Adaptive Spectrum-Aware Feature Disentangled Network for Small Object Detection

Yang Guo[1,2], Zihan Yang[3], Feifei Kou[2], Yulan Hu[4], Ran Zhang[5], and Siyuan Yao[1⋆]

[1] Shenzhen Campus of Sun Yat-sen University
[2] Beijing University of Posts and Telecommunications, Beijing, China
[3] Hangzhou International Innovation Institute, Beihang University, Hangzhou, China
[4] Renmin University of China [5] China CITIC Bank

guoyang4409@gmail.com zihanyang@buaa.edu.cn koufeifei000@bupt.edu.cn
huyulan@ruc.edu.cn zhangran4@citicbank.com yaosiyuan04@gmail.com

Abstract. Small Object Detection (SOD) is a fundamental yet challenging problem in computer vision due to its limited spatial resolution and weak visual cues. Although recent approaches have achieved remarkable advances, the background distractors in different frequency spectra still degrade the performance. In this paper, we propose a novel small object detection framework termed **SFDNet**, which is capable of detecting small objects via efficient spectrum-aware feature disentanglement. Specifically, we propose an Adaptive Spectrum Disentanglement (ASD) module that decomposes backbone features into multiple complementary spectral components, aiming to construct discriminative object-relevant representations by discarding the background distractors for each component. Afterwards, to strengthen the semantic consistency of the similar objects in the same class, we propose a Class-Wise Prototype Distillation (CPD) procedure, which establishes class prototypes for the object instances and enforces the compact representation by efficient prototype distillation. Extensive experiments on multiple challenging benchmarks show that SFDNet outperforms existing state-of-the-art methods by a large margin. Our code is available at https://github.com/ManOfStory/SFDNet.

Keywords: Small Object Detection · Spectrum Disentanglement · Prototype Distillation

1 Introduction

Small Object Detection (SOD) is a critical research task aiming to accurately identify objects with limited size, which plays a vital role in a wide range of applications such as surveillance, drone-based inspection, and autonomous driving, etc. Compared to the classical object detection, the small objects are featured by fewer pixels and are prone to being confused with backgrounds, making it infeasible to adapt generic object detectors directly to the SOD task.

⋆ Corresponding Author

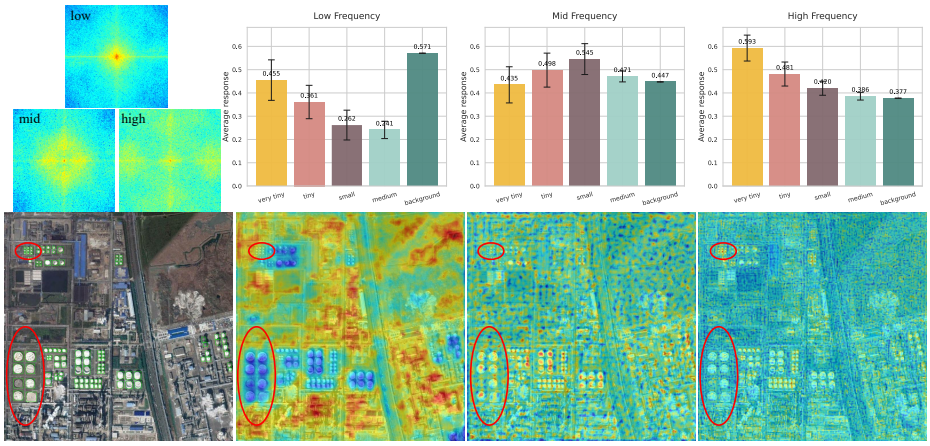


Fig. 1. The first row (left) illustrates the disentangled low, mid, and high frequency spectra, while the first row (right) reports the corresponding average response statistics within target regions at different scales. The second row presents the ground-truth annotations along with the heatmaps derived from low, mid, and high frequency features, respectively. Please zoom in for details. Regions highlighted by red circles emphasize salient differences.

From a technical perspective, existing small object detectors [49,50,21,58] primarily focus on learning semantically discriminative features to mitigate spatial information attenuation. Despite remarkable progress, these methods still struggle to reliably detect small objects in extremely complex scenarios. First, due to the limited spatial extent and weak visual cues of small objects, current detectors are easily confused by surrounding distractors. The most advanced approaches [35,37,20] attempt to alleviate this issue by suppressing noise in specific frequency bands, thus the models are encouraged to pay more attention to the informative small-object features. However, these methods overlook the fact that noise is distributed across the entire spectrum, which is crucial for robust small object detection. As shown in Fig. 1, different frequency spectra exhibit distinct sensitivities to various scales of objects: the low-frequency spectrum tends to emphasize background structures and large-scale distractors, the mid-frequency spectrum demonstrates a wider range of response to the objects from tiny to medium-sized targets, while the high-frequency spectrum is particularly sensitive to extremely tiny objects. Due to the absence of comprehensive full-spectrum modeling, the performance of existing SOD methods[55,35,24] is inherently constrained. Besides, many small objects share similar appearance semantics, while the intrinsic relationship information of these objects has not been sufficiently explored. Although some methods, e.g., [43,17,55,52], attempt to enhance the appearance representation of small objects via super-resolution, feature mimic learning, or memory-based retrieval mechanisms, they still fail to effectively exploit the semantic similarity among objects. These approaches typically utilize the grouped samples with predefined metrics, e.g., IoU or cosine

similarity, to construct instance-level relationships, but the generalization ability is still limited due to the lack of class-wise supervision.

To address these challenges, we propose SFDNet, which adaptively disentangles features into multiple complementary spectral components and enhances each component with tailored scanning strategies to better capture frequency-specific cues. By adaptively regulating the contributions of different spectrum components, SFDNet explicitly separates object-relevant information from background distractors within different spectra to facilitate accurate small object detection. In addition, we introduce a Class-Wise Prototype Distillation (CPD) procedure to model the shared appearance semantics among small objects within the same category. Through the distillation of class-specific semantic prototypes, CPD promotes compact and discriminative feature representations for each class. Extensive experiments on several widely used benchmarks demonstrate the effectiveness of the proposed method. In summary, the main contributions of this paper are as follows:

- We propose SFDNet, a novel small object detection framework that introduces an Adaptive Spectrum Disentanglement (ASD) module to adaptively disentangle features into multiple complementary spectral components, effectively mitigating background distractors.
- We introduce a Class-Wise Prototype Distillation (CPD) procedure, which performs category-aware prototype distillation to enforce compact representations, leading to more accurate and robust bounding box estimation.
- Extensive experiments on several popular benchmarks demonstrate the effectiveness and robustness of SFDNet in detecting extremely small objects under both densely and sparsely distributed scenarios.

2 Related Work

2.1 Small Object Detection

Recent advances in small object detection primarily focus on three aspects: data augmentation, label assignment, and feature enhancement. For data augmentation, [18,5] augment small object instances via random copy-paste. DS-GAN [3] utilizes a GAN-based generator to produce high-quality synthetic training samples. Fang et al. [12] propose a controllable diffusion-based augmentation method to generate high-quality training samples. For label assignment, NWD-RKA [46] and RFLA [47] leverage Wasserstein distance and KL divergence to mitigate the scale-variant IoU issue and increase positive samples for small objects in the training phase, while CFINet [55] adopts a coarse-to-fine RPN with dynamic IoU thresholds. It also generates more effective positive samples across object scales. For feature enhancement, DN-FPN [24] employs contrastive learning to suppress noise in the top-down pathway of FPN by enhancing the discriminability of features at each level. HS-FPN [35] first introduces a High-Frequency Perception (HFP) module to suppress low-frequency components that often dominate small

object features. It also proposes a Spatial Dependency Perception (SDP) module to reduce the feature misalignment across FPN levels. However, its frequency filtering strategy relies on manually designed hyperparameters, which limit its ability to adaptively suppress background distractors in complex scenes.

