Small Object Detection Using Multi-scale Detail Enhancement and Decoupled Detection Head*

Yixin Qiao[#,a], Xinyuan Zhou[#,a], Shiyong Lan[a,*], Wenwu Wang[b], Yao Li[a] and Guonan Deng[a]

[a]College of Computer Science, Sichuan University, Chengdu, 610065, China
[b]Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK

ARTICLE INFO

Keywords:
Small object detection
Feature pyramid network
Multi-scale contextual information
Decoupled detection head

ABSTRACT

Small object detection aims to identify and locate objects in an image that occupy a small proportion of pixels. Despite its critical role across various domains, this technology encounters challenges, such as a lack of sufficient feature information due to sparse pixels and vulnerability to environmental noise. To address these challenges, we present a novel network, termed Detail and Context Guided Small Object Detection Network (DCGNet). Firstly, in contrast to the traditional Feature Pyramid Network (FPN) which is unable to provide sufficient information for small objects, we introduce a Detail and Context Enhanced Feature Pyramid Network (DCE-FPN) module. The DCE-FPN transforms feature from the temporal domain into the frequency domain and then performs denoising on the high-frequency components to highlight the detailed information of small objects. Moreover, the DCE-FPN leverages multiple branches with different receptive fields to capture multi-scale contextual information to further differentiate small objects from the background. Secondly, to alleviate ambiguity from the limited features of small objects and the conflicting demands of classification and localization in existing detection heads, we propose an Enhanced Interaction Decoupled RCNN (EDI-RCNN) module, which independently improves task-specific features and promotes effective feature interaction to improve both classification accuracy and localization precision. Finally, extensive experiments on well-known public datasets (i.e., VisDrone2019, AI-TODv2, and DOTAv1.0) demonstrate that our proposed method achieves superior performance in detecting small and tiny objects. Compared to the optimal baseline, our proposed method achieves relative gains of 10.2%/1.7% and -%[1]/2.4% in the average accuracy performance indicators of small objects (AP_S) and tiny objects (AP_t) on the VisDrone2019/AI-TODv2 dataset, respectively, without compromising the detection performance for objects of regular size.

1. Introduction

Object detection is a critical technology with widespread applications in everyday life, including autonomous driving, security surveillance, and medical image analysis. By accurately identifying and localizing objects within images, this technology provides robust support for a wide range of applications and plays a critical role in improving safety and operational efficiency.

With the advancement in deep learning, various detection methods, such as Faster R-CNN [1], Cascade R-CNN [2], FCOS [3], SSD [4], and DETR [5], have been developed to learn robust feature representations for object detection. These methods outperform traditional approaches such as TDFP [6] and EFPN [7], which rely on manually engineered features and often struggle to capture complex backgrounds and diverse object characteristics. Although effective in many cases, deep learning–based methods remain limited for small object detection. They achieve strong results on objects of regular sizes but offer limited performance on objects smaller than 32×32 pixels [8]. As illustrated in Fig. 1, small objects often contain limited discriminative feature information and are highly vulnerable to environmental interference, such as variations in lighting and fog, leading to degraded detection performance.

Figure 1: A demonstration of characteristics of small objects. The images in the first row depict small objects with a significant absence of critical information, while the images in the second row illustrate the vulnerability of these small objects to interference from environmental noise.

*This research is supported by National Natural Science Foundation of China project 62371324. The codes are available at https://github.com/SYLan2019/DCGNet2.

*Corresponding author: lanshiyong@scu.edu.cn. # Equal contribution.

[1]The "-%" denotes that no AP_t result released on VisDrone2019 dataset.

To address these challenges, researchers have explored a range of strategies, such as multi-scale techniques. Models such as the Feature Pyramid Network (FPN) [9] and its variants, including TDFP [6] and EFPN [7], have shown notable effectiveness. By propagating rich semantic information from high-level features to low-level representations, these models substantially improve detection performance across objects of different scales, especially small objects. During feature fusion, however, upsampling techniques like bilinear interpolation can introduce redundancy, while channel reduction risks losing discriminative target features [10]. Although low-level feature maps contain rich detailed information, the initial responses for small objects remain extremely weak. The process of upsampling high-level features can introduce smoothing effects or aliasing artifacts. When these upsampled features are fused with their low-level counterparts, the resulting redundant information can function as noise. This added noise ultimately overwhelms the already weak high-frequency signals of small objects, causing a severe degradation in their feature representation.

Once robust multi-scale features are obtained, an effective detection head is essential for small object classification and localization. Traditional coupled detection heads, which share features between classification and regression [11], often introduce task interference [12], further weakening small object features and exacerbating environmental noise. Building on YOLOX [12], Ni et al. [13] propose a three-branch decoupled head with an additional confidence prediction branch, achieving improved performance for small object detection. However, these methods often overlook the distinct feature requirements of classification and localization, as well as the potential for collaborative interaction between them. Typically, classification relies on global contextual information to mitigate environmental noise [11], whereas localization depends on fine-grained details to accurately delineate object boundaries. Furthermore, given the extreme feature scarcity inherent to small objects, completely decoupling these two tasks can result in the loss of critical complementary information. Specifically, robust semantic cues from the classification branch can help the regression branch filter out background clutter and reduce false positives. Conversely, precise boundary cues from the regression branch can guide the classifier to focus on the core regions of the targets. Therefore, when feature information is limited, it is essential to effectively harness this complementarity to foster task interaction.

In this study, we propose a novel detection network, termed the Detail and Context Guided Small Object Detection Network (DCGNet), built upon the two-stage Cascade R-CNN framework [2]. DCGNet is designed to provide rich fine-grained details for precise small object localization while mitigating environmental noise interference to enhance small object recognition. More specifically, we propose a Detail and Context Enhanced Feature Pyramid Network (DCE-FPN), which addresses the loss of fine-grained details in the features of small objects extracted using a conventional FPN due to upsampling and channel reduction operations. DCE-FPN employs Discrete Wavelet Transform (DWT) [14] to decompose features into frequency components, applies adaptive denoising to enhance high-frequency details, and reconstructs refined features via Inverse DWT (IDWT) [14]. Multi-scale contextual information is captured using parallel branches with different receptive fields and fused effectively. Additionally, to reduce feature confusion introduced by coupled detection heads, we design an Enhanced Interaction Decoupled RCNN (EDI-RCNN), where classification features are enriched by a Global Context Enhancement Module (GCEM) and collaborative feature interaction is achieved through a Feature Interaction Module (FIM), leading to improved small object detection performance. Our main contributions are as follows:

(1) We propose DCGNet, which aims to enhance the detailed information of small objects and mitigate the impact of environmental noise, thereby significantly improving the detection performance for small objects.
(2) A new DCE-FPN is presented to enrich features with fine details in the frequency domain and to gather multi-scale contextual information through a multi-branch architecture, thereby enhancing small object representations and suppressing background noise.
(3) The EDI-RCNN module is introduced to enhance task-specific features and promote their interaction, thereby effectively reducing interference between tasks.
(4) The performance of our method is evaluated on three publicly available datasets of small objects, VisDrone2019, AI-TODv2, and DOTAv1.0, respectively.

This paper is an extension of the conference version [15], with the following significant improvements. First, we provide a comprehensive analysis of the critical challenges inherent in existing small object detection methods (Sections 1 and 2). Second, we present new extensions of several modules, including variants of GCEM and FIM, which are meticulously elaborated in Subsection 3.2. Third, extensive ablation experiments are conducted to evaluate the contribution of each module, including these aforementioned variants. In addition, the detection results are visualized in various complex scenarios, accompanied by comparative analyses with multiple state-of-the-art (SOTA) models (Section 4). These contributions further demonstrate the robustness of the proposed method in detecting small objects under challenging environmental conditions.

2. Related work

2.1. Object detection

Object detection is a fundamental task in the field of computer vision, aimed at identifying and localizing multiple objects within an image, with SOTA results given by deep learning-based methods, such as convolutional neural network (CNN)-based detectors and Transformer-based detectors, as discussed below.

CNN-based Detectors. CNN based detectors are conventionally categorized into one-stage and two-stage detectors. One-stage architectures, such as SSD [4], FCOS

[3], and the YOLO series including the recent YOLOv10 [16], predict object categories and locations through a unified and single-step pipeline. While their design prioritizes computational speed and efficiency for real-time applications, they typically entail a slight compromise in detection accuracy. Conversely, two-stage detectors, such as Faster R-CNN [1] and Cascade R-CNN [2], utilize a sequential pipeline that first generates candidate regions via a Region Proposal Network (RPN) before applying dedicated classification and regression heads. Motivated by their superior detection accuracy, we adopt the two-stage detector architecture as our primary baseline. Furthermore, to explicitly tackle the unique challenges of remote sensing, such as drastic scale variations and diverse background contexts, novel architectures have emerged. Notably, the Poly Kernel Inception Network (PKIN) [17] employs multi-scale convolution kernels without dilation and integrates a context anchor attention mechanism. This design enables highly effective multi-scale feature extraction and contextual information capture. Similarly, LTDNet [18] explores highly efficient lightweight structures by designing a dedicated backbone that redistributes computational resources to early stages alongside a detection head incorporating deformable convolutions. This tailored design effectively preserves the subtle features of tiny objects without introducing excessive computational overhead. Additionally, PointOBB [19] addresses the complex orientations and heavy annotation burden of aerial targets by proposing a single point supervision paradigm. It successfully generates pseudo oriented bounding boxes by leveraging geometric consistency across multiple augmented views and analyzing scale distributions, thereby achieving accurate detection without relying on costly dense annotations.

Transformer-based Detectors. Transformer-based detectors have formulate object detection as a direct set prediction problem. Pioneering models like DETR [5] eliminate the necessity for manually designed components, including anchor generation and non maximum suppression, by leveraging global bipartite matching. Building upon this foundational framework, several advanced models have emerged to optimize training efficiency and feature representation. Specifically, Deformable-DETR [20] drastically accelerates convergence and efficiently processes multi-scale features by applying sparse attention mechanisms that focus exclusively on a limited number of critical sampling points. DAB-DETR [21] enhances the interpretability of spatial queries by utilizing dynamic anchor boxes to supply explicit positional priors, which significantly improves localization accuracy. Furthermore, H-DETR [22] introduces a hybrid matching scheme that seamlessly integrates one to one and one to many assignments, thereby enhancing both training efficiency and feature discriminability. Concurrently, Align-DETR [23] refines the overall label assignment process by incorporating an intersection over union aware loss function to explicitly resolve the misalignment between classification scores and localization precision. Collectively, these advanced methodologies effectively resolve the inherent convergence and alignment limitations of early designs, establishing a highly competitive approach for modern visual detection tasks.

2.2. Small object detection

In contrast to general object detection, small object detection poses a significant challenge due to the scarcity of distinctive features inherent to small objects. Presently, the primary algorithms for small object detection encompass: sample-based methods, context-based methods, attention-based methods, and multi-scale feature fusion methods.

