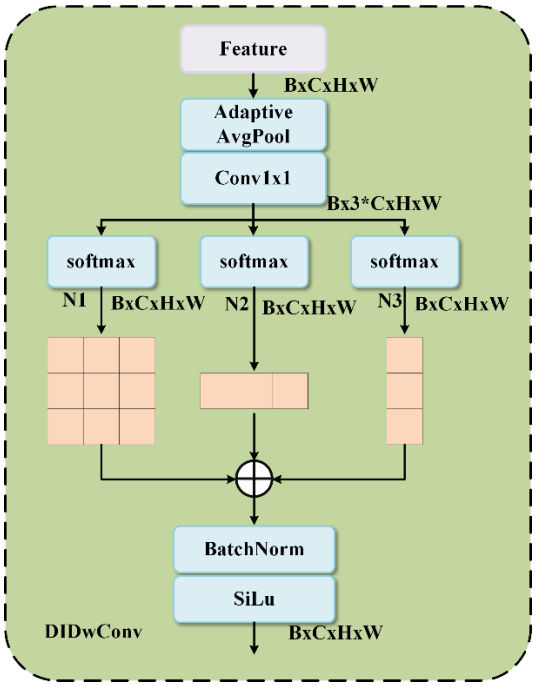

Figure 3. DIDwConv Block Diagram

To ensure module compatibility and lightweight design, DIEBlock inherits CSP's core splitting and aggregation mechanisms without significantly altering the overall network architecture. Specifically, the module first maps input channel c1 to a target number of channels (determined by c2 and scaling factor e) via a 1×1 convolution. It then splits the input into a shortcut branch and a main feature branch. Subsequently, the shortcut branch directly preserves original feature information for subsequent cross-layer residual fusion to mitigate gradient vanishing in deep networks. The main feature branch is fed into stacked InceptionMixerBlock modules for deep feature enhancement. Finally, the enhanced features from the main branch are concatenated with the original features from the shortcut branch. A 1×1 convolution achieves channel-dimensional fusion and normalization, outputting high-quality feature maps rich in both detail and semantic information.

The InceptionMixerBlock serves as the core innovative unit of the DIEBlock, employing a two-stage serial residual architecture of "Norm-Mixer-MLP." Through two sequentially connected processing stages, it achieves dynamic cross-scale mixing in the spatial dimension and nonlinear transformation in the channel dimension, specifically addressing the core issues of traditional modules. The first stage focuses on dynamic cross-scale spatial feature mixing, centered around the DynamicInceptionMixer module. It achieves input-driven multi-scale feature extraction through a "channel grouping - dynamic depthwise convolution - branch aggregation" workflow. This process first applies BatchNorm to normalize input feature channels, eliminating internal covariate shifts and providing stable input distributions for dynamic convolutions to ensure training convergence. Subsequently, input feature channels are uniformly split at a 1:1 ratio, with the number of groups matching the length of the predefined convolution kernel size list. This reduces computational overhead within a single stage while enabling parallel feature processing in the channel dimension. Each channel group corresponds to an independent DynamicInceptionDWConv module (as shown in Figure 3). This module incorporates three complementary deep convolution branches: square convolutions capture local neighborhood features, horizontal strip convolutions capture lateral long-range dependencies, and vertical strip convolutions capture longitudinal long-range dependencies This approach covers multi-scale and multi-directional feature requirements. Concurrently, the Dynamic Kernel Weight (DKW) mechanism is introduced: adaptive global average pooling combined with 1×1 convolutions generates dynamic weights. After dimension reshaping and Softmax normalization, these weights are assigned as adaptive inputs to the three convolutional branches, enabling "dynamic adjustment of kernel contribution based on input feature variations." Subsequently, the outputs from the three convolutional branches are multiplied by their respective dynamic weights and summed to complete multi-scale feature fusion. A 1×1 convolution then concatenates the enhanced

features across channel groups, enabling cross-group information exchange to mitigate information isolation caused by channel grouping. Finally, a learnable layer scaling parameter (initial value 0.01) is introduced. Through dimension expansion, it enables per-channel feature scaling to control the magnitude of residual gradient propagation during early training, preventing gradient vanishing. Simultaneously, the DropPath strategy randomly discards some residual transmission paths during training, achieving model regularization to enhance generalization capability.

Phase Two builds upon Phase One's output features, focusing on channel-dimensional feature transformation. Convolutional Gated Linear Units [21] filter and transform channel features, addressing insufficient channel information fusion and redundancy. This stage first re-performs batch normalization on the first stage's output features to provide stable input for channel transformation. Subsequently, ConvolutionalGLU replaces traditional fully-connected layers (MLP), retaining feature space information while filtering effective channel features through gating mechanisms to suppress redundant information and improve channel information utilization. Its nonlinear expressive capability surpasses that of ordinary 1×1 convolutions, enabling better capture of inter-channel dependencies. Finally, LayerScale and DropPath designs are employed to control feature magnitude and enforce regularization. Residual connections fuse the transformed features with the module's input features, ensuring seamless information transfer between both processing stages.

