Less is Better: Sparse Instance Learning for Cross-Domain Few-Shot Object Detection

Yali Huang[1*], Jie Mei[4*], Ziyi Wu[1], Yiming Yang[1], Hongru Zhao[1,2,3]
Mingyuan Jiu[1,2,3†], Hichem Sahbi[5]

[1]School of Computer and Artificial Intelligence, Zhengzhou University, China
[2]Engineering Research Center of Intelligent Swarm Systems, Ministry of Education, China
[3]National SuperComputing Center in Zhengzhou, Zhengzhou, China
[4]Dongfeng Commercial Vehicle Co.,Ltd, China
[5]Sorbonne University, CNRS, LIP6, F-75005, Paris, France
{hyl2024,yym3537}@gs.zzu.edu.cn, {wolletjohn,wuziyi5817}@gmail.com, {iemyjiu,zhaohongru}@zzu.edu.cn,
hichem.sahbi@lip6.fr

Abstract

Cross-Domain Few-Shot Object Detection (CD-FSOD) is an extremely challenging task due to the inherent data scarcity and substantial domain shift between the source and target domains. Existing methods often suffer from overfitting and noisy feature representations, which hinder the construction of discriminative class prototypes in the target domain. In this paper, we propose a novel framework with sparse instance learning (SI-ViTO) for CD-FSOD, which leverages instance sparsity to achieve a better detection with less representation. SI-ViTO adopts a dual-stage sparsity module, consisting of instance feature sparsity not only on the few-shot support images but also on the query images. This dual sparsity enables the model to effectively preserve salient foreground semantics and simultaneously to filter out redundant or noisy information. Furthermore, a new prototype calibration strategy is also used to dynamically refine the class prototypes with query instances to accelerate prototype adaptation. Extensive experimental results on CD-FSOD benchmarks show that SI-ViTO outperforms the state-of-the-art methods, demonstrating that less discriminative representations yield better cross-domain few-shot object detection performance than more abundant ones.

Code — https://github.com/hyali/SI-ViTO.git

Introduction

Cross-Domain Few-Shot Object Detection (CD-FSOD) aims to rapidly adapt models trained on large-scale source-domain data to novel object categories in a target domain using only a small number of labeled examples (Fan et al. 2020), as illustrated in Fig. 1. The core difficulties of CD-FSOD arise from two aspects: the high risk of overfitting caused by limited supervision, and feature distribution misalignment induced by domain shifts between the source and target domains (Fu et al. 2024).

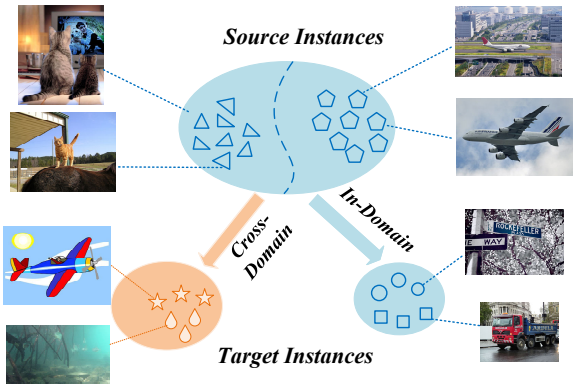

Figure 1: Illustration of Cross-Domain Few-Shot Object Detection task.

Prior research has explored various methodologies to tackle this intricate problem; however, each presents notable limitations. Meta-learning approaches, such as Meta R-CNN (Yan et al. 2019), aim to enhance generalization through meta-training on simulated tasks, but they often exhibit task-specific biases, leading to performance degradation when confronted with substantial domain discrepancies (Wang et al. 2025). Similarly, prevalent transfer learning techniques, including TFA (Wang et al. 2020), typically involve freezing backbone networks and fine-tuning detection heads, a strategy that frequently results in negative transfer due to its inherent neglect of explicit cross-domain feature alignment. Furthermore, advanced contrastive learning strategies, such as FSCE (Sun et al. 2021), enhance inter-class discriminability but amplify background noise interference under low-shot conditions due to their dense feature interaction mechanisms. Collectively, these limitations highlight critical shortcomings in the existing CD-FSOD methods, demonstrating insufficient suppression of redundant features within support samples, weak activation of critical regions in query images, and a notable absence of dynamic calibration mechanisms for class prototypes across

*These authors contributed equally.
†Correspondence author

varying domains. Consequently, these approaches exhibit substantial performance degradation, particularly in scenarios characterized by extreme domain shifts.