Different from previous works, our method performs adaptive spectrum disentanglement by decomposing representations into different complementary spectral components, which explicitly separates object-relevant signals from background distractors, enabling accurate small object detection.

2.2 Spectral-Aware Representation

Spectral modeling has recently gained increasing attention in computer vision. Recent advances have incorporated spectral priors into convolutional architectures to enhance feature representation. For instance, FADC [7] adaptively adjusts dilation rates based on local spectral characteristics, while FDConv [6] reformulates convolutional filters in the Fourier domain to enrich filter diversity and representation capacity. Beyond convolutional adaptations, spectral decomposition has also been integrated into sequence modeling and robustness-oriented frameworks. TinyVim [29] introduces a Laplace-based mixer to decouple spectral components, strengthening visual representation learning within sequence modeling architectures. Bi et al. [1] leverage the Discrete Sine Transform to partition features into illumination-sensitive and illumination-insensitive spectral bands, thereby improving robustness under illumination variations. Furthermore, Wang et al. [42] employ spectral decomposition to disentangle domain-invariant and domain-specific representations, enhancing cross-domain generalization in UAV-based object detection.

In this study, we disentangle backbone features into multiple complementary spectral components and conduct spectrum-specific aggregation within each frequency spectrum. By adaptively fusing the resulting spectral representations, we obtain more discriminative features tailored for small object detection.

2.3 State Space Model

State Space Models (SSMs), originally developed in control theory, have recently been incorporated into deep learning to model sequential data by mapping a one-dimensional function or sequence to another sequence through an implicit latent state. However, classical SSMs are constrained by Linear Time Invariance (LTI), which imposes fundamental limitations on their capacity to model complex and input-dependent patterns. To address this issue, Mamba [13] introduces a Selective State Space Model that removes the LTI constraint while preserving linear-time efficiency. Building upon Mamba, [26,59] adapts SSMs from one-dimensional sequences to two-dimensional visual data by employing multiple scanning routes to convert 2D feature maps into 1D sequences. MambaVision [15] further integrates Mamba with Vision Transformers (ViTs) by redesigning the Mamba formulation to enhance its effectiveness in modeling visual representations. More recently, Spatial-Mamba [44] observes that sequential

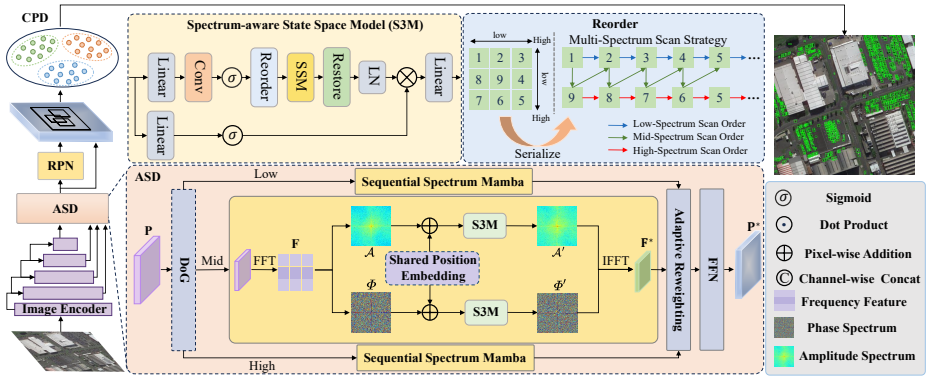


Fig. 2. The proposed architecture consists of two core components: the Adaptive Spectrum Disentanglement (ASD) module and the Class-Wise Prototype Distillation (CPD) procedure. The ASD module disentangles features into multi-spectrum components to perform full-spectrum suppression of background noise. The CPD procedure distills class-wise prototype representations, promoting compact and discriminative feature embeddings for each class.

scanning along any fixed direction inevitably disrupts spatial structure, and thus incorporates local spatial neighborhood features into each state to preserve the inherent 2D spatial relationships. In this work, we utilize SSMs in spectrum modeling and propose a multi-spectrum scan strategy.

3 Methodology

In this section, we present the overall architecture of SFDNet. As illustrated in Fig. 2, the framework consists of two key components: an Adaptive Spectrum Disentanglement (ASD) module and a Class-Wise Prototype Distillation (CPD) procedure. The ASD module decomposes backbone features into multiple complementary spectral components and models the contextual dependencies of each spectrum under a spectrum-specific scanning strategy. The CPD procedure constructs class-wise prototypes and promotes compact and discriminative feature representations by exploiting intra-class similarity, thereby enhancing small object detection performance.

3.1 Adaptive Spectrum Disentanglement

Given an input image $\mathbf{I}_t$, we first feed it into a vision backbone to extract multi-scale feature maps $\{\mathbf{C}_2, \mathbf{C}_3, \mathbf{C}_4, \mathbf{C}_5\}$. These features are subsequently processed by a standard Feature Pyramid Network (FPN) to generate top-down aggregated multi-scale representations $\{\mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4, \mathbf{P}_5\}$. Without loss of generality, we denote an arbitrary pyramid level as $\mathbf{P}$. The aggregated feature $\mathbf{P}$ is then

fed into the proposed Adaptive Spectrum Disentanglement (ASD) module for spectrum-aware processing.

Specifically, the ASD module first disentangles the feature $\mathbf{P}$ into low, mid, and high frequency components using the Difference of Gaussian (DoG) approach, formulated as follows:

$$\begin{aligned}
\mathbf{P}_{\text{low}} &= \mathcal{F}^{-1}\big[\mathcal{F}(\mathbf{P}) \cdot \hat{G}(\omega, k\sigma)\big], \\
\mathbf{P}_{\text{mid}} &= \mathcal{F}^{-1}\big[\mathcal{F}(\mathbf{P}) \cdot (\hat{G}(\omega, \sigma) - \hat{G}(\omega, k\sigma))\big], \\
\mathbf{P}_{\text{high}} &= \mathcal{F}^{-1}\big[\mathcal{F}(\mathbf{P}) \cdot (1 - \hat{G}(\omega, \sigma))\big],
\end{aligned} \tag{1}$$

where $\hat{G}(\omega, \sigma)$ denotes the frequency-domain Gaussian filter with standard deviation σ, $\mathcal{F}$ and $\mathcal{F}^{-1}$ represent the Fourier transform and inverse Fourier transform, $\mathbf{P}_{\text{low}}$, $\mathbf{P}_{\text{mid}}$, and $\mathbf{P}_{\text{high}}$ correspond to the disentangled features of different frequency, respectively.