Sample-based methods. To improve the model's attention to small objects, a series of data augmentation techniques such as [24, 25, 26, 27] have improved detection performance by increasing the number of small object samples in images. Simultaneously, given the high sensitivity of small objects to positional deviations [28], traditional label assignment strategies based on Intersection over Union (IoU) often struggle to provide high-quality positive samples for small objects. Consequently, literatures such as [27, 28, 29, 30] proposed innovative label assignment strategies that effectively improve sample quality, thereby greatly enhancing the accuracy of small object detection. In our experiments, we adopted the Receptive Field based Label Assignment (RFLA) [30] as the label assignment strategy, which measures the similarity between the Gaussian receptive field and ground-truth boxes to improve sample quality.

Context-based methods. In images, small objects often occupy only a small area, making their features less distinct and more vulnerable to the corruption by background noise. To address this challenge, modeling contextual information can help to reveal the association between the targets and their surrounding environment [31], thereby maintaining a high detection accuracy even in complex backgrounds. Yuan et al. [32] designed a context-aware module that provides multiple observation perspectives and contextual information, enhancing the features of small objects. Cui et al. [33] [33] designed a context-aware block (CAB), which effectively captures contextual information at various scales through the use of pyramidal dilated convolutions. The design constructs high resolution feature maps that are semantically rich, significantly enhancing the performance of small object detection. Zhang et al. [34] proposed a method called Dynamic local and global Context Exploration (DCE), which offers local and global contextual information for small objects, effectively modeling the relationships between the objects themselves and between the objects and their environment. Although the aforementioned methods achieve effective modeling of contextual information, they are still limited in the fusion of multi-scale contextual features.

Attention-based methods. Motivated by human visual attention, the attention mechanism enables the model to focus on important areas of an image and ignore the less significant areas [31]. In the field of small object detection,

the attention mechanism aids in enhancing detection accuracy by enabling the model to concentrate on smaller and less salient targets within the image. Specifically, the spatial attention mechanism focuses on key areas of the image, emphasizing important spatial locations through the generation of a weight map, while disregarding less significant parts. However, the channel attention mechanism allows the model to automatically identify the importance of each channel and adjust the representation of the input data accordingly. The Convolutional Block Attention Module (CBAM) [35] integrated channel attention and spatial attention mechanisms, enabling a comprehensive understanding of image information and the enhancement of critical features while reducing the focus on less significant ones. To better capture features of small objects, Li et al. [36] developed a cross-level attention module that models the spatial relationships among pixels at different levels, thereby obtaining a more discriminative feature of small objects.

However, the locality of spatial features can restrict temporal attention methods from effectively distinguishing subtle differences between objects and their backgrounds [37], leaving small objects vulnerable to interference from complex scenes. In contrast, we explore the application of attention mechanisms in the frequency domain for small objects after wavelet transformation [14]. This approach enables more accurate identification of small objects and reduces noise within the signal, while preserving essential details and structural information.

Multi-scale feature fusion methods. In the multi-scale feature maps generated by CNNs [38, 39, 40, 41], the deep low-resolution feature maps are rich in semantic information but may omit detailed information of small objects, while shallow high-resolution feature maps retain more information about small objects but contain significant environmental noise. By effectively fusing features from different levels, the performance of small object detection can be enhanced.

FPN [9] enhances multi-scale object detection by propagating high-level semantics to shallow layers, but its reliance on spatial interpolation (e.g., bilinear upsampling) introduces aliasing that blurs fine details critical for small objects. The Path Aggregation Network (PANet) [42] strengthens information flow with an additional bottom-up path, and the Bi-directional FPN (BiFPN) [43] further improves fusion through learnable weights, accounting for the different contributions of various layers. However, both operate purely in the spatial domain and lack mechanisms to suppress background noise, leading to noise accumulation that can overwhelm weak small-object features. To address this, the authors of [44] proposed a lightweight Enhanced Interlayer Feature Correlation (EFC) strategy within FPN. Its Grouped Feature Focus (GFFs) unit reinforces cross-layer correlations, while the Multilevel Feature Reconstruction (MFR) module reduces redundancy by separating weak and strong features, better preserving small-object information.

Recently, frequency-domain transformations, particularly wavelet transforms, have been introduced to preserve fine-grained details in object detection. For instance, DWT-YOLO [45] and DI-YOLOv5 [46] utilize the DWT during the downsampling stage of the backbone to prevent information loss during feature extraction. Similarly, WT-DETR [47] integrates wavelets into a Transformer architecture to enhance global feature modeling. While effective, these approaches mainly target the backbone or global modeling stages, yet they fail to address feature blurring and noise amplification that arise during multi-scale feature fusion.

To overcome these limitations, we propose DCE-FPN to explicitly strengthen small-object details and reduce environmental noise during feature fusion. Unlike prior frequency-domain approaches that directly leverage wavelet sub-bands, DCE-FPN adopts a dedicated high-frequency denoising strategy. It incorporates the Sobel operator for spatial edge guidance and introduces a Detail Attention Module (DAM) tailored specifically to high-frequency components, effectively suppressing noise while preserving the sharp boundaries of small objects.

2.3. Detection head

General object detectors typically employ two parallel heads to separate the tasks of classification and localization. In many detectors [1, 4, 48], the detection head performs both classification and regression tasks on the same feature. However, this coupled design can lead to misalignment and information interference between classification and localization [11], further confusing the limited feature information of small objects. YOLOX [12] pioneered the design of a decoupled detection head, allowing for independent optimization of classification and localization tasks through distinct feature representations. This approach significantly improves the model's performance in both tasks. Subsequent versions, such as YOLOv6 [49] and YOLOv7 [50], have also adopted the decoupled design. DDOD [51] further improved the model's adaptability to spatial features by integrating deformable convolutions [52] into the decoupled detection head. Although the decoupled detection head increases the model's flexibility, the requirements for features in classification and localization tasks often differ.

Classification predominantly requires rich semantic contextual information, while localization focuses more on detailed information [11]. Therefore, a decoupled RCNN is proposed in TBNet [53] to provide rich texture information for the classification task and detailed boundary information for the regression task. However, the decoupled RCNN independently trained the classification and localization tasks without considering the collaborative interaction between them. In contrast, the proposed EDI-RCNN independently enhances the specific features for each task, thereby facilitating comprehensive feature interaction among them while effectively mitigating inter-task interference and the influence of environmental noise.

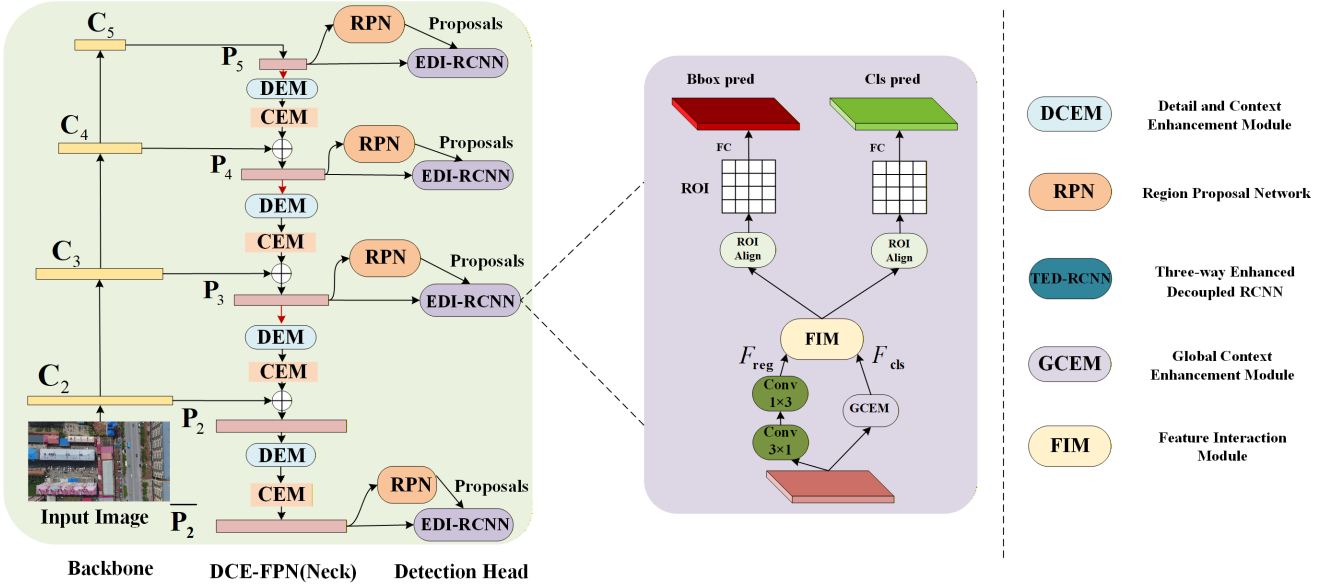


Figure 2: Schematic graph of the overview, which includes the ResNet50 as backbone network, the proposed DCE-FPN as neck, the Region Proposal Network (RPN) and the proposed EDI-RCNN.

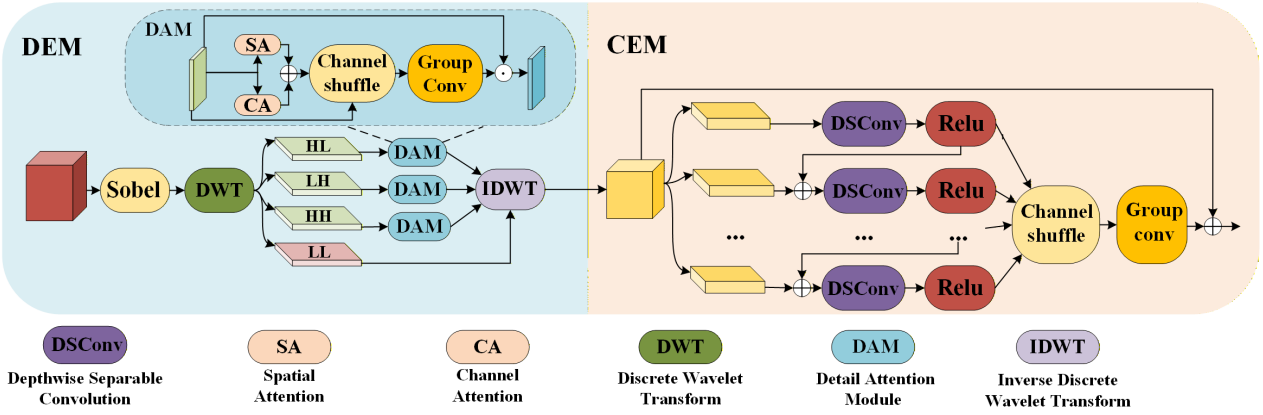


Figure 3: Schematic graph of the DEM and CEM.