Recent efforts incorporating cross-domain alignment and sparse representation techniques partially alleviate these issues but remain inadequate. In cross-domain few-shot learning, mainstream methods extend meta-learning frameworks (e.g., cross-domain meta-feature distillation (Zhang et al. 2024)) or adopt feature disentanglement strategies (e.g., mutual information minimization (Cheng et al. 2023)), yet residual target-domain noise persists. The open-set detector in (Fu et al. 2024) improved domain adaptation but overlooked instance-level redundancy filtering, while a causal disentanglement framework in (Yin et al. 2025) enhanced generalization but inadequately calibrated prototype bias. In sparse representation, SparseFormer (Gao et al. 2023) improves recognition efficiency via latent token pruning but is not optimized for few-shot object detection. Similarly, cross-domain sparse regularization (Ji et al. 2023) is proposed, which benefits classification tasks but proves to be inadequate for the complex spatial localization in the detection. This mainly stems from the independent optimization of support and query features, where support-set noise contaminates prototype construction and spatial redundancy in query features degrades localization performance.

To address these critical gaps, we propose a novel sparse instance learning framework, SI-ViTO, specifically designed for the CD-FSOD task. Inspired by the human cognitive ability to focus on discriminative features while ignoring noise when learning novel concepts, SI-ViTO leverages dual sparse mechanism and prototype calibration to enhance feature robustness under low-shot constraints. Unlike single sparse methods such as SparseFormer (Gao et al. 2023), our proposed dual sparsity jointly mitigates noise sensitivity and information redundancy, demonstrably outperforming traditional dense fusion. At the same time, the prototype calibration strategy dynamically optimizes the alignment between query features and class prototypes using InfoNCE loss. To summarize, the key contributions of this work are:

– **Dual-Sparsity Mechanism.** We introduce a dual-sparsity mechanism that applies instance feature sparsification not only to few-shot support images but also to query images. This enables the model to retain salient foreground semantics while effectively filtering out redundant and noisy features, thereby alleviating overfitting problem and enhancing the robustness in few-shot scenarios.

– **Prototype Calibration Strategy.** We develop a dynamic prototype calibration strategy to refine class prototypes using query instances during adaptation. This enables the model to accelerate prototype alignment to the target domain and thus to improve detection performance in cross domain situation.

– Extensive experiments conducted on multiple CD-FSOD benchmarks demonstrate that SI-ViTO significantly outperforms state-of-the-art methods, confirming that leveraging sparse and discriminative representations, rather than abundant redundant features, lead to superior performance in cross-domain few-shot detection.

Related Work

Cross-Domain Few-Shot Object Detection. Few-shot object detection tackles recognition under severe data scarcity via meta-learning (Zhang et al. 2021; Han et al. 2023; Xin et al. 2024), transfer learning (Wang et al. 2020; Sun et al. 2021), and feature enhancement (Li et al. 2025), yet often fails under domain shifts. This motivates cross-domain few-shot object detection (CD-FSOD),which faces both limited data and domain discrepancy (Yue et al. 2021; Fu et al. 2024; Pan et al. 2025). Recent CD-FSOD research extends meta-learning through cross-domain meta-feature distillation (Zhang et al. 2024), contrastive meta-alignment (Wang et al. 2023), and task-specific lightweight adapters (Chen et al. 2022; Song et al. 2023) to improve parameter transfer efficiency.Feature disentanglement uses mutual information minimization (Cheng et al. 2023) and causal disentanglement (Yin et al. 2025) for domain-invariant representations. Prototype robustness primarily investigate noise suppression and calibration mechanisms (Cheng et al. 2021; Zhao et al. 2023), for instance, constraining prototype consistency through support-query contrastive learning (Fu et al. 2024) and generating domain-invariant category prototypes via modality alignment (Xiao, Wang, and Li 2024). Despite these advances, existing techniques typically lack explicit mechanisms for filtering redundant information during feature and prototype construction (Liang et al. 2022), limiting their effectiveness in preserving discriminative power across domains. This work bridges that gap through dual-sparsity mechanism and prototype calibration.