3.3. Context-Aware Bidirectional Fusion

The CCFF module employed by the real-time object detection framework RT-DETR exhibits significant inherent flaws in its multi-scale feature fusion process, which relies solely on simple scale alignment and channel concatenation: First, the absence of a context-semantic guidance mechanism hinders the precise capture of critical discriminative information—such as edge contours and texture details—of small objects during fusion, resulting in ambiguous feature representations for small targets; Second, the indiscriminate concatenation approach readily introduces redundant background information and invalid feature interference, intensifying noise pollution in the feature space. Third, the unidirectional feature propagation logic limits complementary enhancement between features of different scales, resulting in weak overall feature representation capability.

To overcome these bottlenecks, this paper proposes the Context-Aware Bidirectional Fusion (CABF) module, as shown in Figure 4. Embedded within RT-DETR's multi-scale feature fusion pipeline, it achieves precise performance enhancement for small object detection. Centered on the design philosophy of "lightweight architecture enabling high-precision fusion," this module innovatively constructs an adaptive, context-aware bidirectional feature fusion paradigm. Through the synergistic action of four core mechanisms, it achieves efficient complementarity and enhancement between features of different resolutions (Fea1 and Fea2).

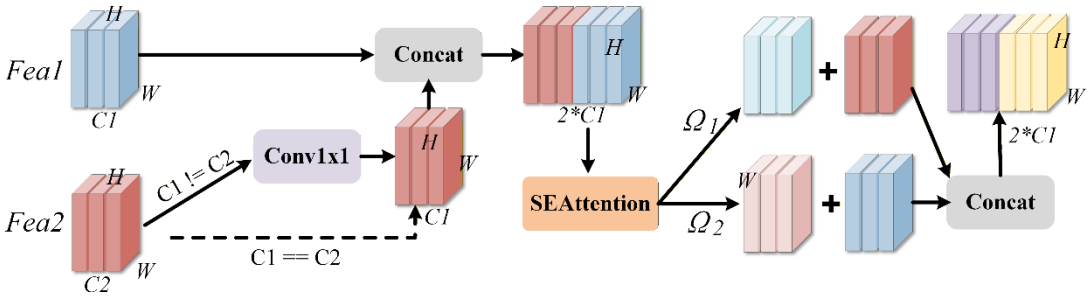

Figure 4. CABF Structure Diagram

Specifically: First, to address the inherent channel dimension differences across feature maps of varying scales, the module incorporates lightweight 1×1 convolution kernels to align the channel dimensions of Fea1 and Fea2. Second, a context-aware SE attention enhancement mechanism is introduced. Abandoning the crude fusion logic of traditional CCFF modules that perform indiscriminate concatenation, after channel alignment, Fea1 and Fea2 are first concatenated along the channel dimension to construct a global feature interaction space. Subsequently, it embeds a Squeeze-and-Excitation (SE) attention module. Through global average pooling, it aggregates global contextual information and adaptively learns dependencies between feature channels, precisely generating differentiated channel importance weights Ω1 (corresponding to Fea1) and Ω2 (corresponding to Fea2). This weight allocation mechanism selectively enhances the expression of critical information such as small object edge details and semantic features while suppressing redundant background and noise interference, achieving efficient purification of the feature space. Furthermore, a bidirectional cross-enhancement fusion strategy is designed. We perform element-wise multiplication between the weight features generated by SE attention and their corresponding original features to obtain weighted-enhanced feature maps. Building upon this, we employ bidirectional cross-addition operations: fusing the weighted features of Fea1 with the original features of Fea2, and the weighted features of Fea2 with the original features of Fea1. This constructs bidirectional flow channels for cross-scale features. This strategy overcomes the limitations of traditional unidirectional feature propagation, enabling strong semantic information from low-resolution features and fine-grained detail information from high-resolution features to mutually guide and reinforce each other. This significantly enhances the semantic discriminative power and detail capture accuracy of the fused features, providing core support for the precise localization and recognition of small targets. The above steps can be expressed as follows:

$$x_0 \in \mathbb{R}^{C_0 \times H \times W}, x_1 \in \mathbb{R}^{C_1 \times H \times W} \quad (1)$$

$$[\alpha, \beta] = \mathrm{SE}([\mathrm{Adjust}(x_0) \parallel x_1]) \quad (2)$$

$$\mathrm{Adjust}(x_0) = \begin{cases} \mathrm{Conv}_{1\times1}(x_0) & C_0 \neq C_1 \\ x_0 & C_0 = C_1 \end{cases} \quad (3)$$

$$y = [\mathrm{Adjust}(x_0) + x_1 \odot \beta \parallel x_1 + \mathrm{Adjust}(x_0) \odot \alpha] \quad (4)$$

Here, x_0 and x_1 denote the two inputs to the fusion

network, $\mathrm{Adjust}(x_0)$ represents the number of x_0, and $\|$ denotes the concatenation operation, $\mathrm{SE}(*)$refers to the SE attention module, which calculates attention weights for the concatenated features to obtain two weights α and β. $\odot$ denotes element-wise multiplication.

