FSDC-DETR: A Frequency-Spatial Domain Collaborative DETR for Small Object Detection

Aiwen Liu[1], Chengguang Zhu[1](✉), Gang Wang[1], Dandan Zhu[2], Haodong Lin[1], Yan Wang[4], Huiyu Zhou[3], and Zhengyi Pan[1]

[1] Micro-Intelligence, Shanghai 201100, China
nevereverinsomnia@gmail.com, chengguang.zhusjtu@gmail.com, {roy.wang,haodong.lin,eric.pan}@micro-i.com.cn
[2] East China Normal University, Shanghai 200241, China
ddzhu@mail.ecnu.edu.cn
[3] University of Leicester, Leicester LE1 7RH, UK
hz143@leicester.ac.uk
[4] Chongqing Normal University, Chongqing 401331, China

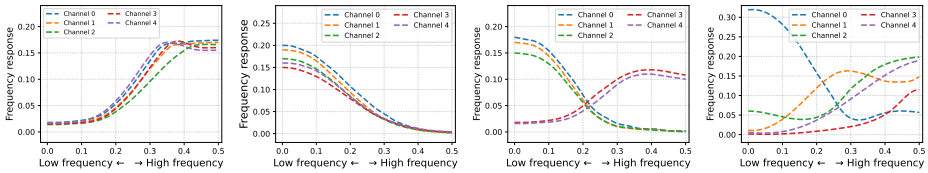
Abstract. Small object detection (SOD) remains a challenging task in real-world applications. Despite recent advances, existing detectors remain limited by rigid processing that entangle spatial aggregation with implicit frequency aliasing and truncation, leading to inadequate preservation of high-frequency components for SOD. To tackle these limitations, we propose a Frequency-Spatial Domain Collaborative Detection Transformer (FSDC-DETR), a novel collaborative framework that explicitly models complementary spatial and frequency representations. Specifically, we first introduce Dual-Branch Frequency-Spatial Adaptive Fusion (DBFSAF) to enhance frequency diversity and adaptively capture frequency-spatial domain discriminative representations. Building on these representations, a frequency-spatial interaction scheme is further explored within the hybrid encoder to enable progressive feature propagation to the decoder. In particular, structure-aware frequency-spatial aggregation is achieved through Shunt Frequency-Spatial Feature Fusion (SFS-FF), establishing bidirectional interaction and progressive cross-scale propagation between frequency and spatial representations for coherent discriminative modeling. Meanwhile, informative high-frequency responses are preserved during scale transitions through Frequency-Spatial Dynamic Downsampling (FSD-Down), thereby minimizing frequency degradation throughout multi-scale fusion for the precise SOD. Experimental results demonstrate that FSDC-DETR achieves state-of-the-art performance, improving AP by 6.4 on VisDrone-DET2019 and 6.6 on AITODv2, with gains of 6.8 and 6.9 AP for small objects. The code is available at https://github.com/nevereverinsomnia/FSDC-DETR.

Keywords: small object detection · detection transformer · frequency-spatial collaborative modeling · multi-scale feature fusion

⋆ Aiwen Liu, Chengguang Zhu, and Gang Wang — Equal contribution.

1 Introduction

Object detection is a fundamental task in computer vision that has witnessed significant progress in recent years, supporting a wide range of applications such as industrial defect detection [12, 37, 79, 90], aerial surveillance [11, 22, 39, 80], remote sensing [2, 34, 53, 54], medical lesion analysis [16, 73, 88], and autonomous systems [13, 52, 75, 83]. Most of these advances have been driven by Convolutional Neural Network (CNN)-based detectors, which typically rely on a series of manually designed components, such as anchor box generation and non-maximum suppression (NMS), and remain heavily dependent on human expertise [18, 27, 49, 63]. These components often require extensive task-specific tuning, making the overall detection pipeline complex and inefficient in real-world applications. More recently, the end-to-end Transformer-based Detection Transformer (DETR) framework has been introduced to largely alleviate these limitations by reformulating object detection as a direct set prediction problem [4]. Subsequent DETR-based variants have further improved the detection accuracy while significantly improving inference efficiency [25, 35, 43, 92]. However, these advances are primarily designed for generic object detection and still face substantial challenges when applied to small object detection (SOD), particularly in scenarios such as industrial defect detection and aerial imagery acquired from unmanned aerial vehicles (UAVs) and satellites.



(a) CNN **(b)** ViT **(c)** DEIMv2 **(d)** FSDC-DETR

Fig. 1: Frequency response analysis of different backbone architectures.

SOD focuses on identifying and localizing objects with extremely limited spatial extent in images [68, 77]. Although spatially limited, these objects play a critical role in real-world applications. The challenges of SOD mainly stem from insufficient feature representation capacity in the model backbone, key information loss in the deeper layers of feature hierarchies, and frequency-domain aliasing during multi-scale feature fusion. As a result, small objects often experience feature loss and inadequate representation, which leads to reduced detection accuracy. RT-DETR [92] improves inference efficiency by decoupling intra-scale and cross-scale interactions, yet it does not explicitly model frequency-domain representations. As a result, high-frequency details are implicitly attenuated during feature aggregation and downsampling, leading to the loss of fine-grained information critical for SOD tasks. As illustrated in Fig. 1(a)-(b), CNN and Vision Transformer (ViT) backbones exhibit substantially different frequency response

characteristics [1, 8, 9, 42, 55, 60, 67, 69]. Although DEIMv2 [24] adopts a dual-branch CNN-ViT backbone, feature fusion is performed via direct concatenation followed by a simple channel projection. In the absence of explicit frequency-aware modeling, this design overlooks the heterogeneous spectral properties of CNN and ViT representations, leading to insufficient preservation of fine-grained high-frequency details, as illustrated in Fig. 1(c).

In this paper, we introduce the Frequency-Spatial Domain Collaborative Detection Transformer (FSDC-DETR), a novel end-to-end transformer-based object detection framework designed to enhance SOD through explicit frequency-spatial modeling. To construct frequency-spatial collaborative representations and mitigate frequency aliasing and truncation in dual-branch backbones, we first propose a Dual-Branch Frequency-Spatial Adaptive Fusion (DBFSAF) module that enhances frequency diversity of the dual-branch representations, adaptively fuses and selects features with increased preservation of high-frequency components, and performs Partial Frequency-Spatial Refinement (PFSR). The corresponding frequency response is illustrated in Fig. 1(d). Building upon these refined representations, we further develop a Shunt Frequency-Spatial Feature Fusion (SFS-FF) strategy within a hybrid encoder to propagate frequency-aware features across scales. By decoupling and collaboratively refining spatial and frequency components, SFS-FF enables structure-aware cross-scale aggregation without suppressing fine-grained details. To preserve frequency integrity during resolution transitions in multi-scale modeling, we introduce a Frequency-Spatial Dynamic Downsampling (FSD-Down) operator based on learnable wavelet decomposition and grouped convolution. FSD-Down retains informative high-frequency responses while maintaining spatial consistency, ensuring minimal frequency degradation throughout the feature hierarchy.

Experimental results on VisDrone-DET2019 [15] and AITODv2 [77] demonstrate that FSDC-DETR establishes a new state-of-the-art for SOD, delivering substantial improvements over strong baselines, especially in the AP_S metric. These results verify that explicit frequency-spatial modeling effectively preserves high-frequency cues critical for SOD. In summary, our main contributions are as follows:

- We propose FSDC-DETR, a frequency-spatial domain collaborative DETR framework that explicitly models complementary spatial and frequency representations to address aliasing and high-frequency degradation in SOD.
- To systematically construct and propagate frequency-aware representations, we design a unified frequency-spatial collaborative pipeline composed of three progressively integrated components. (i) The DBFSAF module mitigates frequency aliasing and truncation phenomenon in dual-branch backbones by enhancing spectral diversity and adaptively preserving informative high-frequency cues. (ii) The SFS-FF strategy within the hybrid encoder propagates frequency-enriched features across scales through structure-aware frequency-spatial aggregation. (iii) The FSD-Down operator, built upon learnable wavelet decomposition and grouped convolution, preserves frequency integrity during scale transitions while maintaining spatial consistency.