Sparse Representation Learning. Sparse representation learning enhances model discriminability and robustness by extracting low-dimensional structures or critical feature subsets from data with theoretical foundations (Elad 2010; Candès, Romberg, and Tao 2006). In recent years, this paradigm has achieved remarkable progress in object recognition and detection tasks. For generic detection, sparse convolutions (Liu et al. 2017), sparse attention mechanisms (Roh et al. 2021), and sparse activation in Transformers (Gao et al. 2023) effectively suppress background interference. In few-shot scenarios, improved prototype networks (Snell, Swersky, and Zemel 2017) incorporate sparse regularization to mitigate overfitting, while dynamic kernel methods enable instance-specific convolutional kernel sparsification (Ma et al. 2022). Addressing cross-domain challenges, recent studies explore domain-adversarial sparse coding (Li et al. 2021), task-adaptive sparse metric learning (Oreshkin, Rodríguez López, and Lacoste 2018), and learnable sparse pruning techniques (Ji et al. 2023). The proposed dual-sparsity mechanism in this work overcomes single-stage limitations through instance-level noise filtering and selective activation of query-specific feature regions. Following the "less is better" principle, by integrating a query-driven calibration strategy, our approach significantly enhances the detection performance for CD-FSOD task.

Sparse Instance Learning Framework

Preliminary. Our proposed framework builds upon CD-ViTO (Fu et al. 2024), a cross-domain detector designed