4. Experiment

4.1. Dataset Introduction

VisDrone2019-Det [22] is a drone-view object detection dataset. Its training set comprises 6,471 images, the validation set contains 548 images, and the test challenge set includes 1,580 images, with a total of 10 object categories annotated. The core challenge of this dataset lies in the extreme variation in object scales due to differing UAV flight altitudes, with most objects being small-sized (pixel dimensions < 32). This demands strong multi-scale feature capture capabilities from detection models.

HIT-UAV [23] is a specialized target detection dataset for high-altitude infrared thermal imaging from UAVs, constructed and open-sourced by a team from Harbin Institute of Technology. Its data originates from the curation of 43,470 frames of high-altitude aerial video footage captured by UAVs, ultimately yielding 2,898 high-quality infrared thermal images. These images cover 5 common urban scene objects: Person, Bicycle, Car, OtherVehicle, and DontCare, with a total of 24,899 annotated objects. The dataset follows a standard partitioning scheme: "Training Set (2,029 images) - Validation Set (290 images) - Test Set (579 images)", fully supporting the entire workflow of detection algorithm training, tuning, and performance evaluation.

Notably, as shown in Figure 5, the core object categories in both the VisDrone2019-Det and HIT-UAV datasets are primarily "Person" and "Car." This provides a consistent foundation at the category level for subsequent cross-dataset model generalization validation.

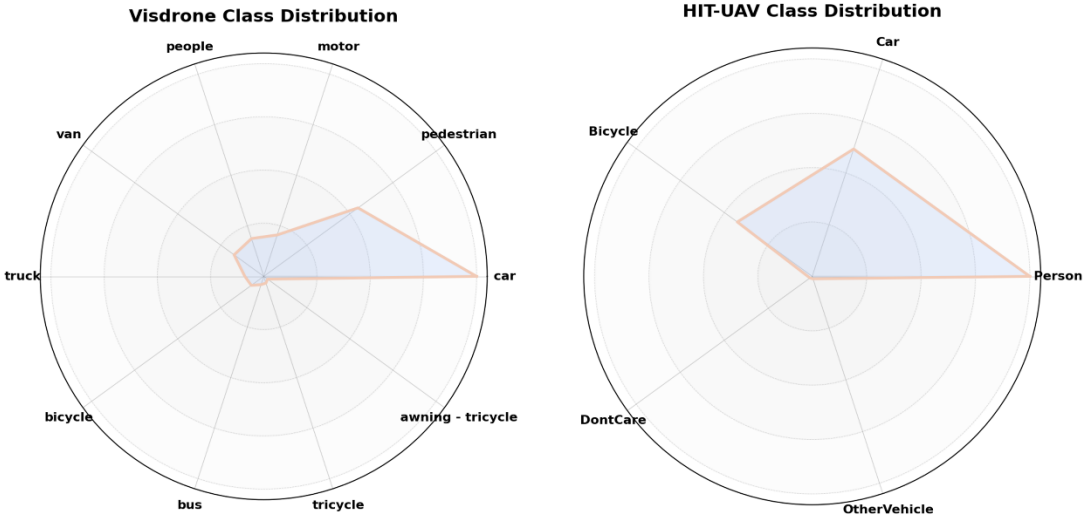

Figure 5. Visdrone and HIT-UAV Category Distribution

4.2. Evaluation Metrics

Average Precision: The performance evaluation metrics for object detection are Average Precision (AP) per class and Mean Average Precision (mAP) across all classes. Typically, AP is the integral of the PRC over recall from 0 to 1, representing a more comprehensive and robust indicator of detection performance. For a detector, a higher mAP clearly indicates superior detection performance. In the design, evaluation, and optimization of deep learning models, GFLOPs and Params are equally two core performance metrics, measuring the model's computational complexity and parameter size respectively. These directly impact training efficiency, inference speed, and deployment costs. The definitions of AP and mAP are as follows:

$$\mathrm{AP} = \int_0^1 \mathrm{P}(\mathrm{R})\, d\mathrm{R} \tag{5}$$

$$\mathrm{mAP} = \frac{1}{N_{\mathrm{cls}}} \sum_{i=1}^{N_{\mathrm{cls}}} \mathrm{AP}_i \tag{6}$$

4.3. Implementation Details

Our experimental setup utilized the PyTorch framework (PyTorch 1.9.0) on Ubuntu 20.04. Experiments for Visdrone and AI-TOD were conducted on a single NVIDIA GeForce RTX 3090 GPU, each equipped with 24 GB of memory. The batch size was set to 4. Input images were uniformly resized to 640 × 640 pixels. Horizontal flipping at 0.5 was employed as the sole data augmentation technique, with no additional methods applied. Following standard practice, all models were trained using the AdamW optimizer with a learning rate of 0.0001 and weight decay of 0.1. Momentum was set to 0.9, representing the weight of the previous gradient update.