3. Proposed method

Our network architecture is built upon the two-stage detector, with the overall process illustrated in Fig. 2. Initially, the backbone network (ResNet50 [41]) extracts the multi-scale features from the input image. Subsequently, the proposed DCE-FPN is used to fuse these multi-scale features and enhance the detailed information of small objects. Following this, the RPN and the proposed EDI-RCNN perform the detection tasks for the first and second stages, respectively. Detailed descriptions of the proposed DCE-FPN and EDI-RCNN are elaborated in Sections 3.1 and 3.2, respectively.

3.1. DCE-FPN

The feature information of small objects is extremely limited, making precise identification of them in complex environment very challenging. Extracting detailed information can improve the localization of small objects. Moreover, utilizing contextual information can establish a connection between small objects and their surrounding environment, which helps mitigate the interference of environmental noise, thereby enhancing the accuracy of detection. To this end, we designed the DCE-FPN, as shown in Fig. 2, to improve the detailed information of small objects between adjacent feature maps and to provide rich contextual information.

DCE-FPN is composed of the detail enhancement module (DEM) and the context enhancement module (CEM), as shown in Fig. 3. Specifically, to supplement fine-grained detailed information essential for small objects, we initially apply the Sobel operator to enhance edge information within the high-level feature maps F_{hl} that have undergone upsampling, prior to combining them with the low-level feature maps:

$$F'_{hl} = Sobel\left(F_{hl}\right) \tag{1}$$

where F'_{hl} and F_{hl} denote the high-level feature maps after and before the Sobel operator, respectively. Edge information facilitates a more accurate localization of small objects. Specifically, we employ a fixed 3×3 convolution kernel initialized with standard Sobel filters to extract gradient information from the feature maps.

Accurate localization of small objects depends strongly on clear edge details, which, from a signal processing viewpoint, correspond to high-frequency components due to their abrupt intensity variations. To better enhance these details and suppress noise, we convert features from the spatial domain to the frequency domain using DWT [14]. Unlike conventional spatial operations that may blur fine structures during fusion, DWT offers superior joint spatial–frequency localization. It explicitly decomposes the feature maps into one low-frequency approximation component F_{low} and three high-frequency detail components F_{high}:

$$\{F_{ll}, F_{lh}, F_{hl}, F_{hh}\} = DWT(F'_{hl}) \tag{2}$$

$$F_{low} = F_{ll}, F_{high} = \{F_{lh}, F_{hl}, F_{hh}\} \tag{3}$$

where F_{ll} represents the low-frequency sub-band containing the fundamental structural information of the image. Meanwhile, the high-frequency sub-bands F_{lh}, F_{hl}, and F_{hh} capture intricate edge information in the horizontal, vertical, and diagonal directions, respectively. This frequency-domain decomposition provides a key advantage: it enables DCE-FPN to explicitly separate and strengthen high-frequency components that encode fine details of small objects, while suppressing environmental noise. Such targeted enhancement effectively addresses the shortcomings of conventional spatial-domain operations. In practice, we employ the Haar wavelet as the basis function and apply a single-level decomposition. The implementations of the DWT and IDWT follow WaveCNet [14] to ensure stable and reliable feature processing.

Given that these three high-frequency components not only carry rich detailed information (edges and textures) but also contain a significant amount of environmental noise, we designed a Detail Attention Module (DAM) to perform noise reduction on these high-frequency components. It initially calculates the coarse attention map M through the element-wise addition of weight maps independently acquired from spatial attention (SA) [35] and channel attention (CA) [54] mechanisms:

$$M = SA\left(F_{high}\right) \oplus CA\left(F_{high}\right) \tag{4}$$

where $\oplus$ denotes the element-wise addition. Subsequently, the attention map M is thoroughly fused with the original high-frequency components F_{high} through channel shuffling and group convolution, resulting in a refined attention map M':

$$M' = GConv\left(CS\left(M, F_{high}\right)\right) \tag{5}$$

where $GConv$ denotes the group convolution and CS denotes the channel shuffling. The attention map M' is then utilized to modulate the high-frequency component F_{high} to yield F'_{high}, which enhances the detailed information and suppresses environmental noise:

$$F'_{high} = M' \odot F_{high} \tag{6}$$

where $\odot$ denotes the element-wise multiplication. Ultimately, the low-frequency component F_{low} and the three refined high-frequency components F'_{high} undergo the IDWT [14] to yield the temporal domain feature F_{de} with enhanced detailed information:

$$F_{de} = IDWT\left(F_{low}, F'_{high}\right) \tag{7}$$

To mitigate the impact of environmental noise on small objects, we propose to enhance the identification of small objects by capturing contextual information at various scales. Specifically, we first split the feature map F_{de} along the channel dimension into k equal parts F^i_{de} ($i \in (1, ..., k)$). Then, each F^i_{de} will undergo a branch with a depthwise separable convolution [55] with different kernel sizes $(3, 5, ...)$ followed by a ReLU activation function to capture contextual information at different scales. It should be noted that, to fully integrate contextual information on different scales, the output of the previous branch serves as part of the input for the next branch. Moreover, to further fuse multi-scale contextual information, we employ channel shuffling and group convolution to integrate the contextual information from the outputs of the k branches, rather than simply adding them together:

$$F^i_{context} = \begin{cases} ReLU\left(DSConv\left(F^i_{de}\right)\right) & , i = 1 \\ ReLU\left(DSConv\left(F^{i-1}_{context} + F^i_{de}\right)\right) & , i > 1 \end{cases} \tag{8}$$

$$F_{dce} = GConv\left(CS\left(F^i_{context}\right)\right) \tag{9}$$

where $F^i_{context}$ denotes the contextual information from the i-th branch and F_{dce} denotes the final feature map with enhanced detailed information and contextual information.

It is noteworthy that, although feature map P_2 (i.e., the feature of the second layer in FPN) is not involved in the upsampling fusion process, we still utilize DEM and CEM to refine its detailed and contextual information and obtain $\overline{P}_2$. Then, the $\overline{P}_2$ is directly employed for the subsequent detection tasks. This is because P_2 inherently possesses a mount of detailed information and environmental noise. Once the refined contextual information for small objects has been acquired, the top-down propagation mechanism within the FPN effectively disseminates this information across all levels, substantially improving the detection accuracy of small objects.

3.2. EDI-RCNN

An object detection task is generally divided into localizing the precise location of the targets and identifying its category where the classification task relies more on global contextual information while the localization task focuses more on detailed information. Consequently, providing the specific information required for each task can better enhance the detection accuracy. Moreover, the common coupled detection head can further blur the already weak features of small objects. To this end, we design the EDI-RCNN, as shown in the middle part of Fig. 2.

As illustrated in Fig. 2, given that DCE-FPN has already enhanced the detailed information, in the branch for bounding box regression, F_{dce} further refines the edge details of the targets by employing a 3×1 convolutional layer and a 1×3 convolutional layer, resulting in the regression feature map F_{reg}. For the classification branch, F_{dce} employs the proposed Global Context Enhancement Module (GCEM) to enhance the global contextual information, thereby obtaining the classification feature F_{cls}:

$$F_{reg} = Conv_{1\times3}(Conv_{3\times1}(F_{dce})) \tag{10}$$

$$F_{cls} = GCEM\left(F_{dce}\right) \tag{11}$$

These asymmetric convolutions facilitate the decomposition of orthogonal features. Unlike standard 3×3 convolutions that mix horizontal and vertical information and blur localization cues, the 3×1 and 1×3 kernels spatially decouple feature extraction. By explicitly capturing directional gradients along independent axes, this design prevents information mixing and enables precise refinement of object boundaries.

The proposed GCEM module is inspired by the work in [56], but differs in that it integrates global contextual information from both spatial and channel dimensions, rather than focusing solely on the spatial dimension. As illustrated in Fig. 4, we propose two implementation schemes for GCEM.

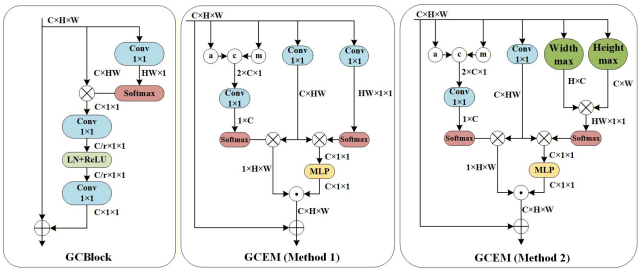


Figure 4: Schematic graph of the GCBlock and GCEM.

Method 1: To construct global contextual information across the channel dimension, the input feature $F_{dce} \in \mathbb{R}^{C\times H\times W}$ is first processed by global max pooling and global average pooling, followed by a 1×1 convolutional layer, and then the Softmax function to obtain the channel score $F_{cs} \in \mathbb{R}^{C\times1\times1}$ (named QK [57]):

$$F_{cs} = Softmax\left(Conv\left(\left[GAP\left(F_{dce}\right), GMP\left(F_{dce}\right)\right]\right)\right) \tag{12}$$

where $Softmax$ denotes the Softmax operation, GAP denotes the global average pooling, GMP denotes the global map pooling, [,] denotes the concatenation, and $Conv$ denotes the 1×1 Convolution. Next, the channel score F_{cs} is matrix-multiplied by the feature $F'_{dce} \in \mathbb{R}^{C\times HW}$, which is obtained through a 1×1 convolution on F_{dce} (named value [57]), to yield the global contextual information across the channel dimension, denoted as $F_c \in \mathbb{R}^{1\times H\times W}$:

$$F'_{dce} = Conv\left(F_{dce}\right) \tag{13}$$

$$F_c = F_{cs} \times F'_{dce} \tag{14}$$

To construct global contextual information cross the spatial dimension, Method 1 adopts the design concept of GCBlock, utilizing a 1×1 convolutional layer on the GCEM's input feature F_{dce} to generate the spatial score $F_{ss} \in \mathbb{R}^{HW\times1\times1}$ (named QK [57]). Finally, the global contextual information across spatial dimension, $F_s \in \mathbb{R}^{C\times1\times1}$, is obtained by multiplying F_{ss} with F'_{dce}:

$$F_{ss} = Softmax(Conv_{1\times1}(F_{dce})) \tag{15}$$

$$F_s = F_{ss} \times F'_{dce} \tag{16}$$

where $Conv_{1\times1}$ denotes the 1×1 convolutional layer.