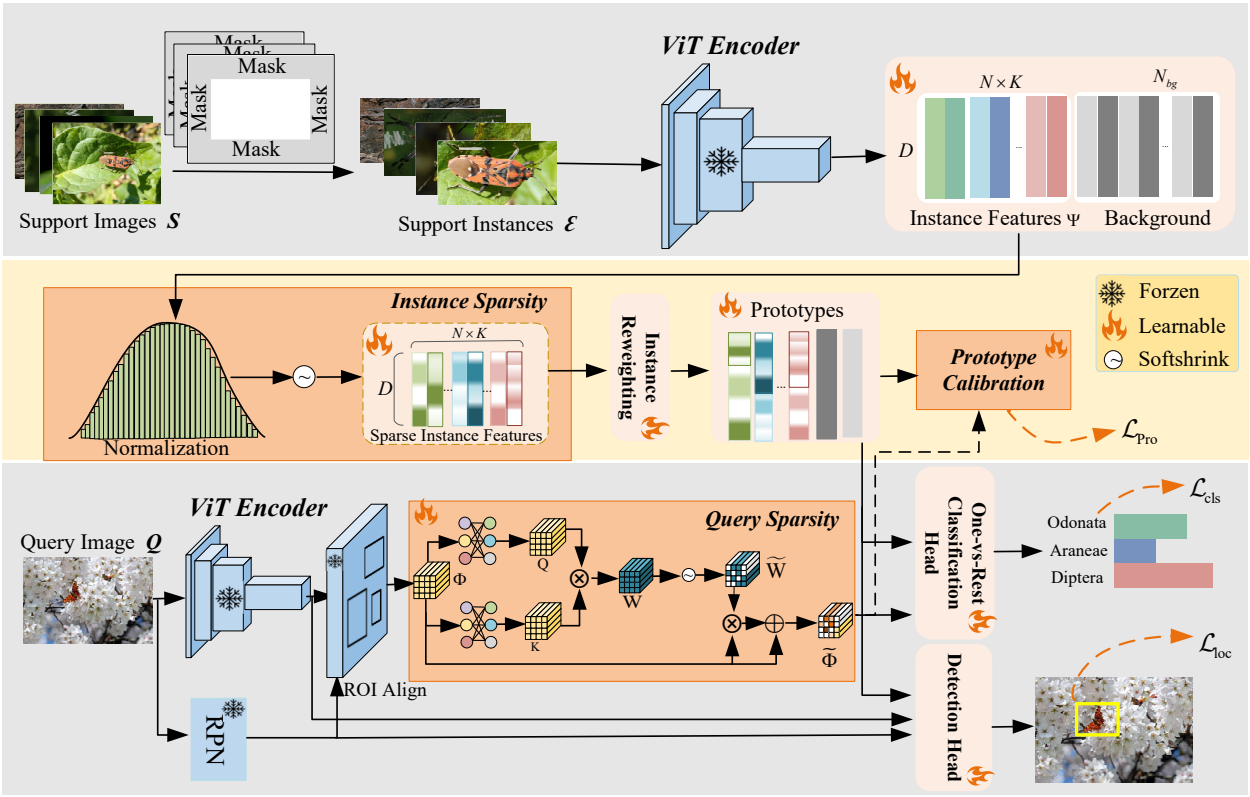

Figure 2: Overview of the proposed SI-ViTO. The novel contributions include Instance Sparsity, Query Sparsity and Prototype Calibration Strategy (i.e. $\mathcal{L}_{\text{Pro}}$).

to disentangle localization and classification tasks by leveraging visual features from a large pre-trained model. CD-ViTO's architecture primarily comprises a pre-trained Vision Transformer (ViT), an instance reweighting module, a prototype module, a Region Proposal Network (M_{RPN}), an ROI Align Module (M_{ROI}), a detection head (M_{DET}) and a one-vs-rest classification head (M_{CLS}). Specifically, given a query image Q and a set of support instances S, CD-ViTO firstly extracts instance features Ψ using the pre-trained ViT. Subsequently, class prototypes are derived via the instance reweighting module. For the query image Q, CD-ViTO sequentially applies the pre-trained ViT, M_{RPN}, and M_{ROI} to generate region proposals, visual features, and ROI features Φ. The M_{DET} then takes these region proposals, ROI features Φ, and the class prototypes as input to compute the localization loss $\mathcal{L}_{loc}$. Meanwhile, the M_{CLS} performs the classification task based on ROI features Φ and the class prototypes, yielding the classification loss $\mathcal{L}_{cls}$. The network is learned by minimizing a combination of loss $\mathcal{L}_{loc}$ and $\mathcal{L}_{cls}$.

To promote the detector to learn fewer key discriminative features and then boost the model robustness for the cross-domain few-shot detection, the proposed framework ameliorates the CD-ViTO by combining a **dual sparsity mechanism**, which is able to effectively filter out redundant and noisy features and improve the representation ability for the new class under few-shot conditions. Furthermore, to further boost the generalization capability of class prototypes in the target domain, a **prototype calibration strategy** is designed to accelerate the adaptation process in the target domain. In the following we respectively illustrate the details of the core components of SI-ViTO.

Dual-Sparsity Mechanism

To address the challenges of redundant and noisy features in CD-FSOD, we introduce a novel dual-sparsity mechanism, as illustrated in Fig. 2. This mechanism employs distinct sparsification strategies tailored to the unique characteristics of support instance and query features. By selectively suppressing irrelevant information, our dual-sparsity approach enables the detector to extract compact and highly informative representations, which is crucial for the novel class detection in the target domain.

Instance Sparsity Given a set of few-shot support images S from the target domain, we first generate semantic masks based on their label annotations. Specifically, the mask is used to suppress the background regions outside the labeled bounding boxes, thereby highlighting the semantic structure of the target objects. These masks are then used to perform semantic-aware augmentations on the support images, resulting in structurally refined and semantically enhanced support instances. We subsequently employ a Vision Transformer (ViT) to encode these instances and extract the initial support representations, which is enriched with source-

domain knowledge. These representations are then used to initialize the learnable instance features $\Psi \in \mathbf{R}^{N\times K\times D}$, where N denotes the number of novel classes, K is the number of support samples per class, and D is the feature dimensionality. This initialization strategy provides the model with a discriminative representation for the target domain support set, facilitating effective adaptation from the source domain to the target domain during the following fine-tuning.