4.4. Experimental Comparison and Analysis

4.4.1. Ablation Studies

To evaluate the performance of each module in the proposed DIF-DETR, ablation studies were conducted on the Visdrone dataset. The effectiveness of each module was tested by disabling it one at a time. Results are shown in Table 1. The baseline model was constructed using RT-DETR with ResNet18 as the backbone network.

This table systematically compares the performance and computational resource consumption of the baseline model versus different combinations of the DynamicIncep-EnhanceBlock (DIENet) and Context-Aware Bidirectional Fusion (CABF) modules for detecting small aerial objects. Model performance was evaluated using the commonly adopted metrics in object detection: Mean Average Precision (mAP) and Mean Average Precision at 0.5 Intersection over Union (mAP_{50}). Resource consumption was quantified through Gigaflops per second (GFLOPs) and total parameters (Params).

Experimental results demonstrate: When using only the Baseline model, performance on aerial small object detection tasks is limited, achieving only 26.7 mAP and 45.3 mAP_{50}, accompanied by high computational redundancy (58.9 GFLOPs) and moderate parameter count (19.4M). Integrating DIENet onto the Baseline enhances the model's feature representation for small objects, boosting mAP to 28.5 and mAP_{50} to 47.5. Computational load significantly decreases to 43.6 GFLOPs, and parameters shrink to 13.2M, demonstrating DIENet's dual advantages of "lightweighting and feature enhancement." Adding the CABF module to the Baseline model further improved mAP to 29.2 and mAP_{50} to 47.6, but computational cost rebounded to 57.0 GFLOPs and parameters increased to 20.0M. This indicates that while CABF optimizes feature fusion, its standalone use increases resource overhead. When the Baseline simultaneously integrates DIENet and CABF, model performance reaches its optimal level (mAP=30.5, mAP_{50}=49.7), while computational cost and parameter count remain within low-overhead ranges (43.6 GFLOPs, 13.3M). — This result clearly demonstrates the synergistic collaboration between DIENet's efficient feature extraction capability and CABF's bidirectional feature fusion mechanism: the former achieves lightweight feature capture tailored for small aerial targets through dynamic branching and weight allocation, while the latter enhances the complementarity of multi-scale features via bidirectional attention transfer. Ultimately, this approach significantly improves small object detection accuracy in complex aerial scenes while maintaining resource efficiency.

Table 1. Performance of Each Module on Visdrone

Baseline	DIENet	CABF	mAP	mAP50	GFLOPs	Params(M)
√			26.7	45.3	58.9	19.4
√	√		28.5	47.5	43.6	13.2
√		√	29.2	47.6	57.0	20.0
√	√	√	30.5	49.7	43.6	13.3

4.4.2. Comparison with State-of-the-Art Methods on the Visdrone-2019-Det val Dataset

Table 2. Performance Comparison of Different Methods on the Visdrone-2019-Det val Dataset

Models	mAP	mAP50	Params (M)	GFLOPS	FPS
QueryDet [24]	28.3	48.1	-	212	7
TOOD [25]	26.1	42.3	32.04	199	33.9
RTMDet-L [26]	29.3	46.8	52.26	79.97	32.1
DTSSNet [27]	24.2	39.9	-	50.34	75.51
YOLOv8-M [28]	27.1	44.4	25.9	78.9	83
YOLOv9-M [29]	27.2	44.6	20.0	76.3	77
YOLOv10-M [30]	26.5	43.1	16.4	63.5	70
YOLOv11-M [31]	27.7	45.2	20.04	67.7	86
YOLOv12-M [32]	26.3	43.1	20.11	67.2	83
MHAF-YOLO [33]	28.1	45.2	15.81	67.6	83
FBRT-YOLO-M [34]	28.4	45.9	7.2	58.7	94
FBRT-YOLO-L [34]	29.7	47.7	14.6	119.2	70
FBRT-YOLO-X [34]	30.1	48.4	22.8	185.8	52
DETR [14]	24.1	40.1	41.56	187	16.8
Deformable-DETR [5]	27.1	49.2	40.7	193	28.5
Conditional-DETR [15]	21.7	39.5	43.4	101	44.4
RT-DETR(R18) [7]	26.7	45.3	19.4	58.9	90
RT-DETR(R34) [7]	27.2	46.0	31.0	92.0	88
RT-DETR(R50) [7]	28.8	48.3	42.0	136.0	65
DIF-DETR (ours)	30.5	49.7	13.4	43.6	96