– Experimental results on VisDrone-DET2019 and AITODv2 demonstrate state-of-the-art performance, improving AP by 6.4 and 6.6, with gains of 6.8 and 6.9 AP on small objects, respectively.

2 Related Work

2.1 Small Object Detection

SOD remains particularly challenging due to limited pixel support, low signal-to-noise ratio, and severe feature attenuation during hierarchical structure. Compared with the general detection scenario, small objects occupy only a few spatial locations, making them highly vulnerable to the spatial reduction and the smoothing representation. Repeated multi-scale aggregation introduces progressive attenuation of high-frequency components, undermining the discriminative signals required for accurate detection.

Conventional CNN-based detectors, such as Fast R-CNN [18, 49] and YOLO [27–29, 51, 61, 63–66, 72, 76], often exhibit performance degradation on small objects due to insufficient fine-grained feature representation and limited long-range dependency modeling. Although data augmentation strategies (e.g., copy-paste techniques [17]) and IoU-aware loss [20, 62, 81, 89] formulations attempt to mitigate scale imbalance, these approaches primarily operate at the data or optimization level, leaving representation degradation comparatively underexplored [25, 35, 86].

Transformer-based detectors, particularly DETR [4] and its variants [24, 25, 30, 35, 38, 43, 48, 50, 70, 84, 91, 92], introduce global attention mechanisms that enhance long-range modeling and remove hand-crafted components such as anchors and NMS. However, vanilla DETR suffers from high computational cost and slow convergence, limiting its applicability in real-time scenarios. Recent variants, including RT-DETR [35, 38, 70, 92], D-FINE [43], Mr. DETR [84], and DEIM [25], improve efficiency and accelerate training, making DETR-style models more practical. Nevertheless, despite architectural refinements, existing hybrid encoder designs still rely heavily on conventional multi-scale fusion and spatial downsampling schemes, where high-frequency information may be weakened during scale transition, posing challenges for SOD [54, 58, 87].

2.2 DETRs with Vision Foundation Model

Vision foundation models, such as the DINO family [5, 40, 56], SAM3 [3], and C-RADIOv4 [47], have substantially alleviated the limitations of feature representation in diverse visual scenarios. RT-DETRv4 [35] introduces a semantic transfer strategy that leverages DINOv3 to supervise the AIFI [92] module. Specifically, the high-level representations produced by DINOv3 are used as supervisory signals for the AIFI output to enhance semantic richness and improve feature discrimination. RF-DETR [50] replaces conventional backbones with a DINOv2-pretrained ViT to inherit strong self-supervised visual priors

and further employs weight-sharing neural architecture search to optimize detection performance under different computational budgets. DEIMv2 [24], built upon DEIM [25], adopts a dual-branch architecture that couples DINOv3 with a lightweight CNN backbone. Although this design aims to integrate global contextual modeling with local structural extraction, how to effectively facilitate the interaction between the two branches still deserves further exploration. In particular, DEIMv2 overlooks the frequency-domain disparity between transformer-based and CNN-based representations. ViTs typically exhibit low-pass filtering characteristics due to global attention aggregation [8, 55, 60, 69], whereas CNNs inherently preserve stronger high-frequency components [1, 9, 42, 67]. Without explicit frequency-aware alignment, directly fusing these heterogeneous representations may introduce aliasing effects and hinder effective information complementarity. Consequently, the potential synergy between the two branches is not fully exploited.

2.3 Frequency Domain Learning

Frequency-domain analysis has long served as a fundamental tool in signal processing [19,44]. Recently, it has been increasingly incorporated into deep learning, where it has influenced model robustness [82] and the generalization capability of computer vision models [9, 54, 67]. Beyond theoretical analysis, frequency-aware designs have been integrated into network architectures to enhance global representation learning [26, 32, 33, 57, 59, 79] and domain generalization [31, 67]. Several works explicitly incorporate frequency modeling into CNNs. For instance, FcaNet [45] models channel attention in the frequency domain, extending global average pooling to multi-spectral channel representations. HS-FPN [54] employs high-pass filtering to capture high-frequency responses and uses them as spatial- and channel-wise masks within FPN to enhance fine-grained representations for SOD. FreqFusion [6] introduces a frequency-aware feature fusion framework that employs adaptive low-pass and high-pass filtering to reduce intra-class inconsistency and enhance boundary details during multi-scale feature fusion. SFM [7] reduces aliasing by modulating high-frequency features before downsampling and restoring them during upsampling to better preserve fine details. FADC [10] improves dilated convolution from a spectral perspective by adaptively adjusting dilation rates and reweighting frequency components to better balance receptive field size and effective bandwidth. SET [58] advances tiny object detection from a spectral perspective by suppressing background-dominated high-frequency noise and introducing adversarial perturbation learning to enhance the saliency of tiny-object representations. Recent studies [41, 93] demonstrate that capturing balanced representations of both high- and low-frequency components can enhance model performance. FCENet [71] exploits cross-field frequency correlations between high- and low-frequency components, dynamically selecting and fusing complementary spectral information for improved NIR-assisted image denoising. Nevertheless, existing frequency-aware methods mostly focus on frequency modeling within individual architectures, leaving the complementary fusion of

high-frequency details [8, 55, 60, 69] and low-frequency semantics [1, 9, 42, 67] remains insufficiently investigated.

3 The Proposed Model

3.1 The FSDC-DETR Overview

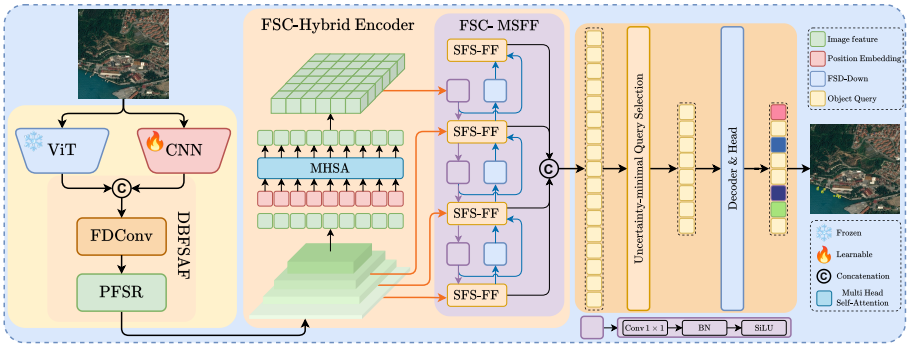


Fig. 2: Overall architecture of the proposed FSDC-DETR

As illustrated in Fig. 2, we propose FSDC-DETR, an enhanced variant of DEIMv2 [24] for SOD through explicit frequency-spatial modeling. The overall framework forms a unified frequency-spatial collaborative pipeline that progressively constructs, propagates, and preserves frequency-aware representations. Specifically, to construct frequency-aware representations, we introduce the DBFSAF module which enhances spectral diversity, adaptively preserves complementary high-frequency cues, and performs frequency-spatial collaborative refinement. Building upon these refined representations, the SFS-FF module propagates frequency-enriched features across scales via structure-aware spectral-spatial aggregation. To preserve frequency integrity during resolution transitions, the FSD-Down module facilitates scale transformation while retaining informative high-frequency responses and maintaining spatial consistency.

