StyleProto: Style-Augmented Prototype Learning for Cross-Domain Few-Shot Object Detection

Xi Yang*, Quantao Xie*

State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China

yangx@xidian.edu.cn, xiequantao@stu.xidian.edu.cn

Abstract

Cross-Domain Few-Shot Object Detection (CD-FSOD) faces significant challenges due to the dual issues of domain shift and limited labeled samples. One major challenge is style bias, caused by limited support samples that fail to represent the target domain's style diversity. Another is feature confusion, which stems from distribution shifts and limited supervision, manifesting as both object-background ambiguity and object-object confusion. To address these challenges, we propose Style-Augmented Prototype Learning (StyleProto), which constructs style-aware prototypes from support samples with diverse visual styles, and refines them via spatial weighting and discriminative fusion. Specifically, our StyleProto consists of three components: (1) Style Generation Augmentation (SGA); (2) Semantic-Focused Prototype Construction (SPC); (3) Hierarchical Prototype Fusion Aggregator (HPFA). SGA synthesizes style-diverse yet semantically consistent training samples by recombining style statistics from the support set, thus improving robustness to unseen styles. SPC aggregates support features using spatial attention to highlight object semantics and suppress background noise, yielding cleaner and more distinctive class prototypes. HPFA leverages query-guided attention to integrate discriminative support features, enhancing prototype representations with richer class-specific details. Extensive experiments on multiple benchmarks demonstrate that StyleProto consistently outperforms existing state-of-the-art methods.

Introduction

Few-shot object detection (FSOD) (Köhler, Eisenbach, and Gross 2023) has emerged as a crucial research topic in computer vision, aiming to detect novel object categories using only a handful of annotated instances. Unlike conventional detectors that rely on large-scale labeled datasets, FSOD seeks to transfer knowledge from base classes to novel ones by exploiting shared semantic structures. Existing FSOD paradigms predominantly fall into two categories. One line of work focuses on meta-learning-based approaches (Han et al. 2021, 2022b; Yan et al. 2019; Bulat et al. 2023; Zhang et al. 2022), which aim to construct task-agnostic feature spaces through episodic training. Another line follows fine-tuning-based methods (Kaul, Xie, and Zisserman 2022; Ma et al. 2023; Qiao et al. 2021; Sun et al. 2021; Wang et al. 2020), which adapt pre-trained detectors to novel classes via hierarchical fine-tuning strategies. While both paradigms have achieved remarkable success under in-domain settings, their performance often deteriorates significantly in cross-domain scenarios due to the compounded effects of domain shift and data scarcity. Cross-Domain Few-Shot Object Detection (CD-FSOD) extends FSOD to more realistic applications, where the source and target domains differ substantially in visual distribution, such as adapting detectors from natural images to industrial defects, remote sensing imagery, or underwater scenes. In addition, there exist entirely new categories in the target domain, and these categories have only a small number of annotated samples.

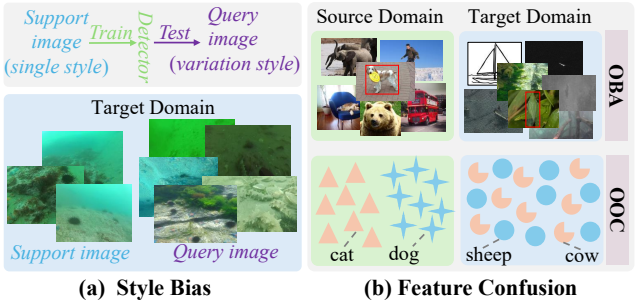

Figure 1: Challenges in CD-FSOD. (a) Style bias: support images exhibit limited style diversity, while query images show broader variations. (b) Feature confusion: the target domain exhibits increased object-background ambiguity (OBA) and object-object confusion (OOC) compared to the source domain.

CD-FSOD introduces new challenges stemming from the interplay between limited annotated data and domain distribution shift, as illustrated in Figure 1. A primary issue lies in the substantial intra-class style variations within the target domain, such as lighting, background context, or camera viewpoint, that cannot be adequately modeled by the limited annotated examples. As shown in Figure 1(a), support samples typically cover only a limited range of styles, while query instances exhibit diverse visual appearances. This style bias hinders the generalization capability of the prototypes derived from the support set. In addition, the

*Corresponding Author

combination of domain shift and limited supervision in the target domain make the model susceptible to feature confusion, as shown in Figure 1(b). This confusion manifests primarily in two forms: (1) object-background ambiguity, and (2) object-object confusion. The former is especially pronounced in visually complex environments, such as underwater or industrial settings, where object and background textures often overlap, making foreground-background separation highly challenging. The latter stems from high visual similarity across different object categories, which reduces inter-class discriminability and exacerbates confusion among target objects.

To address these challenges in CD-FSOD, we propose Style-augmented Prototype Learning (StyleProto), which, unlike conventional methods that extract prototypes from a single-style support set, constructs style-aware prototypes by leveraging support features with diverse visual styles to better capture intra-class variation across domains. These prototypes are further enhanced through spatially-aware weighting and discriminative fusion, resulting in more distinctive representations for CD-FSOD. It comprises three key components: Style Generation Augmentation (SGA), Semantic-Focused Prototype Construction (SPC), and Hierarchical Prototype Fusion Aggregator (HPFA). Specifically, SGA addresses the issue of insufficient style coverage in the support set. By extracting channel-wise style statistics (e.g., mean and variance (Zhao et al. 2022)) from support features and synthesizing new style combinations via statistical interpolation, SGA generates diverse yet semantically aligned feature variants. The generated style-augmented samples implicitly regularize training by exposing the model to broader intra-class variations, without requiring external data. This diversity is particularly crucial for support-based prototype learning, as it lays the foundation for producing more generalizable and domain-invariant representations.