Table 2 evaluates QueryDet, TOOD, RTMDet-L, DTSSNet, the YOLO series (v8-M to v12-M), MHAF-YOLO, FBRT-YOLO series (M/L/X), DETR and its variants (Deformable-DETR, Conditional-DETR, RT-DETR series), and the proposed DIF-DETR (ours). In terms of detection accuracy, DIF-DETR achieved the highest mAP score of 30.5, surpassing high-performance models like FBRT-YOLO-X (30.1) and RTMDet-L (29.3), while also significantly outperforming classic approaches such as QueryDet (28.3) and RT-DETR (R50) (28.8). For mAP50, DIF-DETR achieved 49.7—the highest among all models—surpassing Deformable-DETR (49.2), RT-DETR (R50) (48.3), and QueryDet (48.1), demonstrating superior stability in high-confidence detection scenarios; Regarding model complexity, DIF-DETR's parameter count is only 13.4M, slightly higher than the lightweight model FBRT-YOLO-M (7.2M) but significantly lower than RTMDet-L (52.26M) and RT-DETR (R50) (42.0M), and even falls below mid-weight models like YOLOv11-M (20.04M) and MHAF-YOLO (15.81M). Its computational cost (GFLOPS) is as low as 43.6, ranking among the lowest among all reference models. This represents a significant reduction compared to models like QueryDet (212), Deformable-DETR (193), and FBRT-YOLO-X (185.8), demonstrating an exceptionally prominent resource efficiency advantage; In terms of inference efficiency, DIF-DETR achieves 96 FPS, placing it among the

top models in this comparison. It is only slightly behind FBRT-YOLO-M (94) and RT-DETR (R18) (90). However, the latter two models have mAP values of merely 28.4 and 26.7, respectively, showing a significant gap in detection accuracy compared to DIF-DETR. Meanwhile, FBRT-YOLO-X (mAP=30.1), which has detection accuracy close to DIF-DETR, achieves 180 GFLOPS. 26.7, showing a significant gap in detection accuracy compared to DIF-DETR. Meanwhile, FBRT-YOLO-X (mAP=30.1), which approaches DIF-DETR's accuracy, achieves only 52 FPS—less than 60% of DIF-DETR's inference speed. Overall, DIF-DETR achieves a triple performance advantage in this comparison: optimal detection accuracy, minimal model complexity, and leading inference efficiency. It effectively breaks through the traditional trade-off bottleneck between accuracy, efficiency, and resource consumption in object detection tasks, making it the superior technical solution for high precision, lightweight design, and high inference speed in current scenarios.

Figure 6 presents the visual comparison results of this experiment, covering five models: YOLOv9M, YOLOv11M, YOLOv12M, RT-DETR (R18), and the proposed DIF-DETR model. The results clearly demonstrate that DIF-DETR achieves optimal performance across all time periods, particularly excelling in detecting small objects—a critical challenge in object detection. Specifically, in the first, third, and fourth test image sets, models like YOLOv9M, YOLOv11M, and YOLOv12M exhibited varying degrees of missed detections and false positives for small targets such as motorcycles, bicycles, and riders—objects with low pixel coverage and sparse feature information. RT-DETR (R18) could identify some small objects but still suffered from insufficient localization accuracy. In contrast, DIF-DETR leveraged its innovative feature enhancement mechanism and context information fusion strategy to precisely capture subtle features of small objects and achieve stable detection. In complex scenes with dense object distribution, YOLO series models often suffer from redundant detections and inaccurate bounding boxes due to feature space confusion. RT-DETR (R18) performs relatively well thanks to its efficient query mechanism. DIF-DETR further enhances dense object discrimination and detection efficiency by introducing a dynamic interactive feature calibration module and adaptive anchor matching strategy. The above visual results fully validate DIF-DETR's core innovative value in small object detection and dense scenarios. Its designed feature enhancement and context fusion mechanisms effectively compensate for existing models' performance shortcomings in complex scenarios, achieving high-precision object detection across different time periods and environments.