3.2 Dual-Branch Frequency-Spatial Adaptive Fusion

The representational capabilities of CNNs and ViTs have been widely demonstrated. ViTs exhibit a pronounced low-pass filtering behavior [8, 55, 60, 69], primarily caused by the global attention mechanism, which tends to over-smooth feature representations and weakens local modeling capability. In contrast, CNNs naturally preserve high-frequency components [1, 9, 42, 67] due to their local receptive fields, but often suffer from limited global context modeling. DEIMv2 [24] introduces a dual-branch backbone composed of a DINOv3-pretrained ViT [56],

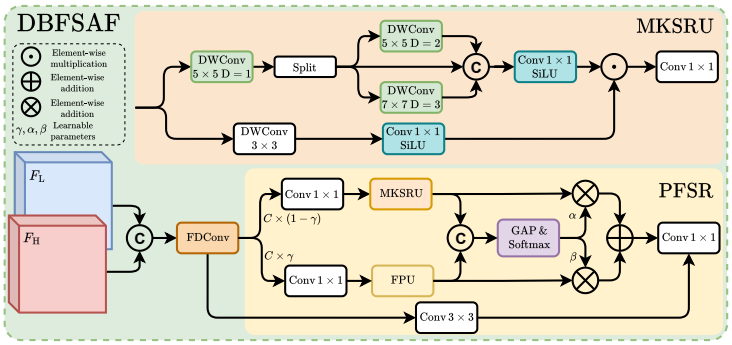


Fig. 3: Illustration of the Dual-Branch Frequency-Spatial Adaptive Fusion.

which provides robust feature representations in various scenarios, and a simple CNN, enabling complementary representation learning and achieving state-of-the-art performance in general object detection. However, DEIMv2 combines heterogeneous representations from the ViT and CNN branches via channel-wise concatenation, followed by a 1×1 convolution prior to multi-scale aggregation. Without explicit frequency-aware adaptation, such a design does not account for the distinct frequency characteristics of ViTs and CNNs, potentially limiting the effective preservation and utilization of high-frequency components, particularly for SOD.

The detection task is crucial to adaptively capture diverse frequency components across the feature map, as different spatial locations exhibit distinct local frequency characteristics. In particular, object boundaries and fine structures rely heavily on high-frequency information for precise localization, whereas homogeneous or smooth regions benefit less from such details.

To address the aforementioned limitation, we introduce a DBFSAF module, inspired by Frequency Dynamic Convolution (FDConv) [9] and Space-Frequency Selection Convolution (SFS-Conv) [33]. As depicted in Fig. 3, DBFSAF first applies FDConv to the concatenated dual-branch features extracted from HGNetv2 [14] and DINOv3 [56], enabling dynamic frequency diversification while maintaining discriminative high-frequency components within the adaptively selected feature representations. Subsequently, a PFSR module performs collaborative frequency-spatial refinement to enhance representation quality with minimal computational overhead. Let F_L^P and F_H^P denote the low-frequency and high-frequency feature maps from HGNETv2 and DINOv3, at stage $\mathbf{P} \in \{2, 3, 4, 5\}$, respectively. The DBFSAF module can be formulated as:

$$F_{\mathbf{C}}^P, F_{\mathbf{FS}}^P = \text{Partial}_\gamma \left(\text{FDConv} \left(\text{Concat} \left[\mathbf{F}_{\text{L}}^P, \mathbf{F}_{\text{H}}^P \right] \right) \right), \tag{1}$$

where $\text{Partial}_\gamma(\cdot)$ denotes a channel partition operation applied to the FDConv output with C channels. Specifically, the feature map is divided according to a hyperparameter γ, which yields $C \times \gamma$ channels feature F_{FS}^P for jointly Frequency-Spatial refinement and $C \times (1-\gamma)$ channels feature F_C^P preserved to alleviate

channel redundancy. This partial design reduces redundant inter-channel interactions and avoids unnecessary computations, thereby achieving a favorable trade-off between robustness and accuracy [23].

Building upon the partially selected representation $F_{\mathbf{FS}}$, we further introduce a Multi-Kernel Spatial Refine Unit (MKSRU) that operates in coordination with the Frequency Processing Unit (FPU) proposed in [33], to jointly enhance spatial structures and frequency-sensitive responses. The refinement process is formulated as follows:

$$F_{\mathbf{S}}^{P} = \mathrm{MKSRU}\big(\mathrm{Conv}_{1\times1}\left(F_{\mathbf{FS}}^{P}\right)\big), \tag{2}$$

$$F_{\mathbf{F}}^{P} = \mathrm{FPU}\big(\mathrm{Conv}_{1\times1}\left(F_{\mathbf{FS}}^{P}\right)\big), \tag{3}$$

$$\hat{F}_{\mathbf{S}}^{P} = \alpha \odot \mathrm{Softmax}\big(\mathrm{GAP}\big(\mathrm{Concat}\big[F_{\mathbf{S}}^{P},\ F_{\mathbf{F}}^{P}\big]\big)\big) \odot F_{\mathbf{S}}^{P}, \tag{4}$$

$$\hat{F}_{\mathbf{F}}^{P} = \beta \odot \mathrm{Softmax}\big(\mathrm{GAP}\big(\mathrm{Concat}\big[F_{\mathbf{S}}^{P},\ F_{\mathbf{F}}^{P}\big]\big)\big) \odot F_{\mathbf{F}}^{P}, \tag{5}$$

$$F^{P} = \mathrm{Conv}_{1\times1}\Big(\mathrm{Concat}\Big[\hat{F}_{\mathbf{S}}^{P} + \hat{F}_{\mathbf{F}}^{P},\ \mathrm{Conv}_{3\times3}\left(F_{\mathbf{C}}^{P}\right)\Big]\Big), \tag{6}$$

where $\odot$ denotes the Hadamard product, α and β are learnable scalar parameters, and $\mathrm{GAP}(\cdot)$ denotes global average pooling. $\hat{F}_{\mathbf{S}}^{P}$ and $\hat{F}_{\mathbf{F}}^{P}$ represent the channel-attended spatial and frequency refinement features $F_{\mathbf{S}}^{P}$ and $F_{\mathbf{F}}^{P}$ generated by MKSRU and FPU, respectively.

Multi-kernel convolution blocks have been widely adopted in various computer vision tasks to enhance and refine spatial representations [36,46,59,85]. Motivated by their strong capability in capturing multi-scale structural cues, we design the developed MKSRU as follows. Let the input feature be $X_{\mathbf{in}} \in \mathbb{R}^{\mathbf{C}\times\mathbf{H}\times\mathbf{W}}$:

$$X_{\mathbf{D}}^{\frac{1}{2}\mathbf{C}},\ X_{\mathbf{D}}^{\frac{1}{8}\mathbf{C}},\ X_{\mathbf{D}}^{\frac{3}{8}\mathbf{C}} = \mathrm{Split}\Big(\mathrm{DWConv}_{5\times5}^{(1)}\left(X_{\mathbf{in}}\right)\Big), \tag{7}$$

$$\hat{X}_{\mathbf{D}} = \mathrm{Concat}\Big[\mathrm{DWConv}_{7\times7}^{(3)}\left(X_{\mathbf{D}}^{\frac{1}{2}\mathbf{C}}\right),\ \mathrm{DWConv}_{7\times7}^{(2)}\left(X_{\mathbf{D}}^{\frac{3}{8}\mathbf{C}}\right),\ X_{\mathbf{D}}^{\frac{1}{8}\mathbf{C}}\Big], \tag{8}$$

$$X_{\mathbf{out}} = \mathrm{SiLU}\Big(\mathrm{Conv}_{1\times1}\left(\hat{X}_{\mathbf{D}}\right)\Big) \odot \mathrm{SiLU}(\mathrm{Conv}_{1\times1}\left(\mathrm{Conv}_{3\times3}\left(X_{\mathbf{in}}\right)\right)), \tag{9}$$

where $\mathrm{DWConv}_{j\times j}^{(i)}$ denotes a depth-wise convolution with kernel size $j \times j$ and dilation rate i, and $\mathrm{SiLU}(\cdot)$ denotes the activation function.