After constructing the initial instance representations Ψ from the target domain support set, we aim to alleviate the adverse effects of distribution shift between the source and target domains. To this end, we perform statistical modeling on the target features extracted by the source-pretrained encoder. Specifically, we assume that the features in the target domain have a normal distribution, the feature-wise mean and variance can be estimated from the support set, and then the features are further normalized to obtain a semantically aligned representation $\hat{\Psi}$ with a standard normal distribution, which is computed as follows:

$$\hat{\Psi}_i = \frac{\Psi_i - \mu}{\sigma}, \quad \text{for} \quad i = 1, 2, \cdots, NK \tag{1}$$

where

$$\mu = \frac{1}{NK}\sum_{i=1}^{NK}\Psi_i, \quad \sigma = \sqrt{\frac{1}{(NK-\epsilon)}\sum_{i=1}^{NK}(\Psi_i - \mu)^2} \tag{2}$$

Here, ϵ is a correction factor that controls the bias of the variance estimation. In the following experiments, ϵ is set to be 1, leading to an unbiased estimation.

Since individual instance features $\hat{\Psi}$ often contain background clutter, textural noise, and redundant weak responses, directly aggregating them to form a class prototype tends to dilute the principal discriminative signal. Therefore, we employ the SoftShrink operator—the proximal mapping of the L_1 regularizer, continuous and differentiable except at the origin—to sparsify instance features $\tilde{\Psi}$. The operator is defined as follows:

$$\begin{aligned}\tilde{\Psi}_i &= SoftShrink(\hat{\Psi}_i, \zeta) \\ &= \text{sign}(\hat{\Psi}_i)\max(|\hat{\Psi}_i| - \zeta, 0)\end{aligned} \tag{3}$$

where ζ denotes the *instance sparsity ratio*, which controls the suppression of small-magnitude features, It drives noise coefficients to zero while only linearly shrinking salient activations, thus preserving discriminative information with minimal distortion. The operator's element-wise form is computationally efficient, requires no sorting, and integrates seamlessly into end-to-end training. This filtering yields more robust class prototypes with reduced estimation bias.

Query Sparsity To address the detection challenges posed by the severe scarcity of novel class samples in CD-FSOD, we not only construct more representative class prototypes via instance sparsity to enhance semantic modeling, but also seek to improve the model's ability to learn discriminative cues from novel-class queries. Motivated by this, we also propose a sparsification strategy within the detector that effectively filters out task-irrelevant redundancy and noise, encouraging the model to focus on more discriminative key information, thereby substantially improving the performance and robustness of object detection of novel class.

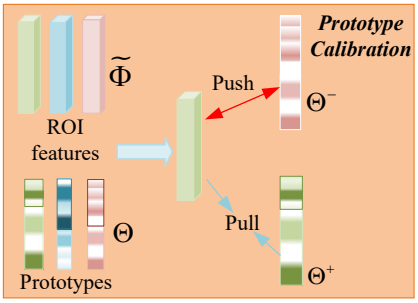

Figure 3: Detail of the prototype calibration strategy.

Given a query image Q, we firstly employ a Vision Transformer (ViT) and a Region Proposal Network (RPN) to extract the image's global feature representation and a set of candidate regions, respectively. Subsequently, the ROI align operation is applied to precisely align these candidate regions on the feature map, resulting in a set of ROI feature representations denoted as Φ. However, since Φ are extracted based on predefined region proposals, they often contain redundant or overlapping information. Additionally, some proposals may correspond to background areas containing noisy or irrelevant features. To mitigate these issues, a sparsification strategy based on an importance weight matrix $\mathbf{W}$ is proposed. Specifically, a relevance score is computed for each ROI and these weights are used to filter or re-weight the ROI features, thereby enhancing the representation of informative regions and suppressing background noise. The sparse ROI features $\tilde{\Phi}$ are computed as follows:

$$\mathbf{W} = \Xi_\alpha(\Phi)^\top \Xi_\beta(\Phi) \tag{4}$$

$$\begin{aligned}\tilde{\mathbf{W}} &= SoftShrink(\mathbf{W}, \eta) \\ &= \text{sign}(\mathbf{W})\max(|\mathbf{W}| - \eta, 0)\end{aligned} \tag{5}$$

$$\tilde{\Phi} = \tilde{\mathbf{W}}\Phi \tag{6}$$

where $\Xi_\alpha(\cdot)$ and $\Xi_\beta(\cdot)$ stand for trainable fully connected (FC) layers, $\mathbf{W} \in \mathbf{R}^{B\times 1\times 1\times 1}$ are the learned importance weights for each ROI (B is the batch size), and η is a predefined sparsity threshold.