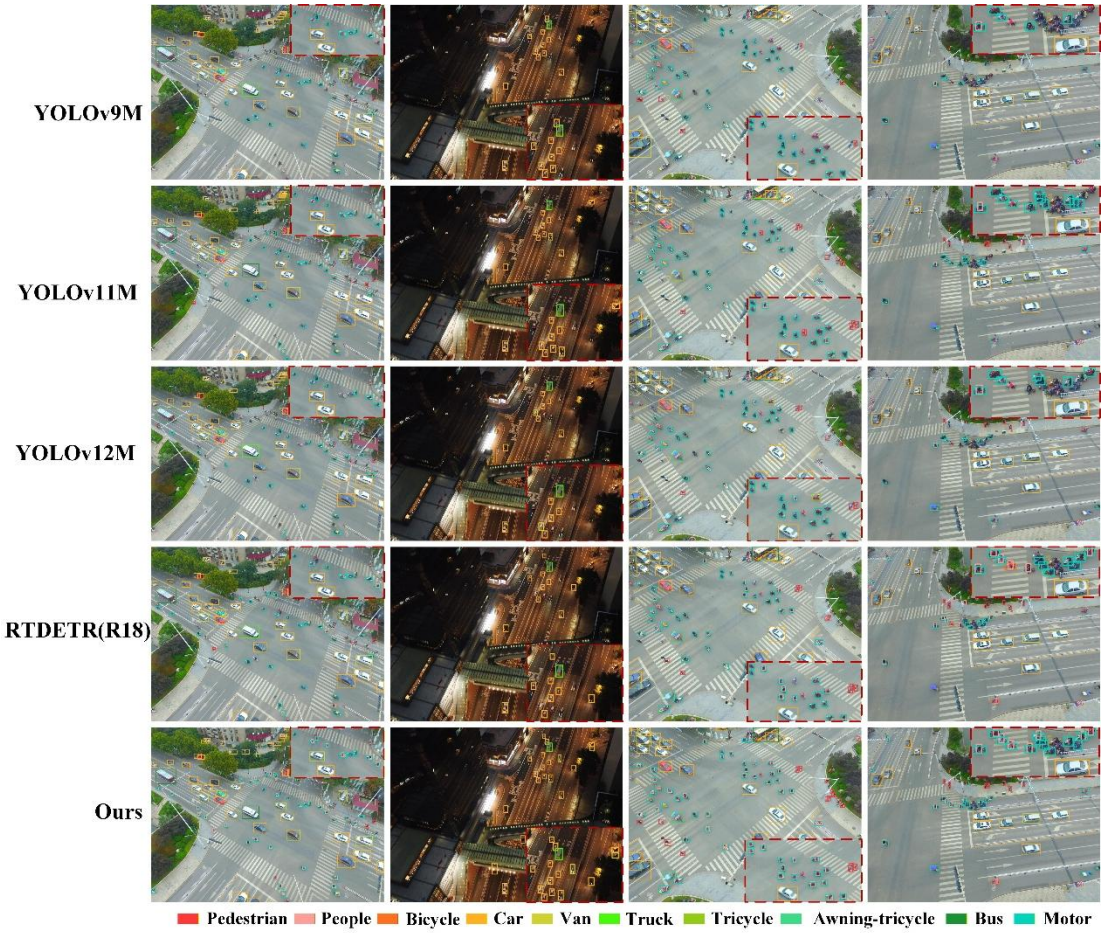

Figure 6. Comparison of DIF-DETR with Other Model Detection Results

Table 3 presents performance comparisons on the HIT-UAV dataset. Results indicate that YOLO series models (v8M to v12M) demonstrate stable performance across various detection categories. Among them, YOLOv11M achieves the best performance with an mAPtest score of 81.7%. while YOLOv12M achieved the highest scores for CAR and BI categories (96.0% and 92.1%), it showed significant decline in OV and DC categories, resulting in an overall mAPtest of only 78.0%. The RT-DETR series progressively improved across multiple metrics as the skeleton depth increased (from R18 to R50). RT-DETR(R50) achieved 94.5% and 95.1% on PE and CAR respectively, with an mAP50test of 82.3% matching DIF-DETR, but its computational cost (136.0 GFLOPs, 42.0M parameters) is significantly higher; In

contrast, DIF-DETR achieved optimal results across CAR (96.2%), OV (65.4%), DC (64.7%), and mAP50test (52.5%), while achieving the lowest computational load (43.6 GFLOPs) and fewest parameters (13.4M) among all models. This highlights its ability to significantly enhance computational efficiency while maintaining or even surpassing the detection accuracy of existing models. Its innovation lies in achieving an exceptional balance between high performance and low resource consumption through effective structural design.

Table 3. Performance Comparison of Different Methods on the HIT-UAV Dataset Where PE denotes Person, BI denotes Bicycle, OV denotes Other Vehicle, and DC denotes Don't Care

Models	PE	CAR	BI	OV	DC	mAPtest	mAP50test	GFLOPs	Params
YOLOv8M [28]	91.7	95.5	91.6	64.4	62.8	81.4	51.7	78.7	25.8
YOLOv10M [30]	90.6	94.7	91.2	63.9	62.6	80.6	51.6	58.9	15.3
YOLOv11M [31]	92.0	95.3	91.8	64.5	63.9	81.7	52.1	67.7	20.0
YOLOv12M [32]	91.4	96.0	92.1	59.9	52.6	78.0	51.5	59.5	19.5
RT-DETR(R18) [7]	92.6	93.3	90.7	65.3	64.5	81.6	51.3	58.9	19.4
RT-DETR(R34) [7]	94.3	94.7	90.8	63.9	64.4	82.0	50.9	92.0	31.0
RT-DETR(R50) [7]	94.5	95.1	91.6	64.3	64.6	82.3	51.8	136.0	42.0
DIF-DETR (ours)	93.9	96.2	90.9	65.4	64.7	82.3	52.5	43.6	13.4

In Figure 7, experiments quantitatively compare four models—YOLOv9M, YOLOv11M, YOLOv12M, and the proposed DIF-DETR—across three core metrics: false positive rate (FPR), false negative rate (FNR), and true positive rate (TPR). These comparisons focus on sparsely populated scenarios and scenarios with densely packed small objects. In sparse object scenarios, YOLOv9M, YOLOv11M, and YOLOv12M all demonstrate detectable capabilities, achieving moderate true positive rates. False positives and false negatives occur only occasionally in edge cases with low contrast or minor occlusions. However, when the scene shifts to complex conditions featuring both dense and small objects, the detection performance of these three models significantly degrades. This manifests as a substantial increase in false negative rates, where numerous small-sized objects such as motorcycles, bicycles, and cyclists are overlooked due to sparse feature information, mutual occlusion between objects, or severe background interference. Concurrently, false positive rates also rise, resulting in background clutter being misclassified as objects or duplicate detections of the same object. In contrast, the proposed DIF-DETR maintains stable and outstanding detection performance across both scenarios, achieving the lowest false positive and false negative rates among all compared models while attaining optimal correct detection rates. This advantage stems from DIF-DETR's innovative feature enhancement and context fusion mechanism. This mechanism effectively extracts subtle feature information from small objects while suppressing target confusion and background interference in dense scenes through a dynamic feature calibration strategy. This fully validates DIF-DETR's design superiority and engineering practicality in complex scenarios.