The proposed DBFSAF block explicitly leverages the complementary frequency characteristics of ViT and CNN representations. Through FDConv and PFSR module, the framework increases frequency diversity while preserving discriminative high-frequency details and suppressing redundant channel interactions. The coordinated frequency-spatial enhancement via MKSRU and FPU further enables multi-scale structural aggregation with minimal computational overhead. The DBFSAF is particularly advantageous for SOD, where fine-grained boundary cues are essential, leading to improved localization accuracy and robustness.

3.3 Shunt Frequency-Spatial Feature Fusion

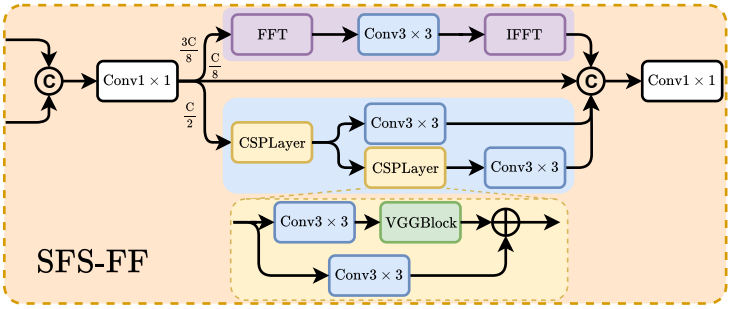


Fig. 4: Illustration of the Shunt Frequency-Spatial Feature Fusion.

Most existing hybrid encoders [24, 25, 35, 38, 43, 70, 92] rely on convolution-based feature fusion and downsampling operations with spatial reduction, which often incur frequency misalignment and attenuate high-frequency components that are essential for SOD. Motivated by this limitation, we revisit the hybrid encoder and propose the Frequency-Spatial Collaborative Hybrid Encoder (FSC-Hybrid Encoder) module jointly the SFS-FF and FSD-Down, which is introduced in Section 3.4. This design realizes Frequency-Spatial Collaborative Multi-Scale Feature Fusion (FSC-MSFF) by explicitly modeling complementary frequency and spatial representations. By collaboratively refining frequency cues and enhancing spatial structure awareness, the proposed hybrid encoder yields more discriminative fine-grained features for SOD.

As illustrated in Fig. 4, given the input features $F_{\mathbf{in}} \in \mathbb{R}^{C\times H\times W}$ and $\hat{F}_{\mathbf{in}} \in \mathbb{R}^{C\times 2H\times 2W}$, the SFS-FF module is formulated as:

$$F_S,\, F_F,\, \acute{F} = \mathrm{Conv}_{1\times 1}\Big(\mathrm{Concat}\Big[F_{\mathbf{in}},\ \mathrm{Down}\Big(\hat{F}_{\mathbf{in}}\Big)\Big]\Big)\,, \tag{10}$$

$$F_{\mathbf{out}} = \mathrm{Conv}_{1\times 1}\Big(\mathrm{Concat}\Big[\mathrm{IFFT}(\mathrm{Conv}_{3\times 3}(\mathrm{FFT}(F_S)))\,, \mathrm{SRM}(F_F), \acute{F}\Big]\Big)\,, \tag{11}$$

where F_S, F_F, and $\acute{F}$ denote the outputs of the 1×1 convolution, with channel allocations of $\frac{C}{2}$ for spatial refinement, $\frac{3C}{8}$ for frequency refinement, and $\frac{C}{8}$ for residual connection, respectively. $\mathrm{FFT}(\cdot)$ and $\mathrm{IFFT}(\cdot)$ denote the fast Fourier transform and the inverse fast Fourier transform, respectively. $\mathrm{SRM}(\cdot)$ denotes spatial refine module [24, 92], $\mathrm{Down}(\cdot)$ represents the proposed FSD-Down.

3.4 Frequency-Spatial Dynamic Downsampling

Conventional multi-scale feature fusion [24, 25, 35, 38, 70, 92] predominantly relies on convolution-based downsampling for spatial reduction, which inherently cou-

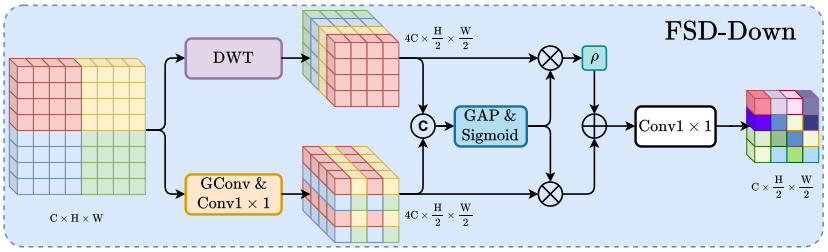


Fig. 5: Illustration of the Frequency-Spatial Dynamic Downsampling.

ples spatial aggregation with implicit frequency aliasing. Such operations introduce aliasing artifacts and attenuate high-frequency responses that are essential for the precise SOD. UAV-DETR [87] proposed a frequency focused downsampling module that applies average pooling followed by a fast Fourier transform to reduce spatial resolution. However, average pooling fundamentally behaves as a low-pass filter, smoothing local variations and suppressing high-frequency components prior to frequency transformation. As a result, fine-grained discriminative details may already be degraded before frequency modeling is applied. Therefore, an effective downsampling strategy should explicitly preserve informative high-frequency cues while maintaining spatial structural consistency during scale transition.

Building upon these observations and prior HWD [78] designs, we develop the FSD-Down module, as illustrated in Fig. 5, which leverages grouped convolution to parameterize wavelet-based decomposition. Specifically, the module employs grouped convolution to dynamically calibrate sub-band responses while preserving spatial structural alignment across feature levels. Given an input feature map $F_{\mathbf{in}} \in \mathbb{R}^{C\times H\times W}$, FSD-Down can be formulated as:

$$F_{\mathbf{CA}} = \sigma(\mathrm{GAP}(\mathrm{Concat}[\mathrm{DWT}(\mathrm{F}_{\mathbf{in}}),\ \mathrm{Conv}_{1\times1}\left(\mathrm{GConv}(\mathrm{F}_{\mathbf{in}})\right)])), \tag{12}$$

$$F_{\mathbf{out}} = \mathrm{Conv}_{1\times1}(\rho \odot \mathrm{F}_{\mathbf{CA}} \odot \mathrm{DWT}(\mathrm{F}_{\mathbf{in}}) + \mathrm{F}_{\mathbf{CA}} \odot \mathrm{Conv}_{1\times1}\left(\mathrm{GConv}(\mathrm{F}_{\mathbf{in}})\right)), \tag{13}$$

where $\mathrm{DWT}(\cdot)$ denotes the 2D discrete wavelet transform that decomposes the input into four sub-bands $\{F_{\mathbf{LL}}, F_{\mathbf{LH}}, F_{\mathbf{HL}}, F_{\mathbf{HH}}\}$ with spatial resolution $\frac{H}{2}\times\frac{W}{2}$, corresponding to the low-frequency approximation and horizontal, vertical, and diagonal high-frequency details, respectively. $\mathrm{GConv}\,(\cdot)$ represents a group convolution with kernel size and stride of 2 and C groups , GAP is the global average pooling, and $\sigma\,(\cdot)$ denotes the Sigmoid activation function. The parameter ρ is a learnable scaling factor that adaptively modulates frequency responses, enabling the model to emphasize frequency patterns and local contrasts that are most discriminative for SOD.

4 Experiments

4.1 Datasets

We perform quantitative evaluations on two widely adopted SOD benchmarks, VisDrone-DET2019 [15] and AITODv2 [77].