The disentangled features are subsequently fed into the Sequential Spectrum Mamba module for spectrum-aware contextual modeling. Supposing $\hat{\mathbf{P}} \in \{\hat{\mathbf{P}}_{\text{low}}, \hat{\mathbf{P}}_{\text{mid}}, \hat{\mathbf{P}}_{\text{high}}\}$ to uniformly denote the frequency-specific features of each spectrum. Specifically, $\hat{\mathbf{P}}$ is first transformed into the frequency domain via the Fast Fourier Transform (FFT), yielding the corresponding frequency-domain representation $\mathbf{F}$. The FFT is formulated as follows:

$$\mathbf{F}(u, v) = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} \hat{\mathbf{P}}(x, y) \cdot \exp\left(-j2\pi \left(\frac{ux}{H} + \frac{vy}{W}\right)\right), \tag{2}$$

where $\hat{\mathbf{P}}(x, y)$ denotes the spatial domain feature at location (x, y), $\mathbf{F}(u, v)$ is its corresponding frequency representation at coordinate (u, v). Afterwards, the frequency features $\mathbf{F}$ are then decomposed into the amplitude spectrum $\mathcal{A}$ and phase spectrum Φ, which can be computed as follows:

$$\begin{aligned}
\mathcal{A}(u, v) &= \sqrt{\text{Re}(\mathbf{F}(u, v))^2 + \text{Im}(\mathbf{F}(u, v))^2}, \\
\Phi(u,v) &= \arctan\left(\frac{\text{Im}(\mathbf{F}(u, v))}{\text{Re}(\mathbf{F}(u, v))}\right),
\end{aligned} \tag{3}$$

where $\text{Re}(\cdot)$ and $\text{Im}(\cdot)$ denote the real and imaginary parts of the complex-valued feature, respectively. The amplitude spectrum $\mathcal{A}$ and phase spectrum Φ are then added with a shared learnable positional embedding, respectively. The resulting features are then fed into Sequential State Space Models (S3M), which aggregate information within different spectra to capture long-range spectral dependencies, as detailed in Sec. 3.2. The complete process is formulated as follows:

$$\begin{aligned}
\mathcal{A}'(u, v) &= \text{S3M}(\mathcal{A}(u, v) + \text{PE}(u, v)), \\
\Phi'(u, v) &= \text{S3M}(\Phi(u, v) + \text{PE}(u, v)),
\end{aligned} \tag{4}$$

where $\mathcal{A}' \in \mathbb{R}^{C \times H \times W}$ and $\Phi' \in \mathbb{R}^{C \times H \times W}$ denote the modified amplitude and phase spectra, respectively. The refined amplitude and phase spectra are then

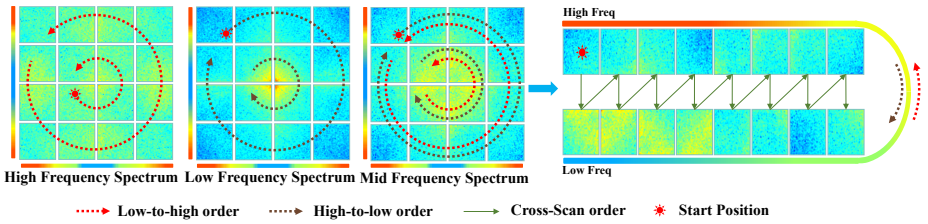


Fig. 3. Illustration of the Multi-Spectrum Scan Strategy. The high and low frequency spectra follow low-to-high and high-to-low scanning orders, respectively, whereas the mid-frequency spectrum adopts an alternating cross-scan strategy.

transformed back into the spatial domain via the inverse Fast Fourier Transform (IFFT), yielding the enhanced feature representation $\mathbf{F}^*$.

To effectively integrate information from different spectra, the spectrum-aware disentangled features are adaptively reweighted according to their learned importance vectors $\boldsymbol{\alpha}_{\mathrm{L}}$, $\boldsymbol{\alpha}_{\mathrm{M}}$, and $\boldsymbol{\alpha}_{\mathrm{H}} \in \mathbb{R}^C$. The reweighted features are subsequently fused through a Feed-Forward Network (FFN) to produce the aggregated representation $\mathbf{P}^*$.

$$\mathbf{P}^* = \mathrm{FFN}\big([\boldsymbol{\alpha}_{\mathrm{L}}, \boldsymbol{\alpha}_{\mathrm{M}}, \boldsymbol{\alpha}_{\mathrm{H}}][\mathbf{F}^*_{\mathrm{low}}, \mathbf{F}^*_{\mathrm{mid}}, \mathbf{F}^*_{\mathrm{high}}]^{\mathrm{T}}\big), \tag{5}$$

The fused representation $\mathbf{P}^*$ is subsequently utilized for downstream detection, serving as a spectrum-adaptively enhanced feature embedding for small object detection.

3.2 Spectrum-aware State Space Model (S3M)

As illustrated in Fig. 2, we propose a Multi-Spectrum Scan Strategy that serializes each spectrum map into a one-dimensional sequence, enabling long-range contextual modeling over the entire spectrum. Since State Space Models (SSMs) process data sequentially by propagating historical information through hidden states, the state vector is inherently dominated by nearby elements and less affected by the remote ones. When modeling along the spectral dimension, this order-sensitive processing induces spectral biases that depend on the chosen scanning order.

To effectively aggregate information emphasized by different spectra, we leverage this property of SSMs to design spectrum-specific scanning strategies tailored to the characteristics of each frequency band. As shown in Fig. 3, for the high-frequency spectrum, we adopt a low-to-high scanning order that unfolds the spectrum in a spiral manner from the spectral center outward along increasing radii, biasing the aggregation toward high-frequency components. Conversely, for the low-frequency spectrum, we follow a high-to-low scanning order, such that low-frequency components dominate the resulting representation. For the mid-frequency spectrum, we employ an alternating cross-scan strategy that interleaves inner and outer spectral regions, mixing low and high frequency signals

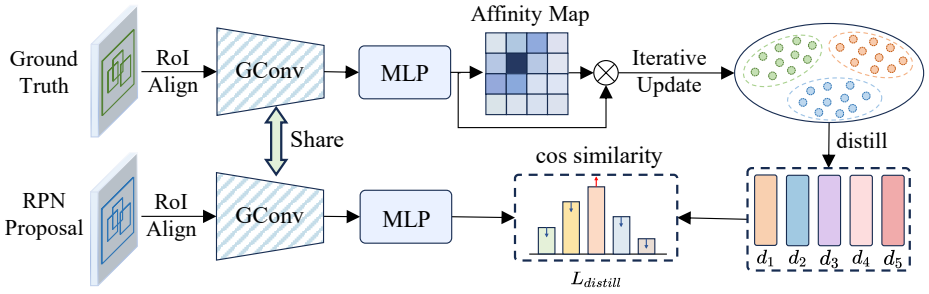


Fig. 4. Overview of the Class-Wise Prototype Distillation (CPD) procedure. CPD distills class-specific prototypes from ground-truth RoI features, which are then used to guide proposal features via contrastive supervision, enhancing semantic consistency within each category.

during serialization to mitigate the inherent inductive bias of SSMs and promote balanced contextual modeling across the spectrum.