Following SGA, the enriched support features, which now include synthesized style variants, are passed into the SPC module to construct initial class prototypes. SPC applies a spatial weighting strategy using soft Gaussian attention to highlight salient object regions while suppressing irrelevant background clutter, effectively mitigating object–background ambiguity caused by domain noise and limited supervision. This spatially focused aggregation boosts prototype quality by improving class specificity and reducing background interference. Crucially, SPC builds directly on the output of SGA to preserve both style diversity and semantic purity in the prototype space. Building on this foundation, HPFA further refines the SPC-generated prototypes by integrating fine-grained, class-discriminative cues from the broader support feature space. Leveraging a learnable, query-driven attention mechanism, it hierarchically fuses relevant information to enhance prototype specificity. This targeted refinement effectively mitigates object–object confusion under domain shifts, improving class separation and overall discriminability. Notably, the feature queries used in HPFA are guided by the style-augmented and spatially-refined structure established by SGA and SPC, making all three modules tightly interlinked. In summary, this end-to-end framework progressively refines support-based prototypes in a style-aware, structure-preserving, and discriminative manner, ultimately achieving robust performance across diverse CD-FSOD domains.

The contributions of this paper are as follows:

- We propose the StyleProto framework for CD-FSOD, which achieves strong generalization and competitive performance across multiple challenging CD-FSOD benchmarks, including an additional exploration on cross-spectral SAR datasets, where our method outperforms most existing approaches.
- We introduce the SGA module to mitigate support-query style bias by synthesizing diverse yet structure-preserving features from limited support data.
- We design the SPC and HPFA modules to address feature confusion by applying semantic-aware spatial weighting and query-guided prototype refinement for more discriminative class representations.

Related Works

Few-Shot Object Detection

FSOD seeks to detect novel categories using only a few labeled instances per class. Existing approaches can be broadly categorized into transfer learning-based and meta-learning-based paradigms.

Transfer learning methods adapt knowledge from base categories to novel ones by fine-tuning pre-trained models under limited supervision. LSTD (Chen et al. 2018) initiates this line of work by proposing a baseline framework and optimization strategy for few-shot settings. To alleviate the domain gap between base and novel classes, TFA (Wang et al. 2020) introduces a cosine classifier, while FSCE (Sun et al. 2021) improves region-level representations by incorporating contrastive learning. Subsequent methods such as DeFRCN (Qiao et al. 2021) further enhance detection performance by regulating gradient flow and refining feature calibration. In contrast, meta-learning-based methods aim to build task-agnostic feature representations that generalize well across novel tasks. These approaches typically learn class-level prototypes and match region features to them during inference. For example, FSRW (Kang et al. 2019) and Meta R-CNN (Yan et al. 2019) aggregate features across classes and utilize similarity-based losses to improve discrimination. FsDetView (Xiao, Lepetit, and Marlet 2022) introduces multi-source feature fusion through channel-wise operations to enrich semantic understanding. RepMet (Karlinsky et al. 2019) adopts a metric-learning framework to model class prototypes explicitly. More recently, DEViT (Zhang et al. 2025) demonstrates the potential of vision transformers in FSOD by building detectors directly from a few support images without relying on extensive language priors, offering a promising route toward open-set few-shot detection.

Cross-Domain Few-Shot Object Detection

To better address real-world constraints, CD-FSOD has emerged as a new task inspired by Cross-Domain Few-Shot Learning (CD-FSL) (Luo et al. 2023; Tang et al. 2022;

Vinyals et al. 2016; Zhang et al. 2023), which poses a dual challenge that combines few-shot learning and domain adaptation. While CD-FSL has been extensively studied in classification (Fu, Fu, and Jiang 2021; Fu et al. 2023; Hu et al. 2022; Zhuo et al. 2022), and segmentation tasks (Nie et al. 2024; Herzog 2024; Tong et al. 2024; Su et al. 2024), its extension to object detection has received comparatively less attention, despite its practical significance in real-world deployments. Recent efforts have begun to systematically investigate this setting. MoFSOD (Lee et al. 2022) provides an in-depth analysis of CD-FSOD, studying the impact of model architectures, fine-tuning strategies, and pretraining datasets. Distill-CD-FSOD (Xiong 2023) introduces a multi-dataset benchmark and proposes a distillation-based baseline. To tackle data scarcity, AcroFOD (Gao et al. 2022) uses domain-aware augmentation and adaptive optimization, while AsyFOD (Gao et al. 2023) applies asymmetric adaptation via source instance partitioning and task-specific supervision. However, both approaches assume a shared label space across domains, limiting their applicability in open-set scenarios. More recently, CD-ViTO (Fu et al. 2024) sets a strong CD-FSOD benchmark across six domains, enhancing transferability through DE-ViT-based instance representations, reweighting, and domain prompts. Building upon these works, we propose StyleProto, which enhances support diversity and constructs more discriminative prototypes to address the core challenges of CD-FSOD.

Method

Overall Framework

We follow the task setting of CD-ViTO (Fu et al. 2024). Specifically, let the source domain be $D_S = \{(I_i, y_i)\}_{i=1}^{N_S}$, where labels $y_i \in C_S$ are drawn from distribution P_S; and the target domain be $D_T = \{(I_j, y_j)\}_{j=1}^{N_T}$, with $y_j \in C_T$ sampled from P_T. The label spaces are disjoint, i.e., $C_S \cap C_T = \emptyset$. Unlike standard FSOD that assumes domain consistency $P_S = P_T$, CD-FSOD addresses the more realistic and challenging case where $P_S \neq P_T$, requiring models to generalize from well-annotated D_S to sparsely labeled D_T, typically with $|D_T| \ll |D_S|$.

As illustrated in Figure 2, the base architecture includes a frozen DINOv2 ViT (Oquab et al. 2023) backbone, a Region Proposal Network (RPN), ROI Align, a Detection Head, and a One-vs-Rest Classification Head. Built upon this foundation, we propose the StyleProto framework, which enhances prototype quality by incorporating style diversity, spatial weighting, and discriminative fusion. The framework is composed of three key modules: SGA, SPC, and HPFA.