VisDrone-DET2019 consists of 14,018 images captured by UAVs, including 6,471 images for training, 548 for validation, and 3,190 for testing. The dataset provides bounding-box (BBox) annotations for 10 object categories and spans a wide range of scene densities, from sparse environments to highly congested scenarios. The distribution of object instances exhibits a heavy-tailed pattern, with an average of 40.7 objects per image and a standard deviation of 46.4, posing significant challenges for accurate detection.

AITODv2 contains 28,036 aerial images with a total of 752,745 annotated instances, divided into 11,214 training images, 2,804 validation images, and 14,018 testing images. The dataset provides BBox annotations for 8 object categories. It is particularly challenging due to the extremely small object scales: the average object size is only 12.7 pixels, with 86% of instances smaller than 16 pixels and even the largest objects not exceeding 64 pixels.

4.2 Implement Details

We train FSDC-DETR for 100 epochs with batch size 64. All input images are normalized and resized to 800×800. Our implementation is based on PyTorch 2.5.1, and all experiments conducted on 8×NVIDIA H200 GPUs. We optimize the model using AdamW with the learning rate of 5×10^{-4}, momentum 0.9, and weight decay 1.25×10^{-4}. Following DEIMv2 [24], we employ standard data augmentations such as color jitter and zoom-out, and additionally apply MixUp, Mosaic, and CopyBlend during training. We disable dense O2O [25] after the first 50% of training epochs, which consistently improves performance. We also turn off all data augmentation in the final two epochs. Additional hyper-parameter settings are provided in Supplement Sec. A.

4.3 Evaluation Metrics

To evaluate FSDC-DETR, we report Average Precision (AP) with at most 300 detections per image. Following the standard protocol, AP is computed by averaging results over IoU thresholds from 0.50 to 0.95 with a step size of 0.05. We also report scale-specific AP, i.e., AP_s, AP_m, and AP_l, for small, medium, and large objects, respectively.

4.4 Main Results

Quantitative comparisons on the test splits of VisDrone-DET2019 and AITODv2 are reported in Tab. 1. Consistent improvements over previous state-of-the-art detectors are achieved by the proposed FSDC-DETR, even when compared against their results reported at different model scales.

Table 1: Quantitative comparison on the official test splits of VisDrone-DET2019 and AITODv2 at input resolution 800×800.

Model	Params.	VisDrone-DET2019						AITODv2				
		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M
YOLOv11-S [28]	9.4	19.2	32.5	19.5	10.2	28.8	42.2	15.6	33.4	12.7	14.6	31.2
YOLOv12-S [61]	9.3	19.5	32.9	20.2	10.5	29.5	39.8	15.1	33.1	11.9	14.2	30.7
YOLOv13-S [29]	9.0	18.3	31.3	18.7	9.4	27.7	41.1	15.7	33.3	12.8	14.8	31.1
FBRT-YOLO-S [76]	2.9	19.5	32.9	20.2	10.1	30.0	39.9	14.3	31.1	11.2	13.2	28.8
D-FINE-S [43]	10.0	23.7	41.6	23.4	14.1	33.9	47.2	24.7	54.1	18.4	23.6	41.3
DEIMv2-S [24]	10.0	23.3	41.4	23.1	13.7	33.3	47.2	23.7	53.1	18.4	22.6	40.9
RT-DETRv4-S [35]	10.0	23.3	41.0	23.1	13.8	33.3	43.7	23.7	53.2	18.0	22.7	39.8
YOLOv11-L [28]	25.3	22.6	37.3	23.5	12.5	33.2	47.8	17.6	35.9	15.1	16.6	33.0
YOLOv12-L [61]	26.4	21.9	36.3	22.7	12.4	32.3	43.8	17.7	36.5	15.3	16.7	33.8
YOLOv13-L [29]	27.6	20.6	34.4	21.1	11.7	30.4	48.7	18.0	36.5	15.5	17.1	32.3
FBRT-YOLO-L [76]	14.6	22.5	37.4	23.6	12.3	34.1	46.4	16.8	35.0	14.0	15.7	32.5
D-FINE-L [43]	31.0	26.1	45.6	26.4	15.8	37.5	53.6	25.1	56.2	18.8	23.9	43.2
DEIMv2-L [24]	32.0	24.7	43.3	24.6	14.2	35.4	52.7	25.7	55.7	20.6	24.4	43.7
RT-DETRv4-L [35]	31.0	26.4	45.7	26.7	16.1	37.3	49.3	26.8	58.3	20.6	23.0	44.5
YOLOv11-X [28]	56.9	21.6	35.8	22.4	12.2	31.3	47.1	18.5	36.7	16.3	17.5	34.3
YOLOv12-X [61]	59.1	22.2	36.8	23.3	13.0	33.1	45.0	18.4	37.2	16.7	17.3	34.3
YOLOv13-X [29]	64.0	21.5	35.9	22.1	12.6	31.7	41.2	18.1	37.4	16.4	17.1	33.6
FBRT-YOLO-X [76]	22.8	22.9	37.6	23.7	12.7	34.6	46.8	17.2	35.7	14.4	16.2	33.0
D-FINE-X [43]	62.0	26.6	46.2	26.8	15.9	38.2	56.0	26.4	58.2	20.5	25.1	45.4
DEIMv2-X [24]	50.0	26.5	45.8	26.7	16.1	37.5	52.6	25.7	59.6	22.1	26.9	48.4
RT-DETRv4-X [35]	62.0	27.1	46.6	27.4	16.5	38.1	53.5	27.1	59.4	21.5	26.1	45.4
VRF-DETR [74]	13.5	23.1	40.3	22.9	14.1	32.8	38.1	16.4	39.1	12.1	16.4	27.9
CSFPR-DETR [21]	14.1	23.3	40.5	23.2	14.9	32.7	40.2	16.4	40.0	12.7	17.2	29.0
UAV-DETR-R18 [87]	20.0	24.2	42.0	24.1	15.0	34.2	43.8	16.7	37.9	11.8	15.7	28.1
UAV-DETR-R50 [87]	42.0	25.6	43.9	25.8	15.6	36.2	45.8	17.5	37.2	13.8	16.7	35.9
FSDC-DETR (Ours)	40.3	**31.1**	**52.3**	**31.9**	**21.0**	**41.8**	**58.1**	**32.3**	**63.9**	**28.7**	**31.3**	**48.5**

On VisDrone-DET2019, built upon the DEIMv2-L, FSDC-DETR achieves 31.1 AP, surpassing DEIMv2-L (24.7 AP) by 6.4 points. The improvement is more pronounced under AP_{50}, where our method reaches 52.3, outperforming DEIMv2-L and RT-DETRv4-L by 9.0 and 6.6 points, respectively. Under stricter localization criteria, AP_{75} increases from 24.6 to 31.9 (+7.3), indicating improved BBox precision. Notably, FSDC-DETR yields substantial gains on small objects. In VisDrone-DET2019, AP_S improves from 14.2 to 21.0, corresponding to a 6.8 point improvement, representing a relative improvement of nearly 48%. These results demonstrate that the proposed frequency-spatial collaborative modeling effectively preserves fine-grained structural details and high-frequency cues during multi-scale aggregation.

The superiority of our design is further validated on AITODv2. FSDC-DETR achieves 31.1 AP and 63.9 AP_{50}, surpassing DEIMv2-L by 6.4 AP and 8.2 AP_{50}, respectively, and outperforming RT-DETRv4-L by 4.7 AP. Moreover, AP_S increases from 24.4 to 31.3 (+6.9) for FSDC-DETR compared to DEIMv2-L, confirming consistent improvements on extremely small objects. Consistent gains across two challenging benchmarks indicate that FSDC-DETR generalizes robustly across object scales, benefiting not only small instances but also medium and large objects through coherent frequency-spatial interaction. More quanti-

tative results are provided in Supplement Sec B, including M-scale comparisons, an enhanced DEIMv2-L baseline with four multi-scale fusion layers, and cross-domain evaluations, further validate the robustness of FSDC-DETR.