The reordered sequences are subsequently fed into the SSM to capture long-range dependencies across frequency signals. Following the discretized formulation of Mamba, the state transition is defined as

$$\begin{aligned} \bar{\mathbf{A}}_t &= \exp(\Delta_t \mathbf{A}) \\ \bar{\mathbf{B}}_t &= (\Delta_t \mathbf{A})^{-1}(\exp(\Delta_t \mathbf{A}) - \mathbf{I})\Delta_t \mathbf{B_t} \\ \mathbf{h}_t &= \bar{\mathbf{A}}_t \mathbf{h}_{t-1} + \bar{\mathbf{B}}_t \mathbf{x}_t, \quad \mathbf{y}_t = \mathbf{C_t} \mathbf{h}_t, \end{aligned} \tag{6}$$

where $\mathbf{A}$ denotes a learnable input-invariant state transition embedding, while $\mathbf{B}_t$, $\mathbf{C}_t$, and Δ_t are input-variant parameters dynamically generated from the input $\mathbf{x}_t$. The hidden state $\mathbf{h}_t$ represents the latent state at step t, and $\bar{\mathbf{A}}_t$ and $\bar{\mathbf{B}}_t$ are the corresponding discretized state transition matrices that enable selective and input-dependent state evolution in the SSM.

Finally, the aggregated spectrum sequences are restored to their original spatial layout via the inverse permutation, ensuring that the underlying frequency basis remains invariant before and after the transformation.

3.3 Class-Wise Prototype Distillation

To better capture the relational semantics among intra-class instances, as shown in Fig.4, we introduce a Class-Wise Prototype Distillation (CPD) procedure that explicitly constructs category-level prototypes from ground-truth annotations. By grouping the prototype representation, the hard samples related to the same class can be mutually enhanced, leading to more discriminative and category-consistent representations.

During training, we extract RoI features from ground-truth bounding boxes, collapse the spatial dimensions using a Global Convolution layer, and project them through an MLP to obtain embedding vectors $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]^\top$. Meanwhile, we establish a class-wise prototype bank $\{\mathbf{d}_c \in \mathbb{R}^C\}_{c=1}^{N_C}$ in the embedding

space. At initialization, each prototype $\mathbf{d}_c^{(0)}$ is computed as the empirical mean of the ground-truth RoI embeddings belonging to class c. To further refine these prototypes and capture intra-class variations, we formulate prototype distillation as a class-conditional, affinity-guided refinement procedure inspired by the hierarchical affinity modeling in [53] and the expectation-maximization paradigm. For each class c, the class-aware affinity and the refined prototype are defined as:

$$\mathbf{A}_{i,c} = \frac{\exp\left(\kappa \mathbf{z}_i^\top \mathbf{d}_c\right) \mathbb{1}(i \in \Omega_c^{gt})}{\sum_{j=1}^{N} \exp\left(\kappa \mathbf{z}_j^\top \mathbf{d}_c\right) \mathbb{1}(j \in \Omega_c^{gt})}, \quad \mathbf{d}_c^* = \sum_{i=1}^{N} \mathbf{A}_{i,c} \mathbf{z}_i, \tag{7}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function, Ω_c^{gt} is the index set of ground-truth instances of class c in the current mini-batch, and κ controls the concentration of the similarity distribution. Afterwards, the prototype bank is updated via a momentum scheme across minibatches to preserve global stability:

$$\mathbf{d}_c \leftarrow \alpha \mathbf{d}_c + (1 - \alpha)\mathbf{d}_c^*. \tag{8}$$

With the refined prototype bank, we further impose contrastive alignment on proposal features. Given object proposals generated by the RPN, we extract RoI features using RoI Align from $\mathbf{P}^*$ and project them into the same embedding space, yielding instance representations $\mathbf{B} \in \mathbb{R}^{N \times C}$. To align these proposals with their corresponding class prototypes, we adopt an InfoNCE-based contrastive objective. For each embedding $\mathbf{b}_i \in \mathbf{B}$ belonging to class c, its similarity to prototype $\mathbf{d}_c$ is maximized against other prototypes $\{\mathbf{d}_k\}_{k=1}^{N_c}$. The loss is formulated as:

$$\mathcal{L}_{\text{distill}} = -\frac{1}{N} \sum_{c=1}^{N_c} \sum_{i \in \Omega_c} \log \frac{\exp\left(\cos(\mathbf{b}_i, \mathbf{d}_c)/\tau\right)}{\sum_{k=1}^{N_c} \exp\left(\cos(\mathbf{b}_i, \mathbf{d}_k)/\tau\right)}, \tag{9}$$

where N_c denotes the number of categories, Ω_c is the index set of RoIs belonging to class c, $\cos(\cdot, \cdot)$ denotes cosine similarity, and τ is a temperature parameter. Finally, the overall loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{IoU}} + \mathcal{L}_{\text{cls}} + \gamma \mathcal{L}_{\text{distill}}, \tag{10}$$

where $\mathcal{L}_{\text{IoU}}$ denotes the IoU loss. $\mathcal{L}_{\text{cls}}$ is the cross-entropy classification loss for candidate proposals. $\mathcal{L}_{distill}$ is the distillation loss. The coefficient γ is a hyperparameter that balances the loss contribution.

4 Experiment

In this section, we first provide the implementation details of our method and comparisons with other state-of-the-art detectors on three popular benchmarks: **AI-TOD** [41], **SODA-D** [9] and **SODA-A** [9]. Next, we conduct ablation studies to evaluate the effectiveness of our approach. Finally, we present qualitative visualizations to analyze the impact on feature representations.

Table 1. Comparison with state-of-the-art detection methods on AI-TOD test set. All metrics follow the COCO-style evaluation protocol, including overall AP and performance across different scene types. **Bold** numbers indicate the best results. † denotes results trained using the official implementation.

Method	Source	AP	AP_{50}	AP_{75}	AP_{vt}	AP_t	AP_s	AP_m
End-to-End Detectors								
DAB-DETR[25]	ICLR2022	4.9	16.0	1.7	1.7	3.6	7.0	18.0
DAB-Deformable-DETR[25]	ICLR2022	16.5	42.6	9.9	7.9	15.2	23.8	31.9
DINO-Deformable-DETR[56]	ICLR2023	23.2	56.6	15.4	9.9	23.1	29.3	37.6
DINO-5scale w/SET [37]	CVPR2025	26.6	57.1	20.8	13.2	27.1	31.5	–
Multi-Stage Detectors								
Faster R-CNN [33]	PAMI2017	11.1	26.3	7.6	0.0	7.2	23.3	33.6
Cascade R-CNN [4]	CVPR2018	13.8	30.8	10.5	0.0	10.5	25.5	36.6
DetectorRS[32]	CVPR2021	14.8	32.8	11.4	0.0	10.8	28.3	38.0
QueryDet[50]	CVPR2022	12.2	29.3	7.9	2.4	10.5	18.5	26.3
CFINet†[55]	ICCV2023	24.7	53.9	18.6	11.7	26.4	28.1	32.2
RFLA[47]	ECCV2022	24.8	55.2	18.5	9.3	24.8	30.3	38.2
DNTR[24]	TGRS2024	26.2	56.7	20.2	12.8	26.4	31.0	37.0
SimD[34]	IROS2024	26.6	55.9	21.2	13.4	27.5	30.9	37.8
DetectorRS w/FIDP [2]	CVPR2025	24.3	54.4	18.3	8.5	24.9	29.8	–
HS-FPN[35]	AAAI2025	25.1	55.7	19.1	12.1	25.3	29.9	36.9
SFDNet	–	29.0	57.9	23.5	14.4	29.0	33.4	40.8
SFDNet*	–	**31.7**	**64.9**	**25.6**	**17.6**	**32.8**	**36.1**	**42.5**

4.1 Implementation Details

All experiments are conducted on a single NVIDIA A100 GPU. For AI-TOD, input patches are fixed to a size of 800×800, whereas for SODA-D and SODA-A, the patches are resized to 1200×1200. The hyperparameters τ and γ are set to 0.07 and 0.1. The DoG parameters are set as $\sigma = 1.0$ and $k = 1.414$. The model is trained for 36 epochs on AI-TOD and 12 epochs on SODA-D and SODA-A. Based on different backbones, we propose two variants of SFDNet: SFDNet and SFDNet*. SFDNet adopts a ResNet-50[16] backbone, while SFDNet* employs a Spatial-Mamba[44] backbone.