Method 2: For the global contextual information across the channel dimension, we employ the same approach as Method 1. For the global contextual information cross the spatial dimension, we adopt a strategy that simultaneously operates on both the width and height dimensions. Specifically, max pooling is performed separately on F_{dce} along the width and height directions to extract the global information from these two directions. Subsequently, the information from these two directions is combined through matrix multiplication to generate the spatial score $F_{ss} \in \mathbb{R}^{HW\times1\times1}$ (named QK [57]):

$$F_{ss} = Softmax\left(M_w\left(F_{dce}\right) \times M_h\left(F_{dce}\right)\right) \tag{17}$$

where M_w and M_h denote max pooling operation along the width and height directions, respectively. Likewise, the global contextual information across spatial dimension, $F_s \in \mathbb{R}^{C\times1\times1}$, is obtained by multiplying F_{ss} with F'_{dce}:

$$F_s = F_{ss} \times F'_{dce} \tag{18}$$

After acquiring the global contextual information across channel and spatial dimensions, we can then enhance the F_{dce} to obtain the feature F_{cls} for classification. Consequently, Eq. (11) can be represented as:

$$F_{cls} = F_c \odot F_s + F_{dce} \tag{19}$$

By constructing global contextual information, the connection between the targets and the global information can be established, thereby enhancing the distinguishability of the targets and further reducing the interference of environmental noise.

After obtaining the features favored by each task, we aim for mutual interaction among tasks rather than their independent execution. Consequently, we designed the Feature Interaction Module (FIM). As illustrated in Fig. 5, we propose multiple implementation schemes for FIM.

Method 1: The FIM concatenates F_{reg} and F_{cls} along the channel dimension and then applies GMP and GAP to obtain $F_{cm} \in \mathbb{R}^{2C\times1\times1}$ and $F_{ca} \in \mathbb{R}^{2C\times1\times1}$, respectively. Subsequently, the concatenated feature of F_{cm} and F_{ca} are processed through a 1×1 convolution layer followed by a sigmoid function to generate the channel attention maps $W_{reg} \in \mathbb{R}^{C\times1\times1}$ and $W_{cls} \in \mathbb{R}^{C\times1\times1}$ for each task:

$$F_{cm} = GMP\left(\left[F_{reg}, F_{cls}\right]\right) \tag{20}$$

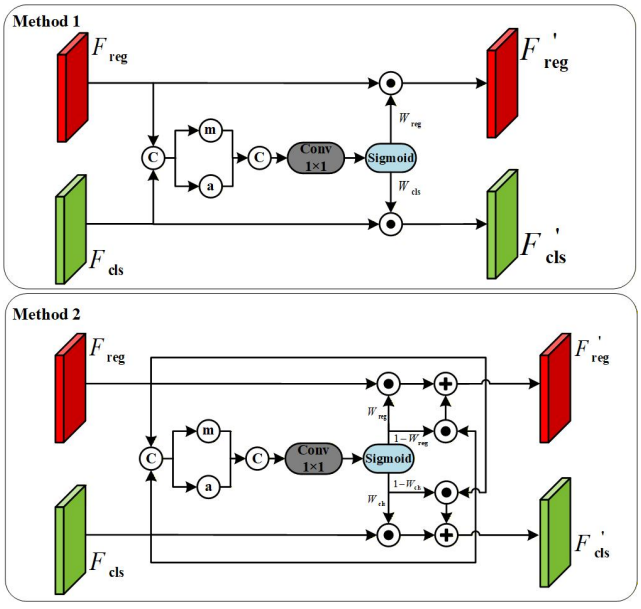


Figure 5: Schematic graph of the FIM.

$$F_{ca} = GAP\left(\left[F_{reg}, F_{cls}\right]\right) \tag{21}$$

$$W_{reg}, W_{cls} = \sigma\left(Conv\left(\left[F_{cm}, F_{ca}\right]\right)\right) \tag{22}$$

where σ denotes the sigmoid function. Ultimately, these attention maps are utilized to weight the corresponding task features, resulting in the final task-specific features F'_{reg} and F'_{cls}:

$$F'_{reg} = W_{reg} \odot F_{reg} \tag{23}$$

$$F'_{cls} = W_{cls} \odot F_{cls} \tag{24}$$

These features are specifically employed for regression and classification, respectively.

Method 2: Compared with Method 1, Method 2 further enhances the interaction and fusion between tasks by more effectively leveraging the complementarity of their features. Specifically, the classification features and regression features are mutually integrated through adaptive weights, thereby achieving complementary advantages between them:

$$F'_{reg} = W_{reg} \odot F_{reg} + \left(1 - W_{reg}\right) \odot F_{cls} \tag{25}$$

$$F'_{cls} = W_{cls} \odot F_{cls} + \left(1 - W_{cls}\right) \odot F_{reg} \tag{26}$$

where "−" represents element-wise subtraction.

4. Experiments and results

4.1. Datasets

We validated the effectiveness of our method on the VisDrone2019, AI-TODv2, and DOTAv1.0 datasets.

VisDrone2019 is a large-scale drone visual dataset that encompasses a multitude of small objects, comprising ten distinct categories of objects. It consists of 10209 high-resolution (2000×1500 pixels) aerial images, in which 6,471 images are designated for training, 3,190 images for testing, and 548 images for the validation. Due to the unavailability of the VisDrone2019 test set (it is not public), we adopt the approach of previous studies [44, 58], utilizing 6471 images for training and 548 images for testing in our experiments.

AI-TODv2 contains a larger number of tiny objects, with an average pixel size of approximately 12.8 [29], significantly smaller than those in other datasets. This dataset spans eight different categories of objects and consists of 11,214 images for the training set, 2,804 images for the validation set, and 14,018 images for the test set. Following the studies [29, 30, 28], we utilize the trainval (training and validation) set for model training and employ the test set to evaluate performance.

DOTAv1.0 is a large-scale dataset specifically designed for object detection in aerial images, comprising 2,806 high-resolution images with sizes ranging from 800 × 800 to 4,000 × 4,000 pixels. It encompasses 15 common object categories and contains 188,282 finely annotated instances. Characterized by complex backgrounds, severe scale variations, and a massive number of densely packed small objects, DOTAv1.0 poses a significant challenge for detection algorithms, making it highly suitable for evaluating model performance in realistic and complex environments.

4.2. Evaluation Metrics

To comprehensively evaluate the effectiveness of our method, we utilize the Average Precision (AP) and the number of parameters (Params) as evaluation metrics. The relevant formula for calculating AP is as follows:

$$AP = \int_0^1 P(R)\, dR \tag{27}$$

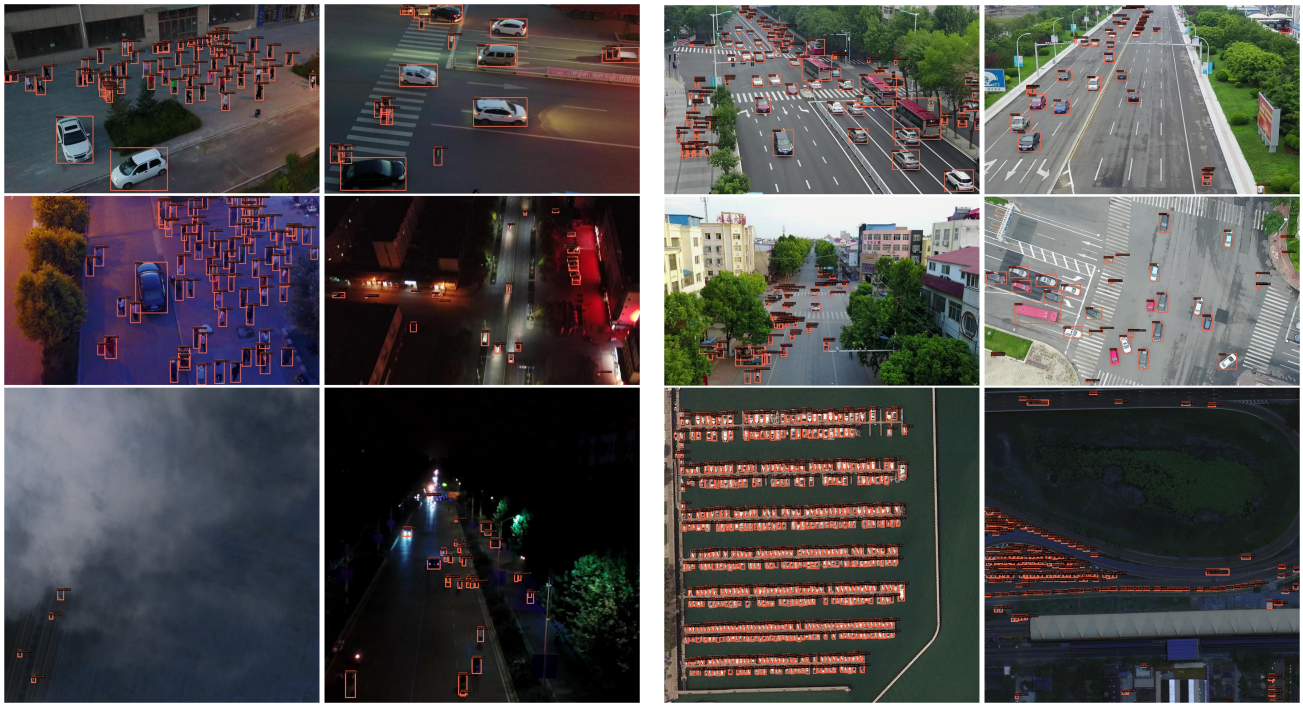
where AP reflects the integral result of Precision (P) and Recall (R) within the confidence threshold range from 0 to 1. Additionally, the calculation formulas for P and R are as follows:

$$P = \frac{TP}{TP + FP} \tag{28}$$

$$R = \frac{TP}{TP + FN} \tag{29}$$

where TP, FP, and FN stands for true positives, false positives, and false negatives, respectively. Additionally, the quantity of FP can be considered as the number of false detection, while the quantity of FN can be seen as the number of missing detection.

For the evaluation on the VisDrone2019 and DOTAv1.0 datasets, we follow the MSCOCO benchmark [59] and use metrics such as AP, AP_{50}, AP_{75}, AP_S, AP_M, and AP_L. AP_{50} and AP_{75} represent the AP of all objects at intersection of union (IOU) thresholds (determining TP) of 0.5 and 0.75, respectively. The AP in here is the average value from AP_{50} to AP_{95}, with an IOU interval of 0.05. Additionally, AP_S, AP_M, and AP_L represent the AP for small, medium, and



(a) (b)

Figure 6: The detection results of DCGNet. (a) detection under the complex environment like low light and fog. (b) detection under the dense scenario. It is evident that our method can effectively detect small objects even under different scenarios.

large objects, respectively. For the AI-TODv2 dataset, we follow the AI-TOD benchmark [60], which additionally uses the AP_t and AP_{vt} evaluation metrics for tiny and very tiny objects, respectively.

In addition to the accuracy metrics, we introduce computational complexity, measured in Floating Point Operations (FLOPs), and inference speed, measured in Frames Per Second (FPS), to comprehensively evaluate the efficiency and practical applicability of the proposed model.