4.5 Ablation Study

Table 2: Comprehensive ablation study on the VisDrone-DET2019 test split at input resolution 800×800.

DBFSAF	SFS-FF	FSD-Down	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
✗	✗	✗	24.7	43.3	24.6	14.2	35.4	52.7
✓	✗	✗	29.7	49.7	30.5	19.8	40.6	55.5
✗	✓	✗	25.9	44.9	26.0	15.5	36.8	53.7
✗	✗	✓	25.4	44.2	25.6	15.1	36.4	52.9
✗	✓	✓	26.3	45.5	26.4	16.4	37.4	54.5
✓	✗	✓	30.0	50.5	30.6	19.8	40.8	55.7
✓	✓	✗	30.3	50.9	30.8	20.0	41.3	56.0
✓	✓	✓	**31.1**	**52.3**	**31.9**	**21.0**	**41.8**	**58.1**

Ablation results on the VisDrone-DET2019 test split are reported in Tab. 2. The baseline (DEIMv2-L) achieves 24.7 AP and 43.3 AP_{50}. Adding DBFSAF raises performance to 29.7 AP (+5.0), with AP_{50} improving to 49.7 and AP_S to 19.8, demonstrating the benefit of frequency-spatial adaptive fusion. Incorporating SFS-FF further improves AP to 30.3, yielding consistent gains in AP_{75} and AP_S. Replacing conventional convolutional downsampling with the proposed FSD-Down results in the best performance of 31.1 AP and 52.3 AP_{50}, corresponding to overall improvements of 6.4 AP and 9.0 AP_{50} over the baseline. Notably, AP_S increases from 14.2 to 21.0, highlighting the effectiveness of frequency-preserving downsampling for SOD. These results verify that DBFSAF, SFS-FF, and FSD-Down contribute complementarily to performance gains. More ablation studies, module-wise parameter statistics, and frequency-response visualizations are provided in Supplement Sec C.

Performance improves as the partial ratio γ in DBFSAF, as defined in Eq. 1, increases from 0 to 0.5, achieving the best results at $\gamma = 0.5$ with 31.1 AP and 21.0 AP_S, as shown in Tab. 3. A further increase in γ slightly degrades performance, indicating that excessive refinement may introduce redundant interactions. The noticeable drop in $\gamma = 0$ further confirms the effectiveness of DBFSAF.

4.6 Visualization Analysis

Visualization comparisons between FSDC-DETR and competing models on the VisDrone-DET2019 and AITODv2 test splits are presented in Fig. 6 and Fig. 7.

Table 3: Comprehensive hyperparameter ablation study on the VisDrone-DET2019 test split at input resolution 800×800.

partial ratio γ in DBFSAF	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
0	29.7	49.7	30.5	19.8	40.6	56.5
0.375	30.8	51.7	31.3	20.4	41.2	57.5
0.5	**31.1**	**52.3**	**31.9**	**21.0**	**41.8**	**58.1**
0.625	30.9	51.8	31.4	20.5	41.2	57.7
1	30.0	50.4	30.7	20.0	41.0	56.4

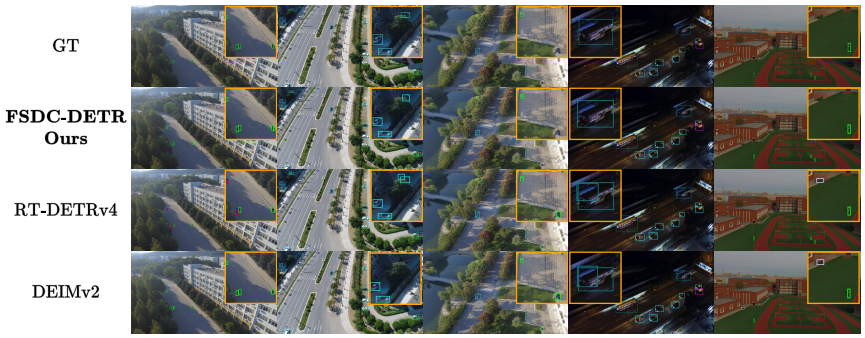


Fig. 6: Visualization comparison of small object detection results on the VisDrone-DET2019 test split. Orange boxes indicate zoomed-in details.

Benefiting from the integration of DBFSAF, SFS-FF, and FSD-Down, FSDC-DETR demonstrates superior capability in SOD.

On the VisDrone-DET2019 test split, FSDC-DETR exhibits more precise localization of small objects, as shown in the first two columns of Fig. 6, where object boundaries are clearer and object regions are more accurately aligned with ground truth (GT). As illustrated in columns 3-5, our model maintains robust detection under cluttered and complex backgrounds while effectively reducing false positives compared to competing methods.

On the AITODv2 test split, as shown in Fig. 7, FSDC-DETR achieves consistently more precise localization of tiny objects. In columns 1-2 and 5-7, small instances are more precisely detected and exhibit better alignment with the GT. In columns 3-4, the proposed method effectively suppresses false positives and maintains stable detection performance under challenging conditions, further demonstrating the robustness of the frequency-spatial collaborative modeling. Supplement Sec. D provides additional heatmap visualizations and failure case analyses of FSDC-DETR.

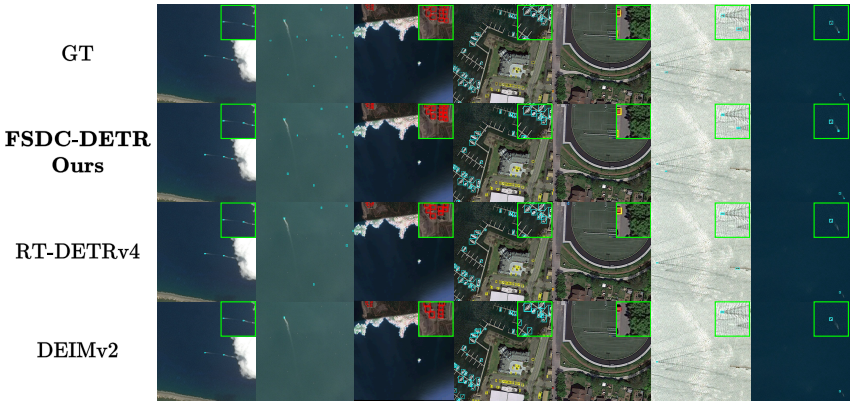


Fig. 7: Visualization comparison of small object detection results on the AITODv2 test split. Green boxes indicate zoomed-in details.

5 Conclusion

In this paper, we present FSDC-DETR for the precise SOD. Unlike conventional detectors that implicitly entangle spatial aggregation with frequency attenuation, FSDC-DETR explicitly constructs, propagates, and preserves frequency-aware representations throughout the detection pipeline. The proposed DBF-SAF module constructs frequency-aware representations by mitigating frequency aliasing and truncation in ViT-CNN backbones through adaptive high-frequency preservation and frequency-spatial collaborative refinement. Building upon these refined representations, the SFS-FF module propagates frequency-enriched features across scales via structure-aware frequency-spatial collaborative aggregation. Furthermore, the FSD-Down preserves frequency integrity during scale transitions while maintaining spatial consistency. By systematically addressing representation construction, cross-scale propagation, and scale-transition preservation from a unified frequency-spatial perspective, FSDC-DETR significantly improves SOD performance. Experimental results on two challenging benchmarks confirm state-of-the-art performance, demonstrating excellent detection capabilities for medium and large objects, with particularly significant improvements in SOD.

Acknowledgements

This work was supported in part by the Changzhou Longcheng Talent Program under Grant No. CQ20250117 and in part by the National Natural Science Foundation of China under Grant No. 62377011.