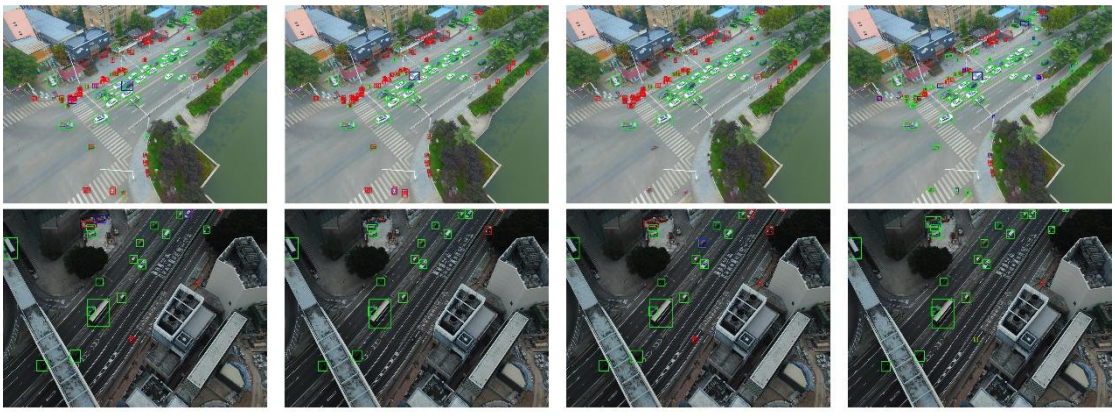

Figure 7. Comparison of false negatives and false positives across different models, where red boxes indicate false negatives, blue boxes indicate false positives, and green boxes indicate correct detections.

5. Conclusion

This paper addresses core challenges in aerial photography scenarios—low detection accuracy for small objects, strong background interference, and excessive computational load of existing high-performance models—by proposing DIF-DETR, a lightweight and efficient Transformer-based object detection network. This model achieves breakthrough performance through three core innovations: First, it designs and integrates the DIENet backbone with embedded Dynamic Inception Enhancement Blocks (DIEBlock). By leveraging multi-branch deep convolutions and input-adaptive weight allocation, it efficiently captures multi-scale and long-range contextual information at extremely low computational cost (43.6 GFLOPs), significantly enhancing the initial feature representation of small objects. Second, it innovatively introduces the Context-Aware Bidirectional Fusion (CABF) module. Guided by the Self-Attention (SE) mechanism, this module implements a bidirectional cross-enhancement strategy that achieves deep semantic and detailed complementary fusion between high- and low-level features. This effectively suppresses background noise and enhances the discriminative features of small targets. Finally, the optimized Dual-Branch Complementary Enhancement

module further ensures the depth and efficiency of feature extraction.

Comprehensive experiments on the VisDrone and HIT-UAV standard aerial datasets fully validate the superiority of DIF-DETR. The model achieves 30.5% mAP and 49.7% mAP50 on the VisDrone validation set, along with 82.3% mAPtest and 52.5% mAP50test on the HIT-UAV test set, matching or surpassing state-of-the-art detection models in accuracy. Simultaneously, DIF-DETR achieves real-time inference speeds of up to 96 FPS with minimal computational complexity (43.6 GFLOPs) and fewest parameters (13.4M), striking an exceptional balance across accuracy, efficiency, and lightweight design. Ablation studies and visual analysis further demonstrate that its innovative modules synergize effectively, significantly reducing false negatives and false positives in dense small-object scenarios. In summary, DIF-DETR offers a highly competitive solution for high-precision real-time aerial target detection in resource-constrained environments, holding significant theoretical implications and practical application value.

Acknowledgment

This work was supported by Key Science and Technology Program of Henan Province (No.252102210091), Postgraduate Education Reform and Quality Improvement Project of Henan Province (YJS2025XQC12), UniversityYoungBackbone Teachers Program of Henan Province (No.2023GGJS053), Fundamental Research Funds for the Universities of Henan Province (No.NSFRF220414) and Excellent YoungTeachers Program of Henan Polytechnic University (No.2019XQG02).

References

[1] Tang Y, Wang B, He W, et al. Pointdet++: an object detection framework based on human local features with transformer encoder [J]. Neural Computing and Applications, 2023, 35(14): 10097-10108.

[2] Zeng K, Ma Q, Wu J, et al. NLFFTNet: A non-local feature fusion transformer network for multi-scale object detection [J]. Neurocomputing, 2022, 493: 15-27.