Prototype Calibration Strategy

Given the prevalent data scarcity in the CD-FSOD, class prototypes derived from source domain knowledge often fail to effectively adapt to the feature distribution of the target domain during the fine-tuning phase on novel classes. This limitation weakens the detector's ability to acquire new class knowledge and deteriorates the overall performance. To address this issue, we propose a prototype calibration strategy that dynamically optimizes the matching relationship between query samples and class prototypes. Additionally, this strategy incorporates a finetuning-driven excitation mechanism to enhance the model's response to query samples, thereby improving the adaptability of prototypes to the target domain under few-shot constraints.

As illustrated in Fig.3, given the sparse ROI feature representation $\tilde{\Phi}$ of query image, and a set of class prototypes

	Method	Backbone	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Avg.
1-shot	Meta-RCNN (Yan et al. 2019)	ResNet50	2.8	-	7.8	-	-	3.6	/
	TFA w/cos (Wang et al. 2020)	ResNet50	3.1	-	8.0	-	-	4.4	/
	FSCE (Sun et al. 2021)	ResNet50	3.7	-	8.6	-	-	3.9	/
	DeFRCN (Qiao et al. 2021)	ResNet50	3.6	-	9.3	-	-	4.5	/
	Distill-cdfsod (Xiong 2023)	ResNet50	5.1	7.6	10.5	-	-	**5.9**	/
	VITDeT-FT (Li et al. 2022)	VIT-B/14	5.9	6.1	12.9	0.9	2.4	4.0	5.4
	Detic (Zhou et al. 2022)	VIT-L/14	0.6	11.4	0.1	0.9	0.0	0.0	2.2
	Detic-FT (Zhou et al. 2022)	VIT-L/14	3.2	15.1	4.1	9.0	3.8	4.2	6.6
	DE-VIT (Zhang et al. 2023)	VIT-L/14	0.4	0.5	2.7	0.4	0.4	1.5	1.0
	DE-VIT-FT (Zhang et al. 2023)	VIT-L/14	10.5	13.0	14.7	19.3	0.6	2.4	10.1
	CD-ViTO (Fu et al. 2024)	VIT-L/14	21.0	17.7	17.8	20.3	3.6	3.1	13.9
	IFC (Huang et al. 2025)	VIT-L/14	21.7	18.3	18.4	19.2	3.9	4.1	14.3
	SI-ViTO(ours)	VIT-L/14	**24.1**	**24.5**	**19.3**	**22.1**	**3.9**	4.0	**16.3**
5-shot	Meta-RCNN (Yan et al. 2019)	ResNet50	8.5	-	17.7	-	-	8.8	/
	TFA w/cos (Wang et al. 2020)	ResNet50	8.8	-	18.1	-	-	8.7	/
	FSCE (Sun et al. 2021)	ResNet50	10.2	-	8.6	-	-	9.6	/
	DeFRCN (Qiao et al. 2021)	ResNet50	9.9	-	18.9	-	-	9.9	/
	Distill-cdfsod (Xiong 2023)	ResNet50	12.5	23.3	19.1	15.5	**16.0**	**12.2**	16.4
	VITDeT-FT (Li et al. 2022)	VIT-B/14	20.9	23.3	23.3	9.0	13.5	11.1	16.9
	Detic (Zhou et al. 2022)	VIT-L/14	0.6	11.4	0.1	0.9	0.0	0.0	2.2
	Detic-FT (Zhou et al. 2022)	VIT-L/14	8.7	20.2	12.1	14.3	14.1	10.4	13.3
	DE-VIT (Zhang et al. 2023)	VIT-L/14	10.1	5.5	7.8	2.5	1.5	3.1	5.1
	DE-VIT-FT (Zhang et al. 2023)	VIT-L/14	38.0	38.1	23.4	21.2	7.8	5.0	22.3
	CD-ViTO (Fu et al. 2024)	VIT-L/14	47.9	41.1	26.9	22.3	11.4	6.8	26.1
	IFC (Huang et al. 2025)	VIT-L/14	49.4	41.7	27.5	**22.3**	11.3	7.2	26.6
	SI-ViTO(ours)	VIT-L/14	**52.1**	**43.2**	**27.6**	22.0	11.8	7.2	**27.3**
10-shot	Meta-RCNN (Yan et al. 2019)	ResNet50	14.0	-	20.6	-	-	11.2	/
	TFA w/cos (Wang et al. 2020)	ResNet50	14.8	-	20.5	-	-	11.8	/
	FSCE (Sun et al. 2021)	ResNet50	15.9	-	21.9	-	-	12.0	/
	DeFRCN (Qiao et al. 2021)	ResNet50	15.5	-	22.9	-	-	12.1	/
	Distill-cdfsod (Xiong 2023)	ResNet50	18.1	27.3	26.5	15.5	**21.1**	14.5	20.5
	VITDeT-FT (Li et al. 2022)	VIT-B/14	23.4	25.6	29.4	6.5	15.8	**15.6**	19.4
	Detic (Zhou et al. 2022)	VIT-L/14	0.6	11.4	0.1	0.9	0.0	0.0	2.2
	Detic-FT (Zhou et al. 2022)	VIT-L/14	12.0	22.3	15.4	17.9	16.8	14.4	16.5
	DE-VIT (Zhang et al. 2023)	VIT-L/14	9.2	11.0	8.4	2.1	1.8	3.1	5.9
	DE-VIT-FT (Zhang et al. 2023)	VIT-L/14	49.2	40.8	25.6	21.3	8.8	5.4	25.2
	CD-ViTO (Fu et al. 2024)	VIT-L/14	60.5	44.3	30.8	22.3	12.8	7.0	29.6
	IFC (Huang et al. 2025)	VIT-L/14	59.6	44.0	**30.9**	**22.3**	13.1	7.9	29.6
	SI-ViTO(ours)	VIT-L/14	**63.0**	**46.6**	30.7	21.9	13.2	7.0	**30.4**