References

1. Abello, A.A., Hirata, R., Wang, Z.: Dissecting the high-frequency bias in convolutional neural networks. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 863–871
2. Bian, J., Feng, M., Dong, W., Wu, F., Luo, J., Wang, Y., Shi, G.: Feature information driven position gaussian distribution estimation for tiny object detection. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 30376–30386
3. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 213–229. Springer International Publishing, Cham
5. Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9630–9640
6. Chen, L., Fu, Y., Gu, L., Yan, C., Harada, T., Huang, G.: Frequency-aware feature fusion for dense image prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 10763–10780
7. Chen, L., Fu, Y., Gu, L., Zheng, D., Dai, J.: Spatial frequency modulation for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(11), 9767–9784
8. Chen, L., Gu, L., Fu, Y.: Frequency-dynamic attention modulation for dense prediction. In: 2025 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22620–22632
9. Chen, L., Gu, L., Li, L., Yan, C., Fu, Y.: Frequency dynamic convolution for dense image prediction. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 30178–30188
10. Chen, L., Gu, L., Zheng, D., Fu, Y.: Frequency-adaptive dilated convolution for semantic segmentation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3414–3425
11. Chen, Y., Yuan, X., Wang, J., Wu, R., Li, X., Hou, Q., et al.: Yolo-ms: Rethinking multi-scale representation learning for real-time object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(6), 4240–4252
12. Cheng, Y., Cao, Y., Yao, H., Luo, W., Jiang, C., Zhang, H., Shen, W.: A comprehensive survey for real-world industrial surface defect detection: Challenges, approaches, and prospects. Journal of Manufacturing Systems **84**, 152–172
13. Cong, R., Chen, Z., Fang, H., Kwong, S., Zhang, W.: Breaking barriers, localizing saliency: A large-scale benchmark and baseline for condition-constrained salient object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **48**(4), 4167–4183
14. Cui, C., Guo, R., Du, Y., He, D., Li, F., Wu, Z., Liu, Q., Wen, S., Huang, J., Hu, X., et al.: Beyond self-supervision: A simple yet effective network distillation alternative to improve backbones. arXiv preprint arXiv:2103.05959
15. Du, D., Zhu, P., Wen, L., Bian, X., Lin, H., Hu, Q., et al.: Visdrone-det2019: The vision meets drone object detection in image challenge results. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 213–226

16. Fan, W., Yang, Y., Qi, J., Zhang, Q., Liao, C., Wen, L., et al.: A deep-learning-based framework for identifying and localizing multiple abnormalities and assessing cardiomegaly in chest x-ray. Nature Communications **15**(1), 1347
17. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2917–2927
18. Girshick, R.: Fast r-cnn. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 1440–1448
19. Gonzalez, R.C.: Digital image processing. Pearson Education India
20. He, J., Erfani, S., Ma, X., Bailey, J., Chi, Y., Hua, X.S.: Alpha-iou: a family of power intersection over union losses for bounding box regression. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21, Curran Associates Inc., Red Hook, NY, USA
21. Hu, L., Yuan, J., Cheng, B., Xu, Q.: Csfpr-rtdetr: Real-time small object detection network for uav images based on cross-spatial-frequency domain and position relation. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–19
22. Hua, W., Chen, Q.: A survey of small object detection based on deep learning in aerial images. Artificial Intelligence Review **58**(6), 162
23. Huang, H., Xia, T., Ren, P., et al.: Partial channel network: Compute fewer, perform better. arXiv preprint arXiv:2502.01303
24. Huang, S., Hou, Y., Liu, L., Yu, X., Shen, X.: Real-time object detection meets dinov3. arXiv preprint arXiv:2509.20787
25. Huang, S., Lu, Z., Cun, X., Yu, Y., Zhou, X., Shen, X.: Deim: Detr with improved matching for fast convergence. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15162–15171
26. Huang, Z., Zhang, Z., Lan, C., Zha, Z.J., Lu, Y., Guo, B.: Adaptive frequency filters as efficient global token mixers. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6026–6036
27. Khanam, R., Hussain, M.: What is yolov5: A deep look into the internal features of the popular object detector. arXiv preprint arXiv:2407.20892
28. Khanam, R., Hussain, M.: Yolov11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725
29. Lei, M., Li, S., Wu, Y., Hu, H., Zhou, Y., Zheng, X., Ding, G., Du, S., Wu, Z., Gao, Y.: Yolov13: Real-time object detection with hypergraph-enhanced adaptive visual perception. arXiv preprint arXiv:2506.17733
30. Li, F., Zeng, A., Liu, S., Zhang, H., Li, H., Zhang, L., et al.: Lite detr : An interleaved multi-scale encoder for efficient detr. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18558–18567
31. Li, H., Wu, Z., Shao, R., Zhang, T., Fu, Y.: Noise calibration and spatial-frequency interactive network for stem image enhancement. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21287–21296
32. Li, J., Wu, H., Qin, J.: Weaveseg: Iterative contrast-weaving and spectral feature-refining for nuclei instance segmentation. In: 2025 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21984–21993
33. Li, K., Wang, D., Hu, Z., Zhu, W., Li, S., Wang, Q.: Unleashing channel potential: Space-frequency selection convolution for sar object detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17323–17332

34. Li, Y., Hou, Q., Zheng, Z., Cheng, M.M., Yang, J., Li, X.: Large selective kernel network for remote sensing object detection. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16748–16759
35. Liao, Z., Zhao, Y., Shan, X., Yan, Y., Liu, C., Lu, L., Ji, X., Chen, J.: Rt-detrv4: Painlessly furthering real-time object detection with vision foundation models. arXiv preprint arXiv:2510.25257
36. Liu, H., Jia, C., Shi, F., Cheng, X., Shi, M., Xie, X., et al.: Lidar: Lightweight adaptive cue-aware fusion vision mamba for multimodal segmentation of structural cracks. In: Proceedings of the 33rd ACM International Conference on Multimedia. p. 1832–1841. MM '25, Association for Computing Machinery, New York, NY, USA
37. Lv, S., Liang, T., Zhang, K., Jiang, S., Ouyang, B., Li, Q., Li, X.: A lightweight hierarchical aggregation task alignment network for industrial surface defect detection. Expert Systems with Applications **263**, 125727
38. Lv, W., Zhao, Y., Chang, Q., Huang, K., Wang, G., Liu, Y.: Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer. arXiv preprint arXiv:2407.17140
39. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., et al.: Conditional detr for fast training convergence. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3631–3640
40. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193
41. Park, N., Kim, S.: How do vision transformers work? arXiv preprint arXiv:2202.06709
42. Paul, S., Chen, P.Y.: Vision transformers are robust learners. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 36, pp. 2071–2081
43. Peng, Y., Li, H., Wu, P., Zhang, Y., Sun, X., Wu, F.: D-fine: Redefine regression task of detrs as fine-grained distribution refinement. In: International Conference on Learning Representations. vol. 2025, pp. 44015–44031
44. Pitas, I.: Digital image processing algorithms and applications. John Wiley & Sons
45. Qin, Z., Zhang, P., Wu, F., Li, X.: Fcanet: Frequency channel attention networks. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 763–772
46. Rahman, M.M., Marculescu, R.: Mk-unet: Multi-kernel lightweight cnn for medical image segmentation. In: 2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 1053–1062
47. Ranzinger, M., Heinrich, G., McCarthy, C., Kautz, J., Tao, A., Catanzaro, B., Molchanov, P.: C-radiov4 (tech report). arXiv preprint arXiv:2601.17237
48. Ren, K., Li, Z., Du, Y., Han, H., Wu, Y.: Fii-detr: Few-shot object detection with fully information interaction. Inf. Fusion **127**(PA)
49. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137–1149
50. Robinson, I., Robicheaux, P., Popov, M., Ramanan, D., Peri, N.: Rf-detr: neural architecture search for real-time detection transformers. arXiv preprint arXiv:2511.09554
51. Sapkota, R., Cheppally, R.H., Sharda, A., Karkee, M.: Yolo26: key architectural enhancements and performance benchmarking for real-time object detection. arXiv preprint arXiv:2509.25164