Given a support image, we apply style augmentation to generate diverse versions via channel-wise normalization. Both stylized and original images are fed into the frozen backbone to obtain instance-level support features. SPC constructs initial prototypes P_{init} by emphasizing semantic object regions and suppressing background noise via Gaussian-masked weighting. These prototypes initialize the learnable queries in HPFA, which attend to discriminative regions in $F_{\text{ins}}^{\text{obj}}$ using scaled dot-product attention and generate refined prototypes P_{attend}. The final class-wise prototypes P_{final} are obtained by fusing P_{init} and P_{attend} via a learnable linear combination. For the query image, instance features $F_{\text{ins}}^{\text{obj}}$ and region proposals are extracted through DINOv2 and RPN. The Detection Head utilizes $F_{\text{ins}}^{\text{obj}}$ and P_{final} to localize objects, while the ROI-aligned features are classified using the One-vs-Rest Classification Head. During adaptation, only the Detection Head and Classification Head are fine-tuned following CD-ViT's protocol, while the backbone remain frozen to preserve cross-domain generalization. The overall training objective combines a bounding box regression loss L_{det} and a one-vs-rest classification loss L_{cls}, applied on query predictions with respect to prototype-guided labels.

Stylized Generation Augmentation (SGA)

In the StyleProto framework, support prototypes serve as the foundation for guiding query prediction. However, under the CD-FSOD setting, the limited support samples often fail to capture the rich style diversity of the target domain, including variations in lighting, viewpoint, and background context. This lack of intra-class variation weakens the representational capacity of prototypes and hinders generalization. To alleviate this, we introduce the SGA module, which expands the style diversity of support features while preserving their spatial and semantic structure. Rather than relying on external data or image-level transformation, SGA operates directly in the feature space, enabling structure-aware style synthesis in a compact and efficient manner. Inspired by channel-wise statistical style transfer, SGA extracts per-channel style statistics, specifically the mean μ_c^{s} and standard deviation σ_c^{s}, from support features $F_{\text{ins}} \in \mathbb{R}^{C\times H\times W}$:

$$\mu_c^{\text{s}} = \frac{\sum_{h,w} F_{\text{ins}}}{HW}, \quad \sigma_c^{\text{s}} = \sqrt{\frac{\sum_{h,w}(F_{\text{ins}} - \mu_c^{\text{s}})^2}{HW}}. \tag{1}$$

These statistics serve as a compact encoding of global appearance factors like brightness and contrast. To synthesize novel styles, SGA samples a convex combination of style statistics drawn from the support set using a Dirichlet distribution:

$$\mu_{\text{new}}^{\text{s}} = \sum_{i=1}^{C} w_i \mu^{(i),\text{s}}, \quad \sigma_{\text{new}}^{\text{s}} = \sum_{i=1}^{C} w_i \sigma^{(i),\text{s}}, \tag{2}$$

where $\boldsymbol{w} \sim \text{Dirichlet}(\alpha)$ and $\{\mu^{(i),\text{s}}, \sigma^{(i),\text{s}}\}$ are basis statistics from the support samples. The new style is applied via Adaptive Instance Normalization (AdaIN):

$$\text{AdaIN}(F_{\text{ins}}) = \sigma_{\text{new}}^{\text{s}} \left(\frac{F_{\text{ins}} - \mu(F_{\text{ins}})}{\sigma(F_{\text{ins}})} \right) + \mu_{\text{new}}^{\text{s}}. \tag{3}$$

Crucially, SGA preserves spatial alignment, making it well-suited for detection tasks that rely on precise localization. As a result, the augmented support features retain semantic consistency while covering a broader range of visual styles, enabling the model to learn more robust and domain-invariant prototypes. Integrated within the StyleProto pipeline, SGA enhances the expressive power of support representations without introducing additional supervision. This leads to significant performance gains in low-shot scenarios, particularly under severe domain shifts.

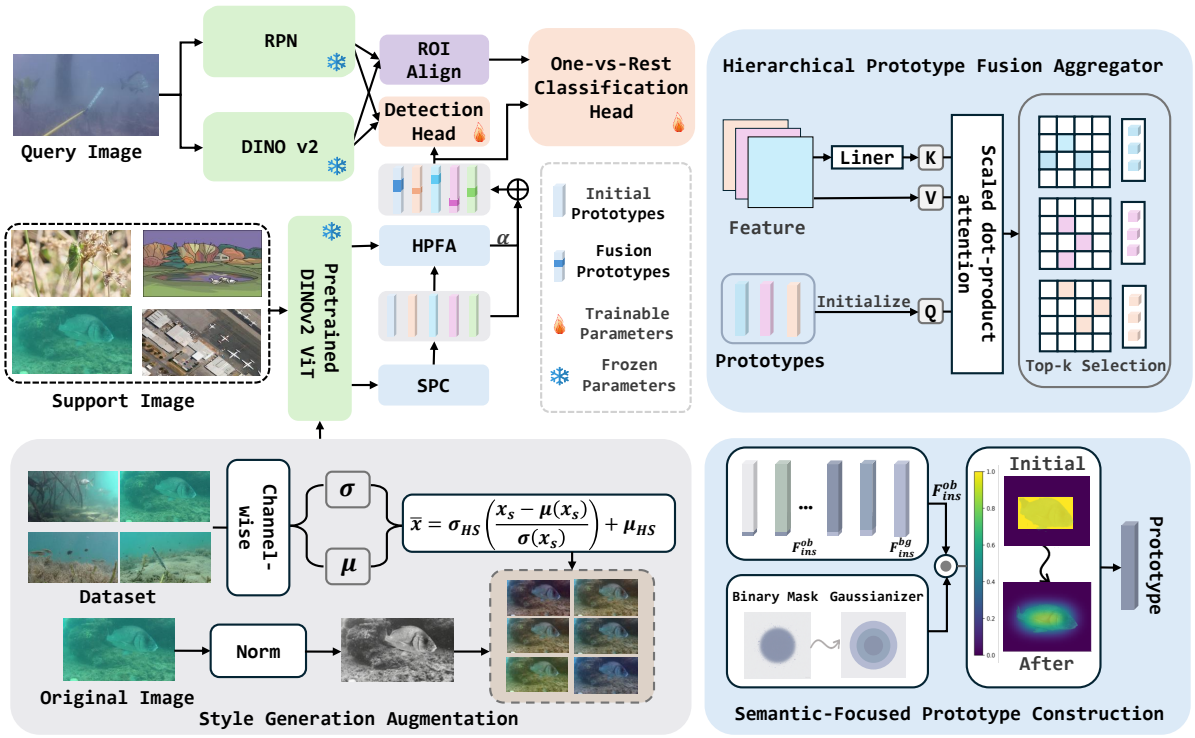

Figure 2: The overall architecture of our StyleProto framework. It includes three key components: Style Generation Augmentation (SGA), Semantic-Focused Prototype Construction (SPC), and Hierarchical Prototype Fusion Aggregator (HPFA).