4.2 Comparison with state-of-the-art

AI-TOD. The AI-TOD dataset contains 700,621 instances in 8 categories in 28,036 aerial images. Compared to existing object detection datasets in aerial images, the mean size of objects in AI-TOD is about 12.8 pixels, which is much smaller than others. We report the results of our proposed SFDNet on AI-TOD test set. As shown in Table 1, SFDNet* achieves the best performance across all metrics, surpassing other state-of-the-art methods by a large margin. Compared with HS-FPN [35], it improves (AP, AP_{50}, AP_{75}, AP_{vt}, AP_t, AP_s, AP_m) by (6.6%, 9.2%, 6.5%, 5.5%, 7.5%, 6.2%, 5.6%), respectively. Besides, SFDNet also achieves state-of-the-art performance.

SODA-A. The SODA-A dataset consists of 2,513 aerial images with 872,069 oriented bounding box annotations, focusing on small object detection. SODA-A has an average resolution of 4761 × 2777 pixels and contains approximately

Table 2. Comparison with oriented object detection methods on SODA-A test set. Metrics follow COCO-style evaluation. **Bold** numbers indicate the best results. † denotes results trained using the official implementation.

Method	Source	AP	AP_{50}	AP_{75}	AP_{es}	AP_{rs}	AP_{gs}	AP_N
Rotated Faster R-CNN[33]	PAMI2017	32.5	70.1	24.3	11.9	27.3	42.2	34.4
Rotated RetinaNet[23]	ICCV2017	26.8	63.4	16.2	9.1	22.0	35.4	28.2
Gliding Vertex[49]	PAMI2020	31.7	70.8	22.6	11.7	27.0	41.1	33.8
Oriented R-CNN[45]	ICCV2021	34.4	70.7	28.6	12.5	28.6	44.5	36.7
S^2A-Net[14]	TGRS2021	28.3	69.6	13.1	10.2	22.8	35.8	29.5
DODet[8]	TGRS2022	31.6	68.1	23.4	11.3	26.3	41.0	33.5
Oriented RepPoints[21]	CVPR2022	26.3	58.8	19.0	9.4	22.6	32.4	28.5
DHRec[31]	PAMI2022	30.1	68.8	19.8	10.6	24.6	40.3	34.6
CFINet[55]	ICCV2023	34.4	73.1	26.1	13.5	29.3	44.0	35.9
DecoupleNet†[28]	TGRS2024	36.6	71.3	33.3	12.2	31.0	47.7	40.2
LEGNet†[27]	ICCVW2025	29.6	58.7	26.4	10.4	26.0	39.6	32.0
GauCho† [30]	CVPR2025	33.2	70.1	25.0	9.9	27.8	44.9	36.4
UGS[36]	ICCV2025	36.0	73.1	30.3	13.7	30.2	47.8	38.1
DCFL[48]	PAMI2026	36.6	72.6	32.4	13.9	30.3	47.4	41.2
Unc-SOD[54]	TIP2026	34.8	73.6	26.4	13.8	29.7	44.7	36.5
SFDNet	–	37.8	73.1	34.5	13.1	32.1	49.4	42.9
SFDNet*	–	**39.2**	**75.0**	**36.6**	**15.5**	**33.7**	**50.4**	**45.6**

347 instances per image, presenting a highly dense object distribution that challenges existing detection models in clustered scenes. Table 2 summarizes the performance on SODA-A benchmark. SFDNet* achieves state-of-the-art results with (39.2%, 75.0%, 36.6%, 15.5%, 33.7%, 50.4%, 45.6%) on (AP, AP_{50}, AP_{75}, AP_{es} AP_{rs}, AP_{gs}, $AP_{\mathbf{N}}$), surpassing state-of-the-art GauCho [30] by (6.0%, 4.9%, 11.6%, 5.6%, 5.9%, 5.5%, 9.2%) across all metrics.

SODA-D. The SODA-D dataset contains 24,828 high-resolution images with 278,433 annotated instances spanning 9 categories. It exhibits rich diversity in terms of locations, weather conditions, period, camera viewpoints, and traffic scenarios. With an average image resolution of 3407 × 2470, the dataset is particularly well-suited for detecting tiny and small objects in complex environments. We report the performance of SFDNet on SODA-D test set. As summarized in Table 3, our SFDNet* achieves state-of-the-art performance across all metrics, with an average AP of 34.2%. Compared with HS-FPN[35], SFDNet* yields consistent improvements of (4.6%, 7.2%, 4.6%, 3.1%, 4.3%, 5.3%, 5.1%) in terms of (AP, AP_{50}, AP_{75}, AP_{es}, AP_{rs}, AP_{gs}, AP_{N}).

4.3 Ablation Study

Ablation of Different Variations. To evaluate the effectiveness of each component, we selectively ablate ASD and CPD from SFDNet. As shown in Table 4, the baseline detector without ASD and CPD achieves (24.8%, 9.3%, 24.8%, 30.3%, 38.2%) in terms of (AP, AP_{vt}, AP_{t}, AP_{s}, AP_{m}), respectively. When ASD is introduced, the performance improves to (28.6%, 13.8%, 28.8%, 33.2%, 39.2%), demonstrating its capability to effectively suppress noise across the full spectrum. Additionally, incorporating CPD into the baseline also leads to per-

Table 3. Comparison with state-of-the-art detection approaches on SODA-D test set. All metrics follow COCO-style evaluation, including overall AP and performance across different scene types. **Bold** numbers indicate the best results. † denotes results trained using the official implementation.