4.3. Implementation Details

All our experiments are based on the MMDetection toolbox [61] and run on the RTX3090ti. During training on the VisDrone2019, AITODv2, and DOTAv1.0 datasets, we utilized a stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 0.001, with a batch size of 2, and conducted a total of 12 epochs of training. The learning rate for VisDrone2019 and DOTAv1.0 was set to 0.01, while for AI-TODv2 it was 0.005, with decaying at the 8th and 11th epochs. During the training process, the Gaussian receptive field based label assignment (RFLA) [30] strategy was employed, which assesses the similarity between the Gaussian receptive field and the ground-truth boxes, effectively enhancing samples quality. In addition, during the training and testing phase, we set the input size as 1333×800, 800×800, and 1024×1024 for VisDrone2019, AI-TODv2, and DOTAv1.0 in the same manner as [27, 28, 44, 30].

4.4. Comparisons With the SOTA Methods

To demonstrate the superiority of our method, we conducted comparisons with numerous state-of-the-art methods on the VisDrone2019, AI-TODv2 and NWPU VHR-10 datasets. These methods include widely-used general object detectors, including Faster R-CNN [1], Cascade R-CNN [2], FCOS [3], DetectoRS [62], ATSS [63], Align-DETR [23], Deformable-DETR [20], H-DETR [22], YOLOv10n [16], and the latest models specifically designed for small object detection, including EFC [44], CEASC [64], NWD [28], FSANet [65], NWD-RKA [66], RFLA [30], SR-TOD [58], ORFENet [67], CFPT [68], LTDNet [18], PFIM [69], and DAB-DETR [21]. It is noteworthy that there are some discrepancies among the comparative models across different datasets. For certain models, we directly cited their reported results on the datasets that have been published. For others, we reimplemented them and obtained results on datasets that had not been reported previously. In these comparisons, we used Cascade R-CNN as our baseline and employed ResNet-50 [41] as the feature extractor. In addition, during the training process, we applied the RFLA [30] as our label assignment strategy. Fig. 6 demonstrates the detection results of our method under different scenarios. Even under complex environmental condition such as low lightning and fog, the proposed DCGNet is still able to accurately identify and detect small objects, which thoroughly verifies that our method effectively mitigates the impact of environmental noise.

Results on VisDrone2019: Table 1 presents the comparative results of our method against other methods on the VisDrone2019. Compared to the main baseline (Cascade

Table 1
Comparisons of our proposed method with other SOTA methods on the VisDrone2019 dataset. All models are trained on the VisDrone2019 training set and tested on the validation set. In the results, the best performance is denoted by red font, while the second-best performance is indicated by blue font. The underlined value represents the optimal result of the baselines. "FIM1+GCEM1" indicates that EDI-RCNN employs both Method 1 of FIM and Method 1 of GCEM, while "FIM1+GCEM2", "FIM2+GCEM1", and "FIM2+GCEM2" correspond to other combinations of these design methods.

Category	Model	Venue	Backbone	Input Size	AP (%)	AP_{50} (%)	AP_{75} (%)	AP_S (%)	AP_M (%)	Params (M)
Trans-based	Align-DETR [23]	Arxiv2023	ResNet-50	1333×800	30.1	51.8	30.1	21.6	40.3	47.50
	Deformable-DETR [20]	ICLR2021	ResNet-50	1333×800	14.3	26.9	13.7	7.8	22.0	40.12
	H-DETR [22]	CVPR2023	ResNet-50	1333×800	27.4	47.5	26.8	19.5	37.4	47.46
	DAB-DETR [21]	ICLR2022	ResNet-50	1333×800	27.5	47.7	26.9	19.8	37.6	-
CNN-based	Faster R-CNN [1]	CVPR2016	ResNet-50	1333×800	24.8	43.1	24.9	15.8	36.5	41.21
	Cascade R-CNN [2]	CVPR2018	ResNet-50	1333×800	26.2	43.6	27.0	16.8	38.3	68.97
	FCOS [3]	ICCV2019	ResNet-50	1333×800	18.0	32.8	17.6	10.5	27.4	31.86
	DetectoRS [62]	CVPR2021	ResNet-50	1333×800	28.3	46.9	29.0	18.9	40.4	123.23
	ATSS [63]	CVPR2020	ResNet-50	1333×800	24.0	40.2	24.8	14.3	35.7	32.13
	EFC† [44]	TGRS2024	ResNet-18	1333×800	30.1	52.1	29.8	-	-	39.98
	FSANet [65]	TGRS2022	ResNet-50	1333×800	28.4	49.7	28.2	20.1	39.0	-
	CEASC† [64]	CVPR2023	ResNet-50	1333×800	28.7	50.7	28.4	-	-	-
	ORFENet [67]	TGRS2024	ResNet-50	1333×800	28.3	50.4	28.1	20.5	38.7	32.77
	Cascade R-CNN w/NWD [28]	ISPR2022	ResNet-50	1333×800	28.9	49.3	29.6	20.6	40.4	68.97
	DetectoRS w/NWD-RKA [66]	ISPR2022	ResNet-50	1333×800	29.6	51.4	29.3	20.9	40.3	123.23
	Cascade R-CNN w/RFLA [30]	ECCV2022	ResNet-50	1333×800	28.4	49.1	28.6	20.4	38.2	68.97
	DetectoRS w/SR-TOD [58]	ECCV2024	ResNet-50	1333×800	28.6	48.5	29.3	20.0	39.5	-
	DetectoRS w/SR-TOD† [58]	ECCV2024	ResNet-50	1280×1280	27.2	47.1	27.2	-	-	-
	YOLOv10n [16]	NeurIPS2024	Modified CSPNet	1333×800	29.4	50.5	29.1	21.1	39.8	2.70
	CFPT† [68]	TGRS2025	ResNet-101	1333×800	29.7	50.0	30.4	19.7	41.9	51.3
	LTDNet [18]	TGRS2025	RepViT-TD	1333×800	29.1	50.2	28.9	20.8	39.6	4.85
	RFLA w/PFIM† [69]	CVPR2025	ResNet-50	1333×800	29.0	50.7	29.0	-	-	-
	Ours (FIM1 + GCEM1)	-	ResNet-50	1333×800	30.5	52.2	30.7	22.9	40.0	79.30
	Ours* [15] (FIM1 + GCEM2)	-	ResNet-50	1333×800	30.7	52.6	31.0	23.2	40.4	79.16
	Ours (FIM2 + GCEM1)	-	ResNet-50	1333×800	30.8	52.8	31.2	23.4	40.6	79.81
	Ours (FIM2 + GCEM2)	-	ResNet-50	1333×800	31.0	53.2	31.5	23.8	40.9	79.98

The Trans-based and CNN-based indicate that their corresponding categories are Transfomer-based models and CNN-based models, respectively.
PS: No AP_t results released on this dataset.
The '†' signifies that the results are referenced from the original paper.
The '-' denotes no result reported. The 'w/' indicates that the corresponding baseline applies the respective method.
The 'w/' indicates that the corresponding baseline applies the respective method.

R-CNN w/RFLA [30]), our method demonstrates superior improvement across all metrics. Specifically, the AP has increased by 2.6% (from 28.4% to 31.0%), AP_{50} has improved by 4.1% (from 49.1% to 53.2%), AP_{75} has risen by 2.9% (from 28.6% to 31.5%), AP_S has improved by 3.4% (from 20.4% to 23.8%), and AP_M has increased by 2.7% (from 38.2% to 40.9%). Additionally, we compared the performance of DCGNet under different design methods of GCEM and FIM. The results demonstrate that the optimal performance (AP=31.0) is achieved when both Method 2 of GCEM and Method 2 of FIM are applied. Compared with other SOTA methods for small object detection, our DCGNet achieves the best results in terms of AP, AP_{50}, AP_{75}, and AP_S. Especially, our DCGNet outperforms the latest small objet detection model CFPT [68] (AP_{50}=50.0%, AP_S=19.7%) by 3.2% in AP_{50} and 4.1% in AP_S, respectively. Furthermore, our method outperforms the second-best model Align-DETR [23] (AP_{75}=30.1, AP_S=21.6) by 1.4% and 2.1% in terms of AP_{75} and AP_S, respectively. Those superior results indicate that our method is particularly effective in detecting small objects.

Results on AI-TODv2: To further validate the effectiveness of our method, we performed additional comparative experiments on the AI-TODv2 dataset. As shown in Table 2, our method achieved the best results across multiple metrics, specifically reaching 25.6%, 58.4%, 18.5%, 25.2%, and 30.6% for AP, AP_{50}, AP_{75}, AP_t, and AP_S, respectively. Moreover, it can be observed that the model achieves better performance when employing GCEM's Method 2 and FIM's Method 2 (AP_{75}=18.5%). Compared to the Cascade R-CNN w/RFLA [30] baseline, our method achieves consistent improvements of 1.2% and 1.4% in AP_{vt} and AP_M, respectively. However, it does not reach state-of-the-art performance on these metrics. This limitation may be attributed to the incorporation of multi-scale contextual information, which tends to dilute the feature representations of tiny and medium-sized objects. Furthermore, compared to recently designed models specifically designed for tiny object detection, our method outperforms the second-best model NWD-RKA [66] (AP_{75}=17.1%, AP_t=24.2%) by 1.4% and 1.0% in terms of AP_{75} and AP_t, respectively.

Results on DOTAv1.0: To comprehensively assess the generalization and transferability of our method across different detection frameworks, we further evaluated it on the DOTAv1.0 dataset. Our modules were incorporated into a representative two-stage detector, Cascade R-CNN, and a typical one-stage detector, FCOS. As reported in Table 3, DCGNet consistently delivers notable performance improvements on both architectures. For Cascade R-CNN, our method raises the overall AP by 1.0% and AP_{50} by 1.2%, with a significant 1.6% gain in small-object detection performance (AP_s). When integrated into FCOS, the improvements are even more substantial, yielding a 1.8% increase in AP, a 2.8% gain in AP_{50}, and a 1.6% boost in AP_s. These consistent gains across both two-stage and one-stage detectors demonstrate the robustness and wide applicability of our feature enhancement and interaction mechanisms in complex remote sensing scenarios.

Table 2
Comparisons of our proposed method with other SOTA methods on the AI-TODv2 dataset. All models are trained on the AI-TODv2 trainval set and tested on the test set. The underlined value represents the optimal result of the baselines. "FIM1+GCEM1" indicates that EDI-RCNN employs both Method 1 of FIM and Method 1 of GCEM, while "FIM1+GCEM2", "FIM2+GCEM1", and "FIM2+GCEM2" correspond to other combinations of these design methods.