Semantic-Focused Prototype Construction (SPC)

Within the StyleProto framework, SPC module transforms the style-augmented features from SGA into robust, semantically consistent prototypes. While SGA addresses intra-class style diversity by generating spatially aligned, diverse support features, SPC ensures that these features contribute to prototypes centered on domain-invariant, discriminative regions.

CD-FSOD suffers from semantic ambiguity and object-background entanglement, especially under significant domain shifts. Directly averaging support features with noisy or blurred boundaries yields prototypes with weakened semantics and sensitivity to domain-specific artifacts. SPC alleviates this via a spatially aware feature aggregation mechanism, emphasizing semantic cores and suppressing background clutter and boundary noise. Concretely, for each support instance, we first obtain a coarse binary mask M to approximate the object's spatial extent. Rather than treating all foreground pixels equally, we transform this binary mask into a soft Gaussian mask $\tilde{M}$ that gradually reduces attention toward object edges:

$$\tilde{M} = \frac{G_{\sigma^g}(M)}{\max(G_{\sigma^g}(M))}, \tag{4}$$

where G_{σ^g} is a Gaussian filter with standard deviation σ^g. This soft mask reflects the intuition that object centers are typically more stable and domain-invariant than peripheral regions, which are more susceptible to appearance changes.

Given a support feature map F_{ins} extracted via a frozen DINOv2 backbone, SPC computes the spatially-weighted features as:

$$F_{\text{weighted}} = F_{\text{ins}} \odot \tilde{M}, \tag{5}$$

where $\odot$ denotes element-wise multiplication. These weighted features are then aggregated across all N support samples of a class to form the initial prototype:

$$P = \frac{1}{N} \sum_{i=1}^{N} F^{(i)}_{\text{weighted}}. \tag{6}$$

By aligning prototype construction with semantic structure, SPC produces cleaner and more discriminative class representations, reducing the impact of domain-specific noise.

Hierarchical Prototype Fusion Aggregator (HPFA)

While the SPC module constructs initial prototypes with strong semantic focus, their expressiveness remains limited in scenarios with complex domain shifts. Specifically, due to the scarcity of support samples and the inherent domain gap, fixed prototypes may fail to capture the fine-grained variations of objects in the target domain. This leads to semantic confusion between similar classes or between foreground and background regions. To address these limitations, we introduce the HPFA module, which complements SPC by dynamically refining prototypes through cross-domain feature interaction.

HPFA builds upon the SPC-initialized prototypes by introducing a set of learnable query embeddings that actively

Method	Backbone	1-shot									
		ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	LEVIR	HRSID	SSDD	Avg.
Meta-RCNN(Yan et al. 2019)	ResNet50	2.8	-	7.8	-	-	3.6	5.6	0.6	0.1	/
TFAw/cos(Wang et al. 2020)	ResNet50	3.1	-	8.0	-	-	4.4	6.1	0.1	0.2	/
FSCE(Sun et al. 2021)	ResNet50	3.7	-	8.6	-	-	3.9	5.8	0.1	0.1	/
DeFRCN(Qiao et al. 2021)	ResNet50	3.6	-	9.3	-	-	4.5	8.2	0.4	0.2	/
Distill-cdfsod(Xiong 2023)	ResNet50	5.1	7.6	10.5	nan	nan	**5.9**	8.9	-	-	/
ViTDeT-FT(Li et al. 2022)	ViT-B/14	5.9	6.1	12.9	0.9	2.4	4.0	10.2	1.1	0.9	4.9
Detic-FT(Zhou et al. 2022)	ViT-L/14	3.2	15.1	4.1	9.0	3.8	4.2	3.5	2.2	1.4	5.2
DE-ViT(Zhang et al. 2025)	ViT-L/14	0.4	0.5	2.7	0.4	0.4	1.5	0.1	0.1	0.2	0.7
DE-ViT-FT(Zhang et al. 2025)	ViT-L/14	10.5	13.0	14.7	19.3	0.6	2.4	14.6	0.7	0.4	8.5
CD-ViTO(Fu et al. 2024)	ViT-L/14	21.0	17.7	17.8	20.3	3.6	3.1	15.9	1.4	5.8	11.8
Our	ViT-L/14	**25.2**	**26.3**	**22.6**	**21.5**	**5.9**	5.7	**18.2**	**3.9**	**8.4**	**15.3**

Method	Backbone	5-shot									
		ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	LEVIR	HRSID	SSDD	Avg.
Meta-RCNN(Yan et al. 2019)	ResNet50	8.5	-	17.7	-	-	8.8	15.8	1.6	0.9	/
TFAw/cos(Wang et al. 2020)	ResNet50	8.8	-	18.1	-	-	8.7	15.4	2.3	1.9	/
FSCE(Sun et al. 2021)	ResNet50	10.2	-	18.7	-	-	9.6	15.8	1.7	2.1	/
DeFRCN (Qiao et al. 2021)	ResNet50	9.9	-	18.9	-	-	9.9	16.3	3.2	1.8	/
Distill-cdfsod(Xiong 2023)	ResNet50	12.5	23.3	19.1	15.5	**16.0**	**12.2**	17.8	13.4	11.9	15.7
ViTDeT-FT(Li et al. 2022)	ViT-B/14	20.9	23.3	23.3	9.0	13.5	11.1	19.6	11.4	11.9	16.0
Detic-FT(Zhou et al. 2022)	ViT-L/14	8.7	20.2	12.1	14.3	14.1	10.4	9.8	12.5	14.4	12.9
DE-ViT(Zhang et al. 2025)	ViT-L/14	10.1	5.5	7.8	2.5	1.5	3.1	3.1	1.2	1.3	4.0
DE-ViT-FT(Zhang et al. 2025)	ViT-L/14	38.0	38.1	23.4	21.2	7.8	5.0	25.4	7.7	5.8	19.2
CD-ViTO(Fu et al. 2024)	ViT-L/14	47.9	41.1	26.9	22.3	11.4	6.8	25.3	12.5	13.3	23.1
Our	ViT-L/14	**53.3**	**43.8**	**27.7**	**23.5**	13.2	8.4	**29.1**	**16.2**	**14.8**	**25.5**