[3] Ding T, Feng K, Wei Y, et al. DeoT: an end-to-end encoder-only Transformer object detector [J]. Journal of Real-Time Image Processing, 2023, 20(1): 1.

[4] Chen G, Mao Z, Wang K, et al. HTDet: A hybrid transformer-based approach for underwater small object detection [J]. Remote Sensing, 2023, 15(4): 1076.

[5] Zhu X, Su W, Lu L, et al. Deformable detr: Deformable transformers for end-to-end object detection [J]. arXiv preprint arXiv:2010.04159, 2020.

[6] Roh B, Shin J W, Shin W, et al. Sparse detr: Efficient end-to-end object detection with learnable sparsity [J]. arXiv preprint arXiv:2111.14330, 2021.

[7] Zhao Y, Lv W, Xu S, et al. Detrs beat yolos on real-time object detection [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 16965-16974.

[8] Ma Y, Chai L, Jin L. Scale decoupled pyramid for object detection in aerial images [J]. IEEE transactions on geoscience and remote sensing, 2023, 61: 1-14.

[9] Han J, Ren Y, Ding J, et al. Few-shot object detection via variational feature aggregation [C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(1): 755-763.

[10] Chen Z, He Z, Lu Z M. DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention [J]. IEEE transactions on image processing, 2024, 33: 1002-1015.

[11] Tang Z, Liu X, Yang B. PENet: Object detection using points estimation in high definition aerial images [C]//2020 19th IEEE international conference on machine learning and applications (ICMLA). IEEE, 2020: 392-398.

[12] Huang Y, Chen J, Huang D. UFPMP-Det: Toward accurate and efficient object detection on drone imagery [C]//Proceedings of the AAAI conference on artificial intelligence. 2022, 36(1): 1026-1033.

[13] Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv:2010.11929, 2020.

[14] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers [C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 213-229.

[15] Meng D, Chen X, Fan Z, et al. Conditional detr for fast training convergence [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 3651-3660.

[16] Zhang H, Li F, Liu S, et al. Dino: Detr with improved denoising anchor boxes for end-to-end object detection [J]. arXiv preprint arXiv:2203.03605, 2022.

[17] Wang Z, Li L, Xue Y, et al. FeNet: Feature enhancement network for lightweight remote-sensing image super-resolution [J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-12.

[18] Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10781-10790.

[19] Zhang G, Lu S, Zhang W. CAD-Net: A context-aware detection network for objects in remote sensing imagery [J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(12): 10015-10024.

[20] Luo Y, Cao X, Zhang J, et al. CE-FPN: enhancing channel information for object detection [J]. Multimedia Tools and Applications, 2022, 81(21): 30685-30704.

[21] Shi D. TransNeXt: Robust Foveal Visual Perception for Vision Transformers [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 17773-17783.

[22] Du D, Zhu P, Wen L, et al. VisDrone-DET2019: The vision meets drone object detection in image cFALlenge results[C]//Proceedings of the IEEE/CVF international conference on computer vision workshops. 2019: 0-0.

[23] Suo J, Wang T, Zhang X, et al. HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection [J]. Scientific Data, 2023, 10(1): 227.

[24] Yang C, Huang Z, Wang N. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection [C]//Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2022: 13668-13677.

[25] Feng C, Zhong Y, Gao Y, et al. Tood: Task-aligned one-stage object detection [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE Computer Society,2021: 3490-3499.

[26] Lyu C, Zhang W, Huang H, et al. Rtmdet: An empirical study of designing real-time object detectors [J]. arXiv preprint arXiv:2212.07784, 2022.

[27] Chen L, Liu C, Li W, et al. Dtssnet: dynamic training sample selection network for uav object detection [J]. IEEE

Transactions on Geoscience and Remote Sensing, 2024, 62: 1-16.

[28] Jocher, G., Chaurasia, A., & Qiu, J. (2023). YOLOv8. Retrieved from https://github.com/ultralytics/ultralytics

[29] Wang CY, Yeh IH, Liao HYM (2024) YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. arXiv preprint arXiv:2402.13616 https://doi.org/10.48550/arXiv.2402.13616

[30] Wang A, Chen H, Liu L, et al. (2024) YOLOv10: Real-time end-to-end object detection. Advances in Neural Information Processing Systems 37 107984–108011 https://doi.org/10.48550/arXiv.2405.14458

[31] Jocher G, Chaurasia A, Qiu J (2024) Ultralytics YOLO. Zenodo https://doi.org/10.5281/zenodo.10983461

[32] Tian Y, Ye Q, Doermann D. Yolov12: Attention-centric real-time object detectors [J]. arXiv preprint arXiv:2502.12524, 2025.

[33] Yang Z, Guan Q, Yu Z, et al. Mhaf-yolo: Multi-branch heterogeneous auxiliary fusion yolo for accurate object detection [J]. arXiv preprint arXiv:2502.04656, 2025.

[34] Xiao Y, Xu T, Xin Y, et al. FBRT-YOLO: Faster and Better for Real-Time Aerial Image Detection [C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(8): 8673-8681.