Method	Source	AP	AP_{50}	AP_{75}	AP_{es}	AP_{rs}	AP_{gs}	AP_N
Faster R-CNN[33]	PAMI2017	28.9	59.7	24.2	13.9	25.6	34.3	43.2
RetinaNet[23]	ICCV2017	28.2	57.6	23.7	11.9	25.2	34.1	44.2
CornerNet[19]	ECCV2018	24.6	49.5	21.7	6.5	20.5	32.2	43.8
FCOS[39]	ICCV2019	23.9	49.5	19.9	6.9	19.4	30.9	40.9
RepPoints[51]	ICCV2019	28.0	55.6	24.7	10.1	23.8	35.1	45.3
ATSS[57]	CVPR2020	26.8	55.6	22.1	11.7	23.9	32.2	41.3
Cascade RPN[40]	NIPS2019	29.1	56.5	25.9	12.5	25.5	35.4	44.7
Deformable-DETR[60]	ICLR2021	19.2	44.8	13.7	6.3	15.4	24.9	34.2
Sparse R-CNN[38]	CVPR2021	24.2	50.3	20.3	8.8	20.4	30.2	39.4
DyHead[10]	CVPR2021	27.5	56.1	23.2	12.4	24.4	33.0	41.9
RFLA[47]	ECCV2022	29.7	60.2	25.2	13.2	26.9	35.4	44.6
CFINet[55]	ICCV2023	30.7	60.8	26.7	14.7	27.8	36.4	44.6
DNTR†[24]	TGRS2024	29.6	57.8	26.5	13.1	26.7	35.5	43.4
CFPT†[11]	TGRS2025	27.5	54.5	23.8	7.1	22.4	36.2	45.9
HS-FPN†[35]	AAAI2025	29.6	56.8	26.7	13.6	26.4	35.3	45.3
Unc-SOD[54]	TIP2026	31.0	60.8	27.1	14.9	27.6	36.9	45.8
SFDNet	–	31.3	62.1	26.8	15.1	27.8	37.3	46.2
SFDNet*	–	**34.2**	**64.0**	**31.3**	**16.7**	**30.7**	**40.6**	**50.4**

Table 4. Ablation studies of the different variants in SFDNet on AI-TOD dataset. The bold number indicates the best result obtained in the experiment.

ASD	CPD	AP	AP_{vt}	AP_t	AP_s	AP_m
×	×	24.8	9.3	24.8	30.3	38.2
✓	×	28.6	13.8	28.8	33.2	39.2
×	✓	27.2	12.0	27.6	32.2	38.3
✓	✓	**29.0**	**14.4**	**29.0**	**33.4**	**40.8**

Table 5. Ablation study of the scanning mechanism in SFDNet on AI-TOD dataset. The bold number indicates the best result obtained in the experiment.

Setting	AP	AP_{vt}	AP_t	AP_s	AP_m
Sweep. [26]	26.7	13.6	27.1	30.6	36.9
Clock. [22]	27.0	13.1	27.4	31.2	36.9
Neigh. [44]	26.2	12.3	26.6	29.7	38.5
Ours	**29.0**	**14.4**	**29.0**	**33.4**	**40.8**

formance gains, achieving (27.2%, 12.0%, 27.6%, 32.2%, 38.3%) for (AP, AP_{vt}, AP_t, AP_s, AP_m), respectively. Furthermore, combining both ASD and CPD yields the best performance, with improvements of (4.2%, 5.1%, 4.2%, 3.1%, 2.6%) over the baseline. These results validate the effectiveness of the proposed ASD module and CPD procedure.

Ablation Study of the Scanning Mechanism. To validate the effectiveness of our Multi-Spectrum Scan Strategy, we conduct ablation experiments on different scanning mechanisms. Specifically, we compare several commonly used scanning methods, including sweeping, clockwise, and neighbor-based orders. As shown in Tab. 5, replacing our proposed scan strategy with any of these alternatives leads to performance degradation, demonstrating the effectiveness of the Multi-Spectrum Scan Strategy.

Fig. 5. Per Class Performance on SODA-D and AI-TOD. Please zoom in for details.

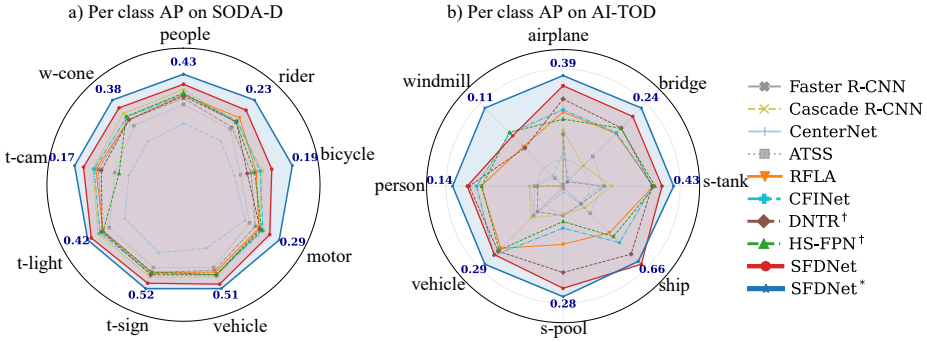


Table 6. Ablation studies of the impact of different spectra in SFDNet on AI-TOD dataset. The bold number indicates the best result obtained in the experiment.

Low	Mid	High	AP_{vt}	AP_t	AP_s	AP_m
×	✓	✓	11.4	24.4	29.4	35.0
✓	×	✓	10.0	23.4	27.3	33.2
✓	✓	×	9.2	21.2	28.5	34.8
✓	✓	✓	**14.4**	**29.0**	**33.4**	**40.8**

Table 7. Ablation studies of the spectrum aggregation mechanism in SFDNet on AI-TOD dataset. The bold number indicates the best result obtained in the experiment.

Setting	AP	AP_{vt}	AP_t	AP_s	AP_m
Attention	26.6	12.3	27.0	30.9	38.3
VMamba	26.4	12.9	26.9	30.3	36.5
Conv	26.3	11.4	26.9	30.1	38.5
S3M	**29.0**	**14.4**	**29.0**	**33.4**	**40.8**

Ablation Study of Different Spectra. To investigate the impact of different spectra, we conduct ablation experiments by selectively discarding each spectrum feature. As shown in Tab. 6, removing any single spectrum component results in performance degradation. Specifically, the removal of the low-frequency spectrum leads to decreases of (3.0%, 4.6%, 4.0%, 4.2%) in terms of (AP_{vt}, AP_t, AP_s, AP_m), respectively. The removal of mid and high frequency components leads to a more pronounced performance degradation, demonstrating their stronger relevance to small objects. These results demonstrate the critical importance of preserving full-spectrum information for small object detection.

Ablation Study of Spectrum Aggregation Mechanism. To evaluate the effectiveness of our proposed spectrum aggregation mechanism, e.g., S3M, we conduct ablation experiments by replacing S3M with alternative mechanisms. As shown in Tab. 7, substituting S3M with Attention, VMamba, or Convolution consistently results in performance degradation. These findings highlight the critical role of S3M in spectrum aggregation for small object detection.