Method	Backbone	10-shot									
		ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	LEVIR	HRSID	SSDD	Avg.
Meta-RCNN(Yan et al. 2019)	ResNet50	14.0	-	20.6	-	-	11.2	18.6	3.4	4.6	/
TFAw/cos(Wang et al. 2020)	ResNet50	14.8	-	20.5	-	-	11.8	18.1	2.9	5.5	/
FSCE(Sun et al. 2021)	ResNet50	15.9	-	21.9	-	-	12.0	17.6	3.5	4.2	/
DeFRCN(Qiao et al. 2021)	ResNet50	15.5	-	22.9	-	-	12.1	19.8	5.2	6.9	/
Distill-cdfsod(Xiong 2023)	ResNet50	18.1	27.3	26.5	15.5	**21.1**	14.5	24.7	16.8	22.1	20.7
ViTDeT-FT(Li et al. 2022)	ViT-B/14	23.4	25.6	29.4	6.5	15.8	**15.6**	28.1	13.6	17.3	19.5
Detic-FT(Zhou et al. 2022)	ViT-L/14	12.0	22.3	15.4	17.9	16.8	14.4	29.8	14.7	18.8	18.0
DE-ViT(Zhang et al. 2025)	ViT-L/14	9.2	11.0	8.4	2.1	1.8	3.1	3.8	3.6	3.2	5.1
DE-ViT-FT(Zhang et al. 2025)	ViT-L/14	49.2	40.8	25.6	21.3	8.8	5.4	29.1	9.6	13.8	22.6
CD-ViTO(Fu et al. 2024)	ViT-L/14	60.5	44.3	30.8	22.3	12.8	7.0	28.7	15.3	18.7	26.7
Our	ViT-L/14	**61.8**	**45.6**	**32.1**	**22.8**	13.5	8.8	**32.8**	**17.2**	**22.9**	**28.6**

Table 1: The 1/5/10-shot main results (mAP) on nine publicly datasets (ArTaxOr, Clipart1K, DeepFish, DIOR, NEU-DET, UODD, LEVIR, HRSID and SSDD).

search and aggregate discriminative regions from support feature maps. This fine-grained adaptation enhances the robustness of prototypes and enables the model to more effectively localize subtle yet crucial differences in the target domain. The refined prototypes are then fused with the initial ones, leading to a more semantically consistent and discriminatively powerful representation.

The HPFA module comprises two main components: learnable query embeddings and attention-based feature aggregation. The query embeddings are initialized from the SPC-generated prototypes, thereby inheriting their semantic-focused characteristics while remaining adaptable. These learnable embeddings act as soft search templates that probe the support feature maps for relevant fine-grained details. To enable spatially-aware feature retrieval, we apply an attention-based aggregation scheme. For each support feature map $F_{\text{ins}} \in \mathbb{R}^{B\times C\times H\times W}$ (where B is batch size, C is feature dimension, and H, W are spatial dimensions), we first perform spatial downsampling using max pooling. The downsampled features are reshaped to $\mathbb{R}^{B\times(H'W')\times C}$ and projected into the query embedding space.

Next, scaled dot-product attention is computed between the query embeddings q and the projected features K:

$$\text{Attn} = \text{softmax}\left(\frac{\mathbf{q}K^T}{\sqrt{k_d}}\right). \tag{7}$$

This attention highlights the most relevant regions in the support feature map, allowing fine-grained localization of key object parts. The resulting attended prototype is $P_{\text{attend}} \in \mathbb{R}^{B\times N_q\times C}$. To synthesize both semantic stability and local adaptability, HPFA fuses the initial and attended prototypes through a learnable combination:

$$P_{\text{final}} = W_{\text{combine}}\left([P_{\text{init}}; P_{\text{attend}}]\right), \tag{8}$$

where $[\cdot;\cdot]$ denotes concatenation and W_{combine} is a linear transformation. This fusion ensures that the final prototypes

Method	10-shot			30-shot		
	nAP	nAP50	nAP75	nAP	nAP50	nAP75
FSRW (Kang et al. 2019)	5.6	12.3	4.6	9.1	19	7.6
Meta R-CNN (Yan et al. 2019)	6.1	19.1	6.6	9.9	25.3	10.8
TFA (Wang et al. 2020)	10	19.2	9.2	13.5	24.9	13.2
FSCE (Sun et al. 2021)	11.9	-	10.5	16.4	-	16.2
Retentive RCNN (Fan et al. 2021)	10.5	19.5	9.3	13.8	22.9	13.8
HeteroGraph (Han et al. 2021)	11.6	23.9	9.8	16.5	31.9	15.5
Meta Faster R-CNN (Han et al. 2022a)	12.7	25.7	10.8	16.6	31.8	15.8
LVC (Kaul, Xie, and Zisserman 2022)	19	34.1	19	26.8	45.8	27.5
Cross-Transformer (Han et al. 2022b)	17.1	30.2	17	21.4	35.5	22.1
NIFF (Guirguis et al. 2023)	18.8	-	-	20.9	-	-
DiGeo (Ma et al. 2023)	10.3	18.7	9.9	14.2	26.2	14.8
FM-FSOD(Han and Lim 2024)	27.7	38.6	30.1	37.0	51.3	39.7
DE-ViT(Zhang et al. 2025)	34.0	53.0	37.0	34.0	52.9	37.2
SCSM(Xin et al. 2025)	22.4	-	23.5	27.8	-	28.6
DE-ViT w/CCL(Chen et al. 2025)	34.4	-	37.5	34.5	-	37.4
CD-ViTO(Fu et al. 2024)	35.3	54.9	37.2	35.9	54.5	38.0
Our	**35.5**	**55.1**	**37.3**	**36.2**	**55.1**	**38.2**

Table 2: Results (nAP, nAP50, and nAP75) on COCO FSOD benchmark. The nAP denotes mAP for novel classes.

maintain SPC's semantic focus while incorporating discriminative local variations crucial for domain generalization.