Table 1: Performance in mAP of different methods with different backbones on the six CD-FSOD benchmarks. **Bold** numbers represent the best performance, "Avg." means the average result.

Θ, we pull $\tilde{\Phi}$ and its corresponding positive prototype (Θ^+) close to maximize their similarity and push $\tilde{\Phi}$ and the prototypes of negative classes (Θ^-) distant to minimize their similarity. This encourages the prototypes to undergo domain-adaptive refinement in the target domain. The optimization is guided by the InfoNCE loss function, which effectively enhances the discriminability of the prototypes and accelerates their adaptation to the target domain. The loss function is defined as follows:

$$\mathcal{L}_{\text{Pro}} = -\log \frac{\exp(\frac{\text{sim}(\tilde{\Phi},\Theta^+)}{\tau})}{\exp(\frac{\text{sim}(\tilde{\Phi},\Theta^+)}{\tau}) + \sum_{j=1}^{K-1} \exp(\frac{\text{sim}(\tilde{\Phi},\Theta_j^-)}{\tau})} \quad (7)$$

where Θ^+ is the positive prototype corresponding to the ground-truth class, $\{\Theta_j^-\}_{j=1}^{K-1}$ are the prototypes of the remaining $K-1$ negative classes, and τ is a temperature parameter that controls the sharpness of the distribution.

Experimental Results

To thoroughly evaluate the performance of the proposed framework, we conduct extensive experiments on the CD-FSOD benchmarks. Specifically, the model is first pretrained on the general-purpose object detection dataset(i.e. COCO), and then is evaluated on six representative cross-domain few-shot detection datasets including ArTaxOr (Drange 2019), Clipart1k (Inoue et al