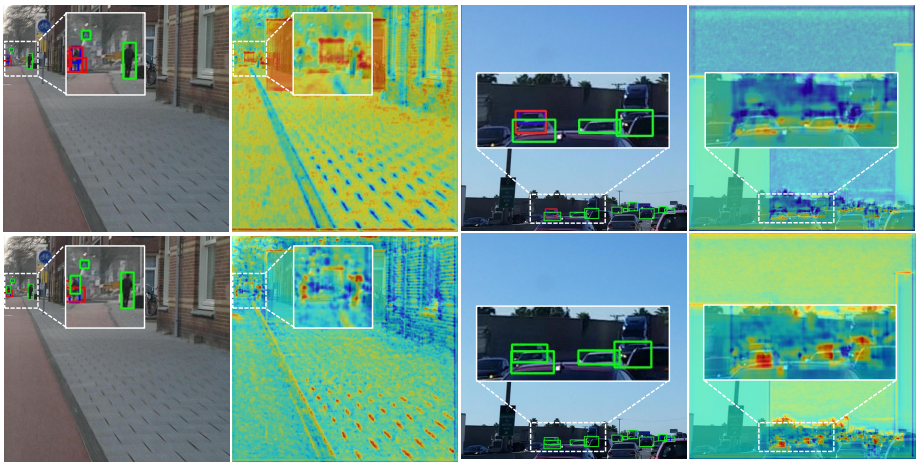
Hyperparameter Sensitivity Analysis. ASD employs DoG-based spectral decomposition to disentangle features into low, mid, and high frequency components. As shown in Tab. 8, the performance is relatively insensitive to σ and k. We attribute this robustness to the limited scale variation of small objects in the

Table 8. Ablation studies on parameters k and σ.

k	1.2	1.4	1.6	1.8
mAP	28.8	29.0	29.1	28.9

σ	0.5	1.0	1.5	2.0
mAP	28.9	29.0	28.8	28.7

Fig. 6. Visualization of feature maps at the P2 level on SODA-D dataset. The first row presents the results of DNTR, while the second row shows those of SFDNet.



evaluated datasets, enabling ASD to generalize well across different input sizes without extensive hyperparameter tuning.

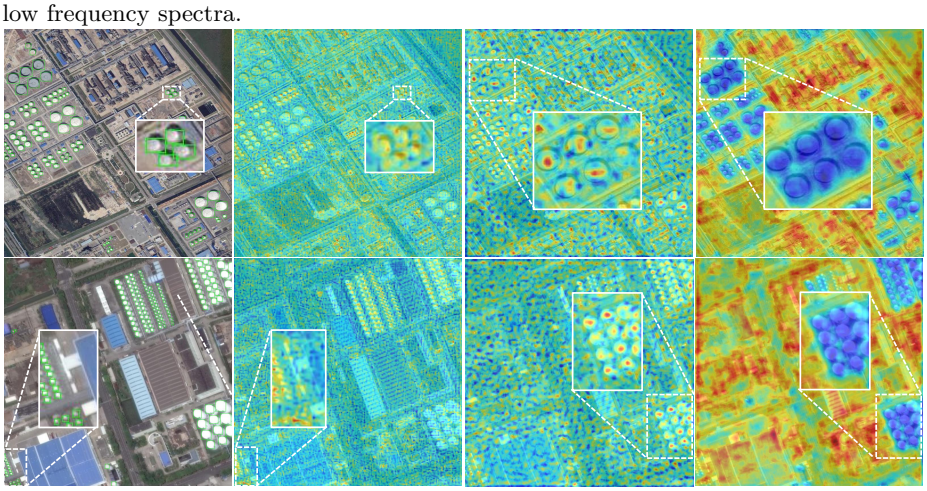
4.4 Qualitative Visualization

Heatmap Visualization. To qualitatively assess the effectiveness of our method, we visualize the feature heatmaps at the P2 level. As shown in Fig.6, the first row shows the results of DNTR, while the second row presents those of SFDNet. Compared with DNTR, SFDNet demonstrates stronger discriminative capability in its feature representations.

Qualitative Analysis of Per-Class Performance. We visualize per-class performance on the SODA-D and AI-TOD datasets. As shown in Fig. 5, our method achieves the best performance across all classes, validating the effectiveness of the proposed Adaptive Spectrum Disentanglement (ASD) module and Class-Wise Prototype Distillation (CPD) procedure.

Heatmap of Different Spectral Components. Fig. 7 visualizes the heatmaps of different spectral components. It can be observed that the low-frequency spectrum exhibits weak responses in the target regions, the mid-frequency spectrum responds strongly to relatively large objects (e.g., the storage tank in the zoomed-in region), and the high-frequency spectrum preferentially attends to smaller objects. These observations demonstrate the effectiveness of our spectrum disentanglement mechanism.

Fig. 7. Visualization of SFDNet feature maps for different spectral components at the P2 level on the AI-TOD test set. From left to right: ground-truth (GT), high, mid, and low frequency spectra.



5 Conclusion

In this paper, we propose a novel small object detection framework, termed **SFDNet**, which effectively enhances the detection of small objects via spectrum-aware feature disentanglement. Specifically, SFDNet decomposes features into multiple spectral components and employs spectrum-specific scanning strategies to capture the key semantic features of small targets within each spectral component, thereby achieving full-spectrum noise suppression and yielding discriminative feature representations. In addition, we design a Class-Wise Prototype Distillation (CPD) procedure that enforces consistent feature representations within each class. Extensive experiments on public benchmarks demonstrate that SFDNet achieves state-of-the-art performance, validating the effectiveness of the proposed method.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (No. 62402055), Beijing Natural Science Foundation (Grant No.L2603037).

References

1. Bi, Q., Yi, J., Huang, H., Zheng, H., Zhan, H., Huang, Y., Li, Y., Wu, X., Zheng, Y.: Nightadapter: Learning a frequency adapter for generalizable night-time scene segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 23838–23849 (2025) 4
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:

IEEE Conference on Computer Vision and Pattern Recognition. pp. 30376–30386 (2025) 10
3. Bosquet, B., Cores, D., Seidenari, L., Brea, V.M., Mucientes, M., Del Bimbo, A.: A full data augmentation pipeline for small object detection based on generative adversarial networks. Pattern Recognition **133**, 108998 (2023) 3
4. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6154–6162 (2018) 10
5. Chen, C., Zhang, Y., Lv, Q., Wei, S., Wang, X., Sun, X., Dong, J.: Rrnet: A hybrid detector for object detection in drone-captured images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 100–108 (2019) 3
6. Chen, L., Gu, L., Li, L., Yan, C., Fu, Y.: Frequency dynamic convolution for dense image prediction. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 30178–30188 (2025) 4
7. Chen, L., Gu, L., Zheng, D., Fu, Y.: Frequency-adaptive dilated convolution for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3414–3425 (2024) 4
8. Cheng, G., Yao, Y., Li, S., Li, K., Xie, X., Wang, J., Yao, X., Han, J.: Dual-aligned oriented detector. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–11 (2022) 11
9. Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., Xie, X., Han, J.: Towards large-scale small object detection: Survey and benchmarks. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(11), 13467–13488 (2023) 9
10. Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., Zhang, L.: Dynamic head: Unifying object detection heads with attentions. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 7373–7382 (2021) 12
11. Du, Z., Hu, Z., Zhao, G., Jin, Y., Ma, H.: Cross-layer feature pyramid transformer for small object detection in aerial images. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–14 (2025) 12
12. Fang, H., Han, B., Zhang, S., Zhou, S., Hu, C., Ye, W.M.: Data augmentation for object detection via controllable diffusion models. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1257–1266 (2024) 3
13. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023) 4
14. Han, J., Ding, J., Li, J., Xia, G.S.: Align deep features for oriented object detection. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–11 (2021) 11
15. Hatamizadeh, A., Kautz, J.: Mambavision: A hybrid mamba-transformer vision backbone. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 25261–25270 (2025) 4
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016) 10
17. Kim, J.U., Park, S., Ro, Y.M.: Robust small-scale pedestrian detection with cued recall via memory learning. In: IEEE International Conference on Computer Vision. pp. 3050–3059 (2021) 2
18. Kisantal, M., Wojna, Z., Murawski, J., Naruniec, J., Cho, K.: Augmentation for small object detection. arXiv preprint arXiv:1902.07296 (2019) 3
19. Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: European Conference on Computer Vision. pp. 734–750 (2018) 12