Category	Model	Venue	Backbone	AP (%)	AP_{50} (%)	AP_{75} (%)	AP_{vt} (%)	AP_t (%)	AP_S (%)	AP_M (%)
Trans-based	Align-DETR [23]	arxiv2023	ResNet-50	24.6	56.1	18.0	8.7	24.6	29.6	37.4
	Deformable-DETR [20]	ICLR2023	ResNet-50	17.0	45.9	8.8	7.2	17.1	22.7	28.2
	H-DETR [22]	CVPR2023	ResNet-50	23.6	54.0	16.7	8.5	23.2	30.1	37.8
	DAB-DETR [21]	ICLR2022	ResNet-50	16.5	42.6	9.9	7.9	15.2	23.8	31.9
CNN-based	Faster R-CNN [1]	CVPR2016	ResNet-50	9.3	21.2	6.9	0.0	4.5	21.0	35.0
	Cascade R-CNN [2]	CVPR2018	ResNet-50	11.3	25.4	8.5	0.0	6.3	23.2	37.0
	FCOS [3]	ICCV2019	ResNet-50	12.0	30.2	7.3	2.2	11.1	.16.6	26.9
	DetectoRS [62]	CVPR2021	ResNet-50	16.2	35.5	12.9	1.0	12.5	28.5	40.1
	ATSS [63]	CVPR2020	ResNet-50	12.8	30.6	8.5	1.9	11.6	19.5	29.2
	EFC [44]	TGRS2024	ResNet-18	24.3	55.6	17.1	7.6	23.9	28.9	39.0
	FSANet [65]	TGRS2022	ResNet-50	22.0	52.7	14.9	6.7	20.7	27.8	36.1
	CEASC [64]	CVPR2023	ResNet-50	20.7	51.7	14.4	5.1	20.1	26.8	35.1
	ORFENet† [67]	TGRS2024	ResNet-50	18.9	44.4	12.7	6.9	18.4	23.4	30.3
	DetectoRS w/NWD [28]	ISPR2022	ResNet-50	22.5	52.8	15.8	5.4	21.8	29.5	39.7
	DetectoRS w/NWD-RKA† [66]	ISPR2022	ResNet-50	24.7	57.4	17.1	9.7	24.2	29.8	39.3
	Cascade w/RFLA [30]	ECCV2022	ResNet-50	23.4	55.2	15.8	7.8	22.6	28.9	38.0
	DetectoRS w/SR-TOD [58]	ECCV2024	ResNet-50	24.4	56.0	17.7	8.4	24.3	29.4	39.4
	YOLOv10n [16]	NeurIPS2024	Modified CSPNet	23.5	55.2	16.1	8.9	24.2	27.4	33.7
	CFPT [68]	TGRS2025	ResNet-101	23.9	56.5	15.9	9.7	23.7	29.6	38.1
	LTDNet† [18]	TGRS2025	RepViT-TD	23.0	54.6	15.5	8.9	23.6	27.2	33.1
	RFLA w/PFIM† [69]	CVPR2025	ResNet-50	23.9	55.8	16.6	7.3	23.5	29.0	-
	Ours (FIM1 + GCEM1)	-	ResNet-50	25.0	57.6	17.7	8.1	24.3	29.7	38.8
	Ours* [15] (FIM1 + GCEM2)	-	ResNet-50	25.5	58.2	18.3	8.8	25.0	30.4	39.2
	Ours (FIM2 + GCEM1)	-	ResNet-50	25.4	58.2	18.1	8.9	24.8	30.3	39.0
	Ours (FIM2 + GCEM2)	-	ResNet-50	25.6	58.4	18.5	9.0	25.2	30.6	39.4

The † signifies that the results are referenced from the original paper.
The 'w/' indicates that the corresponding baseline applies the respective method.

Table 3
Comparisons of our method with different baselines on the DOTAv1.0 test set.

Model	AP (%)	AP_{50} (%)	AP_{75} (%)	AP_s (%)	AP_m (%)	Params (M)	FLOPs (G)	FPS
Cascade w/RFLA (baseline)	43.2	70.1	49.7	24.2	44.6	68.97	239.13	13.1
DCGNet (ours)	44.2(+1.0)	71.3(+1.2)	51.4(+1.7)	25.8(+1.6)	45.5(+0.9)	79.98	342.19	9.8
FCOS w/RFLA (baseline)	33.6	59.9	37.7	19.3	37.7	31.93	557.11	15.4
DCGNet (ours)	35.4(+1.8)	62.7(+2.8)	39.9(+2.2)	20.9(+1.6)	38.5(+0.8)	38.43	664.78	11.8

Table 4
Ablation study of DCE-FPN and EDI-RCNN on the VisDrone2019 validation set. The GCEM design in EDI-RCNN adopts the proposed Method 2, as illustrated in Fig. 4, while the FIM design in EDI-RCNN adopts the proposed Method 2, as illustrated in Fig. 5.

DCE-FPN	EDI-RCNN	AP (%)	AP_{50} (%)	AP_{75} (%)	Params (M)	FLOPs (G)	FPS
×	×	28.4	49.1	28.6	68.97	239.13	13.1
×	✓	29.5	50.7	29.8	72.64	264.75	12.4
✓	×	30.1	51.8	30.4	76.31	316.57	10.7
✓	✓	31.0	53.2	31.5	79.98	342.19	9.8

Table 5
Ablation study of DCE-FPN and EDI-RCNN on the AI-TODv2 test set. The GCEM design in EDI-RCNN adopts the proposed Method 2, as illustrated in Fig. 4, while the FIM design in EDI-RCNN adopts the proposed Method 2, as illustrated in Fig. 5.

DCE-FPN	EDI-RCNN	AP (%)	AP_{50} (%)	AP_{75} (%)	Params (M)	FLOPs (G)	FPS
×	×	23.4	55.2	15.8	68.97	162.21	15.6
×	✓	24.5	56.5	17.0	72.64	181.65	14.6
✓	×	24.7	57.2	17.2	76.31	219.86	12.7
✓	✓	25.6	58.4	18.5	79.98	239.3	11.4

4.5. Ablation Study

1) Effectiveness of each component on VisDrone2019: To validate the effectiveness of the proposed DCE-FPN and EDI-RCNN components, we conducted ablation experiments on the VisDrone2019 validation set. Specifically, we replaced the original FPN with DCE-FPN and the coupled RCNN with EDI-RCNN. As shown in Table 4, replacing the original FPN with the proposed DCE-FPN yields significant improvements in AP, AP_{50}, and AP_{75}, with AP_{50} achieving a notable increase of 2.7%. This substantial gain in localization precision is primarily driven by our frequency-domain enhancement and context aggregation mechanisms. Specifically, the DEM transforms features into the frequency domain to explicitly preserve the high-frequency structural edges of small objects while suppressing the noise typically amplified by traditional upsampling. Simultaneously, the CEM incorporates diverse receptive field information, which facilitates the distinction of small objects from complex backgrounds.

Furthermore, substituting the original coupled RCNN with the EDI-RCNN increases the AP by 1.1%, AP_{50} by 1.6%, and AP_{75} by 1.2%. This performance gain effectively addresses the extreme feature scarcity of small objects, a characteristic that makes them highly susceptible to task interference in a conventional coupled head. By independently enriching the global context for classification through the GCEM and promoting synergistic task interaction via the FIM, the EDI-RCNN ensures that semantic cues assist the regression branch in filtering out background clutter. In contrast, the boundary cues guide the classifier to focus on the core regions of the targets. This design explicitly mitigates the misalignment between classification and localization.

The simultaneous integration of both DCE-FPN and EDI-RCNN yields an overall performance that significantly surpasses the use of either component alone, boosting the

AP by 2.6%, AP_{50} by 4.1%, and AP_{75} by 2.9%. This demonstrates a strong complementary synergy between the two modules. Specifically, the DCE-FPN provides high-quality, noise-suppressed, and detail-rich feature representations at the neck stage. Subsequently, the EDI-RCNN effectively leverages these refined features in the detection head by balancing and facilitating interaction between the tasks. Ultimately, this integrated approach effectively resolves the fundamental challenges of limited information and noise vulnerability in small object detection.

2) Effectiveness of each component on AI-TODv2: To further validate the effectiveness and robustness of our method, we also conducted ablation experiments on the AI-TODv2 test set, with results shown in Table 5. After individually adding DCE-FPN and EDI-RCNN, compared to the baseline, three metrics exhibited significant improvements, especially with AP_{50} increased by 2.0% and 1.3%, respectively. When DCE-FPN and EDI-RCNN were used in combination, all three metrics reached their optimal values, with AP increased by 2.2%, AP_{50} by 3.3%, and AP_{75} by 2.6%. These results further confirm the effectiveness of our method in improving the detection of small objects.

Table 6
Ablation study of detail enhancement and context enhancement on the VisDrone2019 validation set. **w/DE** and **w/CE** denote the detail enhancement and context enhancement, respectively.

w/DE	w/CE	$AP(\%)$	$AP_{50}(\%)$	$AP_{75}(\%)$
×	×	28.4	49.1	28.6
×	✓	28.9	49.9	29.2
✓	×	29.5	51.0	29.7
✓	✓	30.1	51.8	30.4

3) Ablation study on DCE-FPN: To further validate the effectiveness of the proposed DCE-FPN on small object detection by providing detailed and contextual information, we conducted a detailed experimental analysis for the effectiveness of the DEM and CEM modules on the VisDrone2019 validation set. As shown in Table 6, the results clearly indicate that after enhancing the detailed information of small objects, AP is increased by 1.1%, AP_{50} by 1.9%, and AP_{75} by 1.1%. When the contextual information was enhanced, AP is increased by 0.5%, AP_{50} by 0.8%, and AP_{75} by 0.6%. After combining both DEM and CEM, the model performance reached the best, with AP reaching 30.1 (an increase of 6.0%), AP_{50} reaching 51.8% (an increase of 5.5%), and AP_{75} reaching 30.4% (an increase of 6.3%). These results further confirm that providing rich detailed and contextual information for small objects can effectively enhance the model's ability to recognize small objects in complex environments.

To further validate the effectiveness and superiority of the frequency-domain detail enhancement strategy utilized in DCE-FPN, we conducted a quantitative comparative analysis against pure spatial-domain enhancement methods on the VisDrone2019 dataset. The experiment comprises four comparative settings, including the DCE-FPN without detail enhancement (w/o DE), a scheme introducing only the Sobel operator, a scheme combining the Sobel operator with a Spatial Attention (SA) mechanism, and our proposed frequency-domain enhancement strategy. The experimental results are presented in Table 7. Compared to the DCE-FPN without detail enhancement, simply introducing the Sobel operator to extract spatial edges yields a tangible performance improvement (increasing AP from 28.9% to 29.3%). This indicates that edge information positively contributes to the feature representation of small objects. Building upon this, incorporating the Spatial Attention mechanism (Sobel + SA) provides only a marginal performance gain (reaching an AP of 29.6%). In contrast, our proposed frequency-domain enhancement strategy achieves best performance across all evaluation metrics (achieving an AP of 30.1% and an AP_{50} of 51.8%). This significant performance disparity confirms that, compared to pure spatial-domain enhancement methods, processing features in the frequency domain can more effectively extract and strengthen the fine-grained details of small objects while suppressing environmental noise, thereby achieving superior detection accuracy.

Table 7
Comparisons of different detail enhancement strategies for DCE-FPN on the Visdrone2019 dataset.