By dynamically refining prototypes through attention-guided aggregation, HPFA addresses the fine-grained confusion and semantic drift often encountered in CD-FSOD. In tandem with SPC, it forms a two-stage prototype construction pipeline that is both semantically grounded and spatially adaptive, yielding significant gains in cross-domain classification and localization performance.

Experiments

Datasets and Evaluation Metrics

We follow the benchmark protocol established in CD-ViTO for evaluating cross-domain few-shot object detection. Specifically, the model is pre-trained on the source domain COCO and fine-tuned on nine target datasets. The first six target domains are ArTAXOr (Drange 2020), Clipart1k (Inoue et al. 2018), DIOR (Li et al. 2020), DeepFish (Saleh et al. 2020), NEU-DET (Song and Yan 2013), and UODD (Jiang et al. 2021), all of which are standard RGB-based datasets used in prior benchmarks. To further evaluate cross-domain generalization under more severe modality and appearance shifts, we introduce three additional datasets: LEVIR (Zou and Shi 2017), HRSID (Wei et al. 2020), and SSDD (Li, Qu, and Shao 2017), the latter two being SAR (Synthetic Aperture Radar) datasets characterized by significantly different imaging mechanisms and larger domain gaps. This allows a more comprehensive assessment of model robustness in challenging non-RGB scenarios. To ensure fair comparisons with prior work, we evaluate under the standard fine-tuning setting. Performance is measured using the mean Average Precision across IoU thresholds from 0.5 to 0.95 (mAP@0.5:0.95), reported both with and without fine-tuning. All experiments are conducted under 1-shot, 5-shot, and 10-shot settings to assess model robustness across varying levels of supervision.

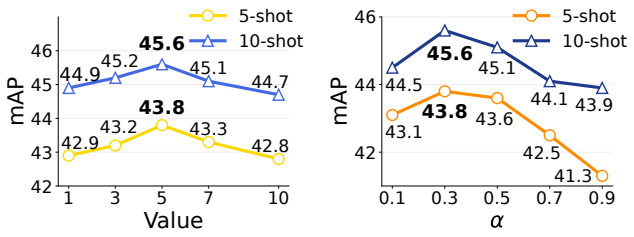

Figure 3: Hyperparameter ablation studies on the number of discriminative queries per class (left) and the fusion coefficient α in HPFA (right).

SGA	SPC	HPFA	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	LEVIR	HRSID	SSDD
			47.9	41.1	26.9	22.3	11.4	6.8	25.3	13.3	13.3
✓			48.6	42.3	27.1	22.9	11.5	7.1	26.5	13.6	13.5
	✓		48.1	41.9	27.0	22.7	11.7	7.4	26.1	13.8	13.4
		✓	49.3	42.1	27.4	22.5	12.5	6.9	27.4	15.1	13.9
✓	✓		50.3	42.9	27.2	23.3	11.8	8.1	27.8	14.4	13.7
✓		✓	52.4	43.1	27.6	23.1	12.1	7.9	28.7	15.3	14.1
	✓	✓	51.2	42.4	27.5	23.0	12.8	7.6	28.3	15.7	14.4
✓	✓	✓	**53.3**	**43.8**	**27.7**	**23.5**	**13.2**	**8.4**	**29.1**	**16.2**	**14.8**

Table 3: Full ablation study on all nine target datasets under the 5-shot setting, evaluating the impact of SGA, SPC, and HPFA.

Implementation Details

All experiments are conducted using PyTorch 2.0 and CUDA 11.8 on four NVIDIA RTX 4090 GPUs. Our implementation is based on Detectron2. We use DINOv2 ViT-L/14 for prototype extraction, and adopt specific architectural and training configurations tailored to each module.

MainResults

Table 1 summarizes the performance of our method under 1-shot, 5-shot, and 10-shot settings across nine target-domain benchmarks. As CD-FSOD is still a nascent task with limited dedicated detectors, we adapt several representative FSOD methods for comparison. However, these baselines generally exhibit limited performance on target domains. For instance, Meta-RCNN and TFA achieve only 14.0% and 14.8% mAP, respectively, on the ArTAXOr dataset under the 10-shot setting. Even Distill-CD-FSOD, a method specifically designed for CD-FSOD, struggles when handling multiple diverse domains and unseen classes.

In contrast, domain-aware approaches such as DE-ViT and CD-ViTO achieve notable improvements (49.2% and 60.5% mAP on ArTAXOr, respectively, under 10-shot). Building upon CD-ViTO, our method consistently outperforms all baselines by a significant margin across all shot settings and datasets, establishing new state-of-the-art results for CD-FSOD. In particular, HRSID and SSDD present greater domain shifts due to their distinct imaging mechanisms, posing more severe cross-domain challenges. Our framework demonstrates consistently strong performance on these datasets, underscoring its capacity to generalize

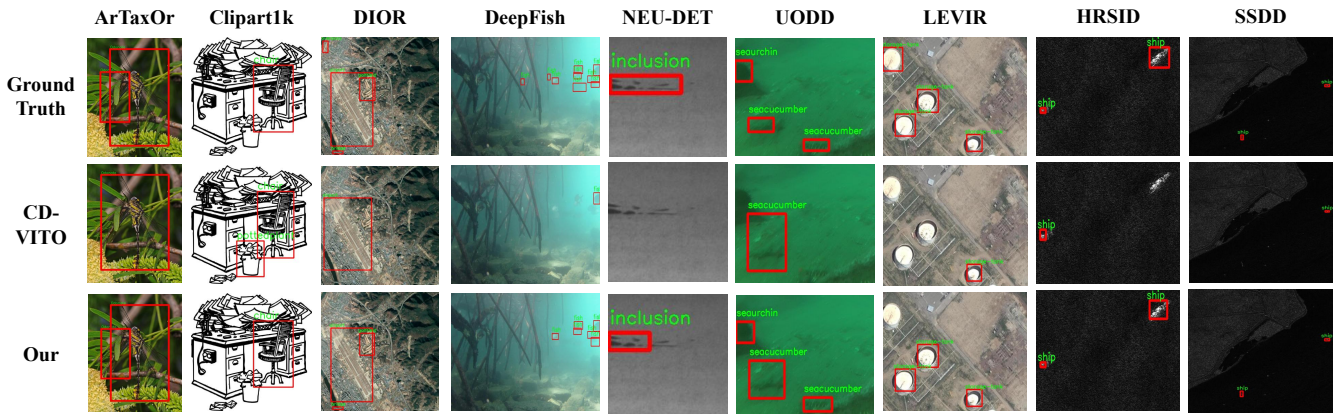

Figure 4: Visualization results of ground truth, CD-ViTO and our StyleProto.

across drastically different visual domains. To further evaluate the method's versatility, we also conduct experiments on COCO, where the target domain partially overlaps with the source (Table 2). Unlike many cross-domain models that sacrifice source-domain accuracy, our method maintains strong performance in the source domain as well, outperforming all baselines. This indicates that StyleProto not only excels in cross-domain generalization but also preserves competitive accuracy in standard FSOD scenarios.