20. Li, C., Dai, P., Li, J., Yao, S., Jiang, Y., Zheng, Z.: Dyfclt: Dynamic frequency-decoupled cross-modal learning transformer for multimodal tiny object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 11313–11323 (2026) 2
21. Li, W., Chen, Y., Hu, K., Zhu, J.: Oriented reppoints for aerial object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1829–1838 (2022) 2, 11
22. Li, X., Fu, K., Zhao, Q.: Mamba-based efficient spatio-frequency motion perception for video camouflaged object detection. arXiv preprint arXiv:2507.23601 (2025) 12
23. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: IEEE International Conference on Computer Vision. pp. 2980–2988 (2017) 11, 12
24. Liu, H.I., Tseng, Y.W., Chang, K.C., Wang, P.J., Shuai, H.H., Cheng, W.H.: A denoising fpn with transformer r-cnn for tiny object detection. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–15 (2024) 2, 3, 10, 12
25. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: DAB-DETR: dynamic anchor boxes are better queries for DETR. In: International Conference on Learning Representations (2022) 10
26. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. In: Advances in Neural Information Processing Systems. pp. 103031–103063 (2024) 4, 12
27. Lu, W., Chen, S.B., Li, H.D., Shu, Q.L., Ding, C.H., Tang, J., Luo, B.: Legnet: A lightweight edge-gaussian network for low-quality remote sensing image object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 2865–2874 (2025) 11
28. Lu, W., Chen, S.B., Shu, Q.L., Tang, J., Luo, B.: Decouplenet: A lightweight backbone network with efficient feature decoupling for remote sensing visual tasks. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–13 (2024) 11
29. Ma, X., Ni, Z., Chen, X.: Tinyvim: Frequency decoupling for tiny hybrid vision mamba. In: IEEE International Conference on Computer Vision. pp. 23519–23529 (2025) 4
30. Marques, J.H.L., Murrugarra-Llerena, J., Jung, C.R.: Gaucho: Gaussian distributions with cholesky decomposition for oriented object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3593–3602 (2025) 11
31. Nie, G., Huang, H.: Multi-oriented object detection in aerial images with double horizontal rectangles. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4932–4944 (2022) 11
32. Qiao, S., Chen, L.C., Yuille, A.: Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 10213–10224 (2021) 10
33. 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 (2016) 10, 11, 12
34. Shi, S., Fang, Q., Xu, X., Zhao, T.: Similarity distance-based label assignment for tiny object detection. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 13711–13718 (2024) 10
35. 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: AAAI Conference on Artificial Intelligence. pp. 6896–6904 (2025) 2, 3, 10, 11, 12

36. Sun, H., Li, Y., Yang, L., Cao, X., Zhang, B.: Uncertainty-aware gradient stabilization for small object detection. In: IEEE International Conference on Computer Vision. pp. 8407–8417 (2025) 11
37. Sun, H., Wang, R., Li, Y., Yang, L., Lin, S., Cao, X., Zhang, B.: Set: Spectral enhancement for tiny object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4713–4723 (2025) 2, 10
38. Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al.: Sparse r-cnn: End-to-end object detection with learnable proposals. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 14454–14463 (2021) 12
39. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: IEEE International Conference on Computer Vision. pp. 9627–9636 (2019) 12
40. Vu, T., Jang, H., Pham, T.X., Yoo, C.D.: Cascade RPN: delving into high-quality region proposal network with adaptive convolution. In: Advances in Neural Information Processing Systems. pp. 1430–1440 (2019) 12
41. Wang, J., Yang, W., Guo, H., Zhang, R., Xia, G.S.: Tiny object detection in aerial images. In: IEEE International Conference on Pattern Recognition. pp. 3791–3798 (2021) 9
42. Wang, K., Fu, X., Huang, Y., Cao, C., Shi, G., Zha, Z.J.: Generalized uav object detection via frequency domain disentanglement. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1064–1073 (2023) 4
43. Wu, J., Zhou, C., Zhang, Q., Yang, M., Yuan, J.: Self-mimic learning for small-scale pedestrian detection. In: Proceedings of the ACM International Conference on Multimedia. pp. 2012–2020 (2020) 2
44. Xiao, C., Li, M., Zhang, Z., Meng, D., Zhang, L.: Spatial-mamba: Effective visual state space models via structure-aware state fusion. In: International Conference on Learning Representations (2025) 4, 10, 12
45. Xie, X., Cheng, G., Wang, J., Yao, X., Han, J.: Oriented r-cnn for object detection. In: IEEE International Conference on Computer Vision. pp. 3520–3529 (2021) 11
46. 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 (2022) 3
47. Xu, C., Wang, J., Yang, W., Yu, H., Yu, L., Xia, G.S.: Rfla: Gaussian receptive field based label assignment for tiny object detection. In: European Conference on Computer Vision. pp. 526–543 (2022) 3, 10, 12
48. Xu, C., Zhang, R., Yang, W., Zhu, H., Xu, F., Ding, J., Xia, G.S.: Oriented tiny object detection: A dataset, benchmark, and dynamic unbiased learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **48**(3), 3167–3184 (2026) 11
49. Xu, Y., Fu, M., Wang, Q., Wang, Y., Chen, K., Xia, G.S., Bai, X.: Gliding vertex on the horizontal bounding box for multi-oriented object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(4), 1452–1459 (2020) 2, 11
50. Yang, C., Huang, Z., Wang, N.: Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 13668–13677 (2022) 2, 10
51. Yang, Z., Liu, S., Hu, H., Wang, L., Lin, S.: Reppoints: Point set representation for object detection. In: IEEE International Conference on Computer Vision. pp. 9657–9666 (2019) 12
52. Yao, S., Guo, Y., Yan, Y., Ren, W., Cao, X.: Unctrack: Reliable visual object tracking with uncertainty-aware prototype memory network. IEEE Transactions on Image Processing **34**, 3533–3546 (2025) 2

53. Yao, S., Sun, H., Xiang, T.Z., Wang, X., Cao, X.: Hierarchical graph interaction transformer with dynamic token clustering for camouflaged object detection. IEEE Transactions on Image Processing **33**, 5936–5948 (2024) 9
54. Yuan, X., Cheng, G., Cheng, J., Yao, R., Han, J.: Unc-sod: An uncertainty learning framework for small object detection. IEEE Transactions on Image Processing **35**, 1127–1142 (2026) 11, 12
55. Yuan, X., Cheng, G., Yan, K., Zeng, Q., Han, J.: Small object detection via coarse-to-fine proposal generation and imitation learning. In: IEEE International Conference on Computer Vision. pp. 6317–6327 (2023) 2, 3, 10, 11, 12
56. 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. In: International Conference on Learning Representations (2023) 10
57. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 9759–9768 (2020) 12
58. Zhou, Z., Zhu, Y.: Kldet: Detecting tiny objects in remote sensing images via kullback–leibler divergence. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–16 (2024) 2
59. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision Mamba: Efficient visual representation learning with bidirectional state space model. In: International Conference on Machine Learning. pp. 62429–62442 (2024) 4
60. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2021) 12