Method	$AP(\%)$	$AP_{50}(\%)$	$AP_{75}(\%)$
wo/DE	28.9	49.9	29.2
Sobel	29.3	50.1	29.4
Sobel + SA	29.6	50.5	29.6
Ours	30.1	51.8	30.4

Table 8
Comparisons of different attention modules on the VisDrone2019 dataset.

Methods	$AP(\%)$	$AP_{50}(\%)$	$AP_{75}(\%)$
SA	29.6	51.2	29.6
CA	29.7	51.2	29.8
DAM (ours)	**30.1**	**51.8**	**30.4**

The proposed Detail Attention Module employs an adaptive soft attention mechanism for dynamic noise suppression in high frequency components. We validate the robustness of this design by replacing it with standard Channel Attention and Spatial Attention modules within the DCE-FPN. As summarized in Table 8, our adaptive denoising strategy significantly outperforms the individual attention mechanisms. This confirms its distinct advantage in preserving critical object details while robustly filtering out environmental noise.

Furthermore, the number of branches k in CEM acts as a hyperparameter that determines the diversity of multi-scale contextual information. Consequently, we conducted further research on the impact of different k values on the

Table 9
Ablation study of different k values on the VisDrone2019 validation set. wo/CE denotes that no context enhancement is performed, which means k equals 0.

k	$AP(\%)$	$AP_{50}(\%)$	$AP_{75}(\%)$
2	30.1	51.8	30.4
3	30.0	51.4	29.8
4	30.1	51.6	30.6
5	29.8	51.3	29.9

performance of DCE-FPN, without employing the proposed EDI-RCNN. Specifically, we tested four distinct k values (2, 3, 4, and 5) on the VisDrone2019 validation set, with the experimental results detailed in Table 9. It is evident from the results that the performance is the worst when k equals 5. It reaches its peak when k equals 2 or 4 with AP=30.1%. The performance deteriorates when k is increased to 5, potentially because the excessively receptive field diminishes the distinctiveness of features associated with small objects. To ensure optimal performance of DCE-FPN while avoiding excessive computational complexity, we have chosen to set k to 2 as the default setting.

Table 10
Ablation study of GCEM and FIM on the VisDrone2019 validation set. The GCEM design adopts the proposed Method 2, as illustrated in Fig. 4, while the FIM design adopts the proposed Method 2, as illustrated in Fig. 5.

GCEM	FIM	$AP(\%)$	$AP_{50}(\%)$	$AP_{75}(\%)$
×	×	28.4	49.1	28.6
×	✓	29.1	50.1	29.2
✓	×	28.9	49.6	29.3
✓	✓	29.5	50.7	29.8

The effective fusion of multi-scale contextual information can significantly enhance the detection performance of small objects. To validate the efficacy of our proposed fusion method, we conducted a comparative evaluation of different fusion strategies on the VisDrone2019 validation set as shown in Table 11. Through channel shuffling and grouped convolution, multi-scale contextual information is effectively fused within the same channels, significantly improving the detection performance for small objects.

Table 11
Ablation study on the different fusion strategies of multi-scale contextual information using the VisDrone2019 validation set. $GConv$ denotes the group convolution and CS denotes the channel shuffling.

Fusion strategy	$AP(\%)$	$AP_{50}(\%)$	$AP_{75}(\%)$
$Concatenation$	29.7	50.7	29.7
Add	29.6	50.7	29.6
Ours (CS + $GConv$)	30.1	51.8	30.4

4) Ablation study on EDI-RCNN: EDI-RCNN not only provides rich global contextual information for the classification task through the proposed GCEM, but also considers the collaborative interaction between classification and localization tasks via the proposed FIM. This approach differs from the original decoupled detection heads proposed in YOLOX [12] and TBNet [53], which allow for fully independent optimization of classification and localization tasks. To more accurately assess the individual effectiveness of GCEM and FIM, we conducted further evaluation of their impact on the performance of EDI-RCNN using the VisDrone2019 validation set. Table 10 presents the experimental results, clearly indicating that the performance enhances with the incorporation of either GCEM or FIM individually, in comparison to the baseline model. Especially, the introduction of GCEM results in a 0.7% improvement in AP_{75} and the introduction of FIM leads to a 1.0% increase in AP_{50}. When both GCEM and FIM are introduced, EDI-RCNN gives best performance, with an AP of 29.5%, an AP_{50} of 50.7%, and an AP_{75} of 29.8%.

Table 12
Comparison of different global context enhancement methods on the VisDrone2019 validation set.

method	$AP(\%)$	$AP_{50}(\%)$	$AP_{75}(\%)$
Baseline	28.4	49.1	28.6
GCBlock	29.0	50.3	29.0
GCEM (Method 1)	29.4	50.5	29.5
GCEM (Method 2)	29.5	50.7	29.8

In the proposed EDI-RCNN, we drew inspiration from the design concept of GCBlock [56] and introduced the GCEM to enhance the global contextual information for the classification task. To verify the effectiveness of GCEM, we conducted a comparative analysis of its impact on the performance of EDI-RCNN against GCBlock on the VisDrone2019 validation set, with the results presented in Table 12. This table indicates that the application of GCBlock resulted in an improvement in the performance of EDI-RCNN, with an increase of 0.6% in AP, 1.2% in AP_{50}, and 0.4% in AP_{75}. When utilizing our designed GCEM, the performance improvement was even more pronounced, with an greatest (the Method 2 of GCEM) increase of 1.1% in AP, 1.6% in AP_{50}, and 1.2% in AP_{75}. It is noteworthy that Method 2 of GCEM captures global spatial context information better in both the length and width dimensions compared to Method 1, resulting in superior detection performance. In summary, above results demonstrate that considering the global contextual information cross the channel and spatial dimensions can further enhance detection accuracy.

In the proposed EDI-RCNN, the FIM module effectively facilitates collaborative interaction between the regression and classification tasks. To further validate the effectiveness of the two FIM variants we designed, we conducted comparative experiments on the VisDrone2019 validation set, with the results presented in Table 13. Clearly, compared to Method 1, Method 2 of FIM more effectively utilizes

(a) Origin images (b) Faster-RCNN [1] (c) FCOS [3] (d) RFLA [30] (e) CFPT [68] (f) DCGNet

Figure 7: Visualization of detection results across various methods on the AI-TODv2. Green, blue, and red boxes denote true positive (TP), false positive (FP), and false negative (FN), respectively. It is evident that under complex background condition, our proposed DCGNet exhibits the smallest number of false detections and missing detections.

the complementarity of features across different tasks, rather than simply performing feature rectification, thereby further enhancing the interaction and fusion between tasks.

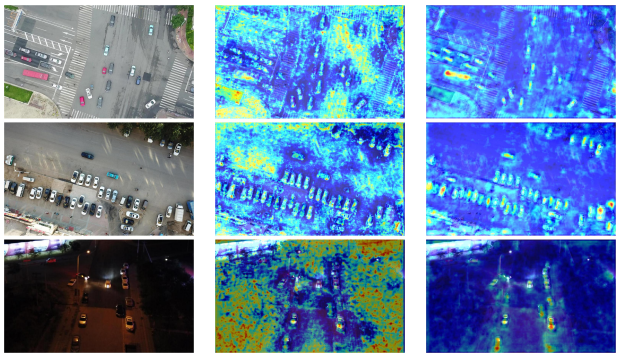
Table 13
Comparison of different FIM designs on the VisDrone2019 validation set.

method	$AP(\%)$	$AP_{50}(\%)$	$AP_{75}(\%)$
FIM (Method 1)	29.3	50.4	29.5
FIM (Method 2)	29.5	50.7	29.8

4.6. Analysis

Visualization of comparative detection results: As shown in Fig. 7, we present a comparative visualization of detection results from the typical one-stage detector FCOS [3], the two-stage detector Faster-RCNN [1], the recent SOTA models including the RFLA baseline [30] and CFPT [68], and our proposed DCGNet. To rigorously evaluate the models under extreme constraints, we specifically selected testing samples featuring high-density distributions (the first two rows) and complex environmental conditions, such as heavy occlusion and low illumination (the last two rows). Small objects are highly susceptible to environmental noise and crowding due to their limited feature representations. As observed in Fig. 7(b) to 7(e), existing methods, including the recent CFPT [68], struggle to separate adjacent targets in high-density scenes and fail to extract valid features in heavily occluded environments, leading to a significant number of missed detections (indicated by red FN boxes). In contrast, as demonstrated in Fig. 7(f), ours proposed DCGNet effectively alleviates this interference by explicitly enhancing detailed and contextual information. Consequently, it successfully delineates dense objects with distinct boundaries and maintains remarkable robustness against heavy occlusion, significantly reducing the instances of both missing detection and false detection in real-world scenarios.

Visualization of features: To intuitively compare the capabilities of the proposed DCE-FPN and the traditional FPN in extracting detailed information of small objects under various complex scenarios, we visualize their output feature maps P_2, in Fig. 8. Darker colors in the maps indicate higher response intensities from the model. As illustrated in Fig. 8(b), the traditional FPN struggles in complex scenes, due to the inherent weakness of small object features and the interference of environmental noise. Specifically, in the dense scenario (the second row), the feature responses of FPN exhibit severe adhesion, where the features of adjacent small objects merge together, making it difficult to distinguish individuals. Furthermore, in the low-light scenario (the third row), the valid features of FPN are largely corrupted by background noise.



(a) Origin images (b) FPN (c) DCE-FPN

Figure 8: Visualization results of feature maps P_2 from both traditional FPN and DCE-FPN. (a) Input images from VisDrone2019. (b) Feature maps P_2 from FPN. (c) Feature maps P_2 from DCE-FPN. The first row to the third row represent the typical small object scenario, the dense scenario, and the low-light scenario, respectively. In these feature maps, the deeper the color, the more attention the model pays to that area.

In contrast, DCE-FPN first utilizes the Sobel operator to enhance the edge details of small objects, and subsequently strengthens these details in the frequency domain while effectively suppressing environmental noise. Simultaneously, by capturing multi-scale contextual information, this module improves the model's ability to separate targets from complex backgrounds. Benefiting from these mechanisms, as shown in Fig. 8(c), DCE-FPN effectively decouples adjacent targets in dense scenes, yielding clear boundaries and isolated feature responses for individual objects. Even under low-visibility conditions, it still guides the model to accurately focus on the target regions. Ultimately, compared to the baseline, the features of small objects extracted by DCE-FPN become significantly more prominent and precise.

5. Conclusions

Small objects typically contain limited feature information because they occupy only a small number of pixels, making them highly susceptible to environmental noise. To tackle this challenge, this paper has presented an effective detection framework, DCGNet, which incorporates the DCE-FPN to enrich fine-grained feature representations in the frequency domain while capturing multi-scale contextual information, thereby reducing noise interference. Furthermore, the EDI-RCNN module enhances detection accuracy by providing task-specific features and encouraging effective feature interaction across detection stages. Experimental results show that DCGNet outperforms recent state-of-the-art approaches in both accuracy and robustness for small object detection. Nevertheless, the method has certain limitations: the multi-branch architecture used for multi-scale feature extraction increases computational complexity, posing challenges for real-time deployment, and the lack of a dedicated loss function for the decoupled detection head limits its full potential. Future work will focus on reducing model complexity and developing loss functions specifically tailored to small object detection to further improve performance.