Ablation Study

To verify the effectiveness of each component in our framework, we conduct comprehensive ablation studies on the nine target datasets under the 5-shot setting. Table 3 reports the mean average precision under various configurations. Our StyleProto approach generates diverse style variants and progressively refines prototypes, consistently boosting performance. Each module contributes clearly to enhancing prototype quality and cross-domain generalization.

Effectiveness of SGA. To evaluate the effectiveness of the SGA, we remove it and train the model using only a few original support samples. This leads to a noticeable drop in performance, highlighting the importance of exposing the model to diverse intra-class styles under limited supervision. The stylized instances promote domain-invariant representation learning and enhance generalization to unseen domains.

Effectiveness of SPC. We then assess the role of SPC in building initial prototypes. Without SPC, the aggregated features are more susceptible to background noise and irrelevant details. By introducing spatially-aware weighting, SPC generates more semantically concentrated and discriminative prototypes, leading to consistent performance gains.

Effectiveness of HPFA. We assess HPFA by comparing the full model with a variant using only SPC-generated prototypes. While SPC offers a strong baseline, it lacks adaptability to fine-grained, class-specific cues. HPFA addresses this by leveraging learnable queries to extract and fuse discriminative details from support features, leading to more robust prototypes. Performance gains confirm HPFA's role in improving inter-class separation under domain shifts.

Parameter Analysis. We conduct hyperparameter ablation studies on two key factors in HPFA, as illustrated in Figure 3. The left plot analyzes the number of discriminative query prototypes per class, where performance initially improves and peaks at five prototypes, then slightly declines as redundant cues introduce noise. The right plot examines the impact of the fusion coefficient α. The model achieves optimal performance at $\alpha = 0.3$, where the fused prototype effectively balances the stability of SPC and the adaptability of HPFA. Extremely low α values underutilize spatial cues, while high values overemphasize noisy attended regions. These results confirm that moderate configurations of both parameters yield the most robust cross-domain representations.

Visualize Detection Results

We visualize representative results on each target dataset in Figure 4, comparing our StyleProto with the vanilla CD-ViTO. StyleProto shows a lower rate of false positives and false negatives, especially in cases involving multiple adjacent objects or significant object-background ambiguity. It more reliably detects small or visually confusing targets that CD-ViTO often misses or misclassifies, reflecting improved discrimination under complex cross-domain conditions.

Conclusion

In this paper, we propose StyleProto, a novel framework for cross-domain few-shot object detection. By modeling style diversity to reduce bias and employing attention-enhanced prototype refinement to resolve feature confusion, our method bridges the gap between source-trained priors and diverse target appearances. The proposed SGA, SPC, and HPFA modules jointly enhance the robustness and adaptability of prototype representations under domain shift. Extensive experiments across multiple datasets demonstrate state-of-the-art performance under various few-shot settings. Future work will explore dynamic prototype adaptation and joint optimization with pseudo-labeled targets to further improve open-set generalization in real-world scenarios.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62372348, in part by the Key Research and Development Program of Shaanxi under Grant 2024GX-ZDCYL-02-10, in part by Shaanxi Outstanding Youth Science Fund Project under Grant 2023-JC-JQ-53.

References

Bulat, A.; Guerrero, R.; Martinez, B.; and Tzimiropoulos, G. 2023. Fs-detr: Few-shot detection transformer with prompting and without re-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11793–11802.

Chen, H.; Wang, Y.; Wang, G.; and Qiao, Y. 2018. Lstd: A low-shot transfer detector for object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Chen, R.; Zhang, H.; Li, J.; Liu, L.; Huang, Z.; and Cao, X. 2025. Generalized Semantic Contrastive Learning via Embedding Side Information for Few-Shot Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Drange, G. 2020. Arthropod Taxonomy Orders Object Detection Dataset.

Fan, Z.; Ma, Y.; Li, Z.; and Sun, J. 2021. Generalized few-shot object detection without forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4527–4536.

Fu, Y.; Fu, Y.; and Jiang, Y.-G. 2021. Meta-fdmixup: Cross-domain few-shot learning guided by labeled target data. In *Proceedings of the ACM International Conference on Multimedia*, 5326–5334.

Fu, Y.; Wang, Y.; Pan, Y.; Huai, L.; Qiu, X.; Shangguan, Z.; Liu, T.; Fu, Y.; Van Gool, L.; and Jiang, X. 2024. Cross-domain few-shot object detection via enhanced open-set object detector. In *European Conference on Computer Vision*, 247–264.

Fu, Y.; Xie, Y.; Fu, Y.; and Jiang, Y.-G. 2023. Styleadv: Meta style adversarial training for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24575–24584.

Gao, Y.; Lin, K.-Y.; Yan, J.; Wang, Y.; and Zheng, W.-S. 2023. Asyfod: An asymmetric adaptation paradigm for few-shot domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3261–3271.

Gao, Y.; Yang, L.; Huang, Y.; Xie, S.; Li, S.; and Zheng, W.-S. 2022. Acrofod: An adaptive method for cross-domain few-shot object detection. In *European Conference on Computer Vision*, 673–690.

Guirguis, K.; Meier, J.; Eskandar, G.; Kayser, M.; Yang, B.; and Beyerer, J. 2023. Niff: Alleviating forgetting in generalized few-shot object detection via neural instance feature forging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24193–24202.