52. Sapkota, R., Karkee, M.: Object detection with multimodal large vision-language models: An in-depth review. Information Fusion **126**, 103575
53. Shi, T., Gong, J., Hu, J., Sun, Y., Bao, G., Zhang, P., et al.: Progressive class-aware instance enhancement for aircraft detection in remote sensing imagery. Pattern Recognition **164**, 111503
54. Shi, Z., Hu, J., Ren, J., Ye, H., Yuan, X., Ouyang, Y., He, J., Ji, B., Guo, J.: Hs-fpn: high frequency and spatial perception fpn for tiny object detection. In: Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'25/IAAI'25/EAAI'25, AAAI Press
55. Shu, Z., Wu, J., Yan, W., Liu, X., Zhang, H., Liu, C., Mao, Y., Chen, J.: Waveformer: Frequency-time decoupled vision modeling with wave equation. arXiv preprint arXiv:2601.08602
56. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., et al.: Dinov3. arXiv preprint arXiv:2508.10104
57. Sun, H., Lv, L., Zhang, P., Tang, T., Tian, F., Sun, W., Lu, H.: Spatial-frequency enhanced mamba for multi-modal image fusion. IEEE Transactions on Image Processing **34**, 7684–7696
58. Sun, H., Wang, R., Li, Y., Yang, L., Lin, S., Cao, X., Zhang, B.: Set: Spectral enhancement for tiny object detection. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4713–4723
59. Sun, Y., Xu, C., Yang, J., Xuan, H., Luo, L.: Frequency-spatial entanglement learning for camouflaged object detection. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 343–360. Springer Nature Switzerland, Cham
60. Tatsunami, Y., Taki, M.: Fft-based dynamic token mixer for vision. In: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'24/IAAI'24/EAAI'24, AAAI Press
61. Tian, Y., Ye, Q., Doermann, D.: Yolov12: Attention-centric real-time object detectors. In: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (eds.) Advances in Neural Information Processing Systems. vol. 38, pp. 78433–78457. Curran Associates, Inc.
62. Tong, Z., Chen, Y., Xu, Z., Yu, R.: Wise-iou: bounding box regression loss with dynamic focusing mechanism. arXiv preprint arXiv:2301.10051
63. Varghese, R., M., S.: Yolov8: A novel object detection algorithm with enhanced performance and robustness. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). pp. 1–6
64. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al.: Yolov10: real-time end-to-end object detection. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. NIPS '24, Curran Associates Inc., Red Hook, NY, USA
65. Wang, A., Liu, L., Chen, H., Lin, Z., Han, J., Ding, G.: Yoloe: Real-time seeing anything. In: 2025 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 24591–24602
66. Wang, C.Y., Yeh, I.H., Mark Liao, H.Y.: Yolov9: Learning what you want to learn using programmable gradient information. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 1–21. Springer Nature Switzerland, Cham

67. Wang, H., Wu, X., Huang, Z., Xing, E.P.: High-frequency component helps explain the generalization of convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8681–8691
68. Wang, J., Yang, W., Guo, H., Zhang, R., Xia, G.S.: Tiny object detection in aerial images. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 3791–3798
69. Wang, P., Zheng, W., Chen, T., Wang, Z.: Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. arXiv preprint arXiv:2203.05962
70. Wang, S., Xia, C., Lv, F., Shi, Y.: Rt-detrv3: Real-time end-to-end object detection with hierarchical dense positive supervision. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1628–1636
71. Wang, Y., Wang, H., Wang, L., Wang, X., Zhu, L., Lu, W., et al.: Complementary advantages: Exploiting cross-field frequency correlation for nir-assisted image denoising. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12679–12689
72. Wang, Z., Li, C., Xu, H., Zhu, X., Li, H.: Mamba yolo: a simple baseline for object detection with state space model. In: Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'25/IAAI'25/EAAI'25, AAAI Press
73. Wei, J., Li, Y., Fan, X., Ma, W., Qiu, M., Chen, H., Lei, W.: Sam-swin: Sam-driven dual-swin transformers with adaptive lesion enhancement for laryngo-pharyngeal tumor detection. Medical Image Analysis **109**, 103906
74. Wenbin, L.: An efficient aerial image detection with variable receptive fields. arXiv preprint arXiv:2504.15165
75. Xia, Q., zheng, L., Zhao, S., Huang, X., Wu, H., Wen, C., Wang, C.: Dota++: Unsupervisely and collaboratively detect objects from multi-agent observations with multi-modal prior constraints. IEEE Transactions on Pattern Analysis and Machine Intelligence **48**(7), 7467–7484
76. Xiao, Y., Xu, T., Xin, Y., Li, J.: Fbrt-yolo: Faster and better for real-time aerial image detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8673–8681
77. Xu, C., Wang, J., Yang, W., Yu, H., Yu, L., Xia, G.S.: Detecting tiny objects in aerial images: A normalized wasserstein distance and a new benchmark. ISPRS Journal of Photogrammetry and Remote Sensing **190**, 79–93
78. Xu, G., Liao, W., Zhang, X., Li, C., He, X., Wu, X.: Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. Pattern Recognition **143**, 109819
79. Yan, F., Jiang, X., Lu, Y., Cao, J., Chen, D., Xu, M.: Wavelet and prototype augmented query-based transformer for pixel-level surface defect detection. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23860–23869
80. Yang, C., Huang, Z., Wang, N.: Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13658–13667
81. Yang, J., Liu, S., Wu, J., Su, X., Hai, N., Huang, X.: Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection. In: Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence

and Fifteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'25/IAAI'25/EAAI'25, AAAI Press
82. Yin, D., Lopes, R.G., Shlens, J., Cubuk, E.D., Gilmer, J.: A fourier perspective on model robustness in computer vision. Curran Associates Inc., Red Hook, NY, USA
83. Zhan, J., Luo, Y., Guo, C., Wu, Y., Meng, J., Liu, J.: Yolopx: Anchor-free multi-task learning network for panoptic driving perception. Pattern Recognition **148**, 110152
84. Zhang, C.B., Zhong, Y., Han, K.: Mr. detr: Instructive multi-route training for detection transformers. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9933–9943
85. Zhang, F., Gu, Z., Wang, H.: Decoding with structured awareness: integrating directional, frequency-spatial, and structural attention for medical image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 12421–12429
86. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605
87. Zhang, H., Zhang, H., Liu, K., Gan, Z., Zhu, G.N.: Uav-detr: Efficient end-to-end object detection for unmanned aerial vehicle imagery. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 15143–15149
88. Zhang, X., Zhu, X., Tang, K., Zhao, Y., Lu, Z., Feng, Q.: Ddtnet: A dense dual-task network for tumor-infiltrating lymphocyte detection and segmentation in histopathological images of breast cancer. Medical Image Analysis **78**, 102415
89. Zhang, Y.F., Ren, W., Zhang, Z., Jia, Z., Wang, L., Tan, T.: Focal and efficient iou loss for accurate bounding box regression. Neurocomputing **506**, 146–157
90. Zhang, Z., Zhou, M., Wan, H., Li, M., Li, G., Han, D.: Idd-net: Industrial defect detection method based on deep-learning. Engineering Applications of Artificial Intelligence **123**, 106390
91. Zhao, C., Sun, Y., Wang, W., Chen, Q., Ding, E., Yang, Y., et al.: Ms-detr: Efficient detr training with mixed supervision. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17027–17036
92. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., et al.: Detrs beat yolos on real-time object detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16965–16974
93. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5122–5130