References

[1] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (6) (2016) 1137–1149.

[2] Z. Cai, N. Vasconcelos, Cascade R-CNN: Delving into high quality object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6154–6162.

[3] Z. Tian, C. Shen, H. Chen, T. He, Fcos: Fully convolutional one-stage object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9627–9636.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, SSD: Single shot multibox detector, in: Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, October 11–14, 2016, Springer, 2016, pp. 21–37.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer, 2020, pp. 213–229.

[6] H. Liang, Y. Yang, Q. Zhang, L. Feng, J. Ren, Q. Liang, Transformed dynamic feature pyramid for small object detection, IEEE Access 9 (2021) 134649–134659.

[7] C. Deng, M. Wang, L. Liu, Y. Liu, Y. Jiang, Extended feature pyramid network for small object detection, IEEE Transactions on Multimedia 24 (2021) 1968–1979.

[8] T. Shi, J. Gong, J. Hu, X. Zhi, W. Zhang, Y. Zhang, P. Zhang, G. Bao, Feature-enhanced centernet for small object detection in remote sensing images, Remote Sensing 14 (21) (2022) 5488.

[9] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings Of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[10] H.-I. Liu, Y.-W. Tseng, K.-C. Chang, P.-J. Wang, H.-H. Shuai, W.-H. Cheng, A denoising fpn with transformer r-cnn for tiny object detection, IEEE Transactions on Geoscience and Remote Sensing 62 (2024) 1–15.

[11] J. Zhuang, Z. Qin, H. Yu, X. Chen, Task-specific context decoupling for object detection, arXiv preprint arXiv:2303.01047 (2023).

[12] Z. Ge, YOLOX: Exceeding YOLO series in 2021, arXiv preprint arXiv:2107.08430 (2021).

[13] L. Ni, X. Pan, X. Wang, D. Bao, J. Zhang, J. Shi, Small-object detection model for optical remote sensing images based on tri-decoupling++ head, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2024).

[14] Q. Li, L. Shen, S. Guo, Z. Lai, WaveCNet: Wavelet integrated cnns to suppress aliasing effect for noise-robust image classification, IEEE Transactions on Image Processing 30 (2021) 7074–7089.

[15] Y. Qiao, S. Lan, W. Wang, H. Chen, Y. Li, G. Deng, DCGNet: Detail and context guided small object detection network with decoupled detection head, in: IEEE International Conference on Multimedia and Expo (ICME), 2025, pp. 1–6.

[16] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, G. Ding, YOLOv10: Real-time end-to-end object detection, arXiv preprint arXiv:2405.14458 (2024).

[17] X. Cai, Q. Lai, Y. Wang, W. Wang, Z. Sun, Y. Yao, Poly kernel inception network for remote sensing detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 27706–27716.

[18] D. Liu, J. Zhang, Y. Qi, Y. Xi, J. Jin, Exploring lightweight structures for tiny object detection in remote sensing images, IEEE Transactions on Geoscience and Remote Sensing 63 (2025) 1–15.

[19] J. Luo, X. Yang, Y. Yu, Q. Li, J. Yan, Y. Li, Pointobb: Learning oriented object detection via single point supervision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16730–16740.

[20] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR: Deformable transformers for end-to-end object detection, in: International Conference on Learning Representations, 2021.

[21] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, L. Zhang, DAB-DETR: Dynamic anchor boxes are better queries for DETR, arXiv preprint arXiv:2201.12329 (2022).

[22] D. Jia, Y. Yuan, H. He, X. Wu, H. Yu, W. Lin, L. Sun, C. Zhang, H. Hu, DETRs with hybrid matching, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19702–19712.

[23] Z. Cai, S. Liu, G. Wang, Z. Ge, X. Zhang, D. Huang, Align-DETR: Improving detr with simple iou-aware bce loss, arXiv preprint arXiv:2304.07527 (2023).

[24] M. Kisantal, Augmentation for small object detection, arXiv preprint arXiv:1902.07296 (2019).

[25] X. Zhang, E. Izquierdo, K. Chandramouli, Dense and small object detection in uav vision based on cascade network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0.

[26] B. Bosquet, D. Cores, L. Seidenari, V. M. Brea, M. Mucientes, A. Del Bimbo, A full data augmentation pipeline for small object detection based on generative adversarial networks, Pattern Recognition 133 (2023) 108998.

[27] S. Shi, Q. Fang, T. Zhao, X. Xu, Similarity distance-based label assignment for tiny object detection, arXiv preprint arXiv:2407.02394 (2024).

[28] J. Wang, C. Xu, W. Yang, L. Yu, A normalized gaussian wasserstein distance for tiny object detection, arXiv preprint arXiv:2110.13389 (2021).

[29] C. Xu, J. Wang, W. Yang, L. Yu, Dot distance for tiny object detection in aerial images, in: Proceedings Of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1192–1201.

[30] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, G.-S. Xia, RFLA: Gaussian receptive field based label assignment for tiny object detection, in: European Conference on Computer Vision, Springer, 2022, pp. 526–543.

[31] G. Cheng, X. Yuan, X. Yao, K. Yan, Q. Zeng, X. Xie, J. Han, Towards large-scale small object detection: Survey and benchmarks, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (11) (2023) 13467–13488.

[32] Y. Yuan, Y. Zhao, D. Ma, NACAD: A noise-adaptive context-aware detector for remote sensing small objects, IEEE Transactions on Geoscience and Remote Sensing 61 (2023) 1–13.

[33] L. Cui, P. Lv, X. Jiang, Z. Gao, B. Zhou, L. Zhang, L. Shao, M. Xu, Context-aware block net for small object detection, IEEE Transactions on Cybernetics 52 (4) (2020) 2300–2313.

[34] Z. Zhang, P. Gong, H. Sun, P. Wu, X. Yang, Dynamic local and global context exploration for small object detection, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[35] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.

[36] Y. Li, Q. Huang, X. Pei, Y. Chen, L. Jiao, R. Shang, Cross-layer attention network for small object detection in remote sensing imagery, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14 (2020) 2148–2161.

[37] Y. Sun, C. Xu, J. Yang, H. Xuan, L. Luo, Frequency-spatial entanglement learning for camouflaged object detection, arXiv preprint arXiv:2409.01686 (2024).

[38] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems 25 (2012).

[39] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[40] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[41] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[42] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768.

[43] M. Tan, R. Pang, Q. V. Le, Efficientdet: Scalable and efficient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10781–10790.

[44] Y. Xiao, T. Xu, X. Yu, Y. Fang, J. Li, A lightweight fusion strategy with enhanced interlayer feature correlation for small object detection, IEEE Transactions on Geoscience and Remote Sensing 62 (2024) 1–11.

[45] D. Wu, W. Wu, H. Zheng, F. Wang, G. Gao, Dwt-yolo: Wavelet-based detection to enhance sensor resolution for small targets, Concurrency and Computation: Practice and Experience 37 (27-28) (2025) e70419.

[46] X. Shao, S. Diao, L. Li, X. Zhao, Y. Mei, Z. Zhu, Wt-detr: Wavelet-enhanced detr for robust tiny object detection via multi-scale feature optimization, Journal of Real-Time Image Processing 22 (5) (2025) 186.

[47] Z.-X. Li, Y.-L. Wang, F. Wang, Di-yolov5: An improved dual-wavelet-based yolov5 for dense small object detection, IEEE/CAA Journal of Automatica Sinica 12 (2) (2025) 457–459.

[48] J. Redmon, YOLOv3: An incremental improvement, arXiv preprint arXiv:1804.02767 (2018).

[49] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, et al., YOLOv6: A single-stage object detection framework for industrial applications, arXiv preprint arXiv:2209.02976 (2022).

[50] C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7464–7475.

[51] Z. Chen, C. Yang, Q. Li, F. Zhao, Z.-J. Zha, F. Wu, Disentangle your dense object detector, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4939–4948.

[52] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 764–773.

[53] Z. Li, Y. Wang, D. Xu, Y. Gao, T. Zhao, TBNet: A texture and boundary-aware network for small weak object detection in remote-sensing imagery, Pattern Recognition 158 (2025) 110976.

[54] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[55] A. G. Howard, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 (2017).

[56] Y. Cao, J. Xu, S. Lin, F. Wei, H. Hu, Gcnet: Non-local networks meet squeeze-excitation networks and beyond, in: Proceedings of the IEEE/CVF International Conference on Computer Vision workshops, 2019, pp. 0–0.

[57] A. Vaswani, Attention is all you need, Advances in Neural Information Processing Systems 30 (2017).

[58] B. Cao, H. Yao, P. Zhu, Q. Hu, Visible and clear: Finding tiny objects in difference map, arXiv preprint arXiv:2405.11276 (2024).

[59] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 740–755.

[60] J. Wang, W. Yang, H. Guo, R. Zhang, G.-S. Xia, Tiny object detection in aerial images, in: 25th International Conference on Pattern Recognition, IEEE, 2021, pp. 3791–3798.

[61] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, et al., Mmdetection: Open mmlab detection toolbox and benchmark, arXiv preprint arXiv:1906.07155 (2019).

[62] S. Qiao, L.-C. Chen, A. Yuille, Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution, in: Proceedings Of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10213–10224.

[63] S. Zhang, C. Chi, Y. Yao, Z. Lei, S. Z. Li, Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9759–9768.

[64] B. Du, Y. Huang, J. Chen, D. Huang, Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13435–13444.

[65] J. Wu, Z. Pan, B. Lei, Y. Hu, FSANet: Feature-and-spatial-aligned network for tiny object detection in remote sensing images, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–17.

[66] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, G.-S. Xia, Detecting tiny objects in aerial images: A normalized wasserstein distance and a new benchmark, ISPRS Journal of Photogrammetry and Remote Sensing 190 (2022) 79–93.

[67] D. Liu, J. Zhang, Y. Qi, Y. Wu, Y. Zhang, Tiny object detection in remote sensing images based on object reconstruction and multiple receptive field adaptive feature enhancement, IEEE Transactions on Geoscience and Remote Sensing 62 (2024) 1–13.

[68] Z. Du, Z. Hu, G. Zhao, Y. Jin, H. Ma, Cross-layer feature pyramid transformer for small object detection in aerial images, IEEE Transactions on Geoscience and Remote Sensing 63 (2025) 1–14.

[69] J. Bian, M. Feng, W. Dong, F. Wu, J. Luo, Y. Wang, G. Shi, Feature information driven position gaussian distribution estimation for tiny object detection, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 30376–30386.