Han, G.; He, Y.; Huang, S.; Ma, J.; and Chang, S.-F. 2021. Query adaptive few-shot object detection with heterogeneous graph convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3263–3272.

Han, G.; Huang, S.; Ma, J.; He, Y.; and Chang, S.-F. 2022a. Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 780–789.

Han, G.; and Lim, S.-N. 2024. Few-shot object detection with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28608–28618.

Han, G.; Ma, J.; Huang, S.; Chen, L.; and Chang, S.-F. 2022b. Few-shot object detection with fully cross-transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5321–5330.

Herzog, J. 2024. Adapt before comparison: A new perspective on cross-domain few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23605–23615.

Hu, S. X.; Li, D.; Stühmer, J.; Kim, M.; and Hospedales, T. M. 2022. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9068–9077.

Inoue, N.; Furuta, R.; Yamasaki, T.; and Aizawa, K. 2018. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5001–5009.

Jiang, L.; Wang, Y.; Jia, Q.; Xu, S.; Liu, Y.; Fan, X.; Li, H.; Liu, R.; Xue, X.; and Wang, R. 2021. Underwater species detection using channel sharpening attention. In *Proceedings of the ACM International Conference on Multimedia*, 4259–4267.

Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; and Darrell, T. 2019. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Vonference on Computer Vision*, 8420–8429.

Karlinsky, L.; Shtok, J.; Harary, S.; Schwartz, E.; Aides, A.; Feris, R.; Giryes, R.; and Bronstein, A. M. 2019. Repmet: Representative-based metric learning for classification and few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5197–5206.

Kaul, P.; Xie, W.; and Zisserman, A. 2022. Label, verify, correct: A simple few shot object detection method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14237–14247.

Köhler, M.; Eisenbach, M.; and Gross, H.-M. 2023. Few-shot object detection: A comprehensive survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9): 11958–11978.

Lee, K.; Yang, H.; Chakraborty, S.; Cai, Z.; Swaminathan, G.; Ravichandran, A.; and Dabeer, O. 2022. Rethinking few-shot object detection on a multi-domain benchmark. In *European Conference on Computer Vision*, 366–382.

Li, J.; Qu, C.; and Shao, J. 2017. Ship detection in SAR images based on an improved faster R-CNN. In *2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA)*, 1–6.

Li, K.; Wan, G.; Cheng, G.; Meng, L.; and Han, J. 2020. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159: 296–307.

Li, Y.; Mao, H.; Girshick, R.; and He, K. 2022. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, 280–296.

Luo, X.; Wu, H.; Zhang, J.; Gao, L.; Xu, J.; and Song, J. 2023. A closer look at few-shot classification again. In *International Conference on Machine Learning*, 23103–23123.

Ma, J.; Niu, Y.; Xu, J.; Huang, S.; Han, G.; and Chang, S.-F. 2023. Digeo: Discriminative geometry-aware learning for generalized few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3208–3218.

Nie, J.; Xing, Y.; Zhang, G.; Yan, P.; Xiao, A.; Tan, Y.-P.; Kot, A. C.; and Lu, S. 2024. Cross-domain few-shot segmentation via iterative support-query correspondence mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3380–3390.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Qiao, L.; Zhao, Y.; Li, Z.; Qiu, X.; Wu, J.; and Zhang, C. 2021. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8681–8690.

Saleh, A.; Laradji, I. H.; Konovalov, D. A.; Bradley, M.; Vazquez, D.; and Sheaves, M. 2020. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific Reports*, 10(1): 14671.

Song, K.; and Yan, Y. 2013. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Applied Surface Science*, 285: 858–864.

Su, J.; Fan, Q.; Pei, W.; Lu, G.; and Chen, F. 2024. Domain-rectifying adapter for cross-domain few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24036–24045.

Sun, B.; Li, B.; Cai, S.; Yuan, Y.; and Zhang, C. 2021. Fsce: Few-shot object detection via contrastive proposal encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7352–7362.

Tang, H.; Yuan, C.; Li, Z.; and Tang, J. 2022. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognition*, 130: 108792.

Tong, J.; Zou, Y.; Li, Y.; and Li, R. 2024. Lightweight frequency masker for cross-domain few-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 37: 96728–96749.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, 29.

Wang, X.; Huang, T. E.; Darrell, T.; Gonzalez, J. E.; and Yu, F. 2020. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*.

Wei, S.; Zeng, X.; Qu, Q.; Wang, M.; Su, H.; and Shi, J. 2020. HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation. *IEEE Access*, 8: 120234–120254.

Xiao, Y.; Lepetit, V.; and Marlet, R. 2022. Few-shot object detection and viewpoint estimation for objects in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3090–3106.

Xin, Z.; Wu, T.; Zou, Y.; Chen, S.; Fu, D.; and You, X. 2025. Few-Shot Object Detection via Spatial-Channel State Space Model. *arXiv preprint arXiv:2507.15308*.

Xiong, W. 2023. CD-FSOD: A benchmark for cross-domain few-shot object detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Yan, X.; Chen, Z.; Xu, A.; Wang, X.; Liang, X.; and Lin, L. 2019. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9577–9586.

Zhang, G.; Luo, Z.; Cui, K.; Lu, S.; and Xing, E. P. 2022. Meta-DETR: Image-level few-shot detection with inter-class correlation exploitation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 12832–12843.

Zhang, J.; Gao, L.; Luo, X.; Shen, H.; and Song, J. 2023. Deta: Denoised task adaptation for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11541–11551.

Zhang, X.; Liu, Y.; Wang, Y.; and Boularias, A. 2025. Detect Everything with Few Examples. In *Conference on Robot Learning*, 3986–4004.

Zhao, Y.; Zhong, Z.; Zhao, N.; Sebe, N.; and Lee, G. H. 2022. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *European Conference on Computer Vision*, 535–552.

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, 350–368.

Zhuo, L.; Fu, Y.; Chen, J.; Cao, Y.; and Jiang, Y.-G. 2022. Tgdm: Target guided dynamic mixup for cross-domain few-shot learning. In *Proceedings of the ACM International Conference on Multimedia*, 6368–6376.

Zou, Z.; and Shi, Z. 2017. Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images. *IEEE Transactions on Image Processing*, 27(3): 1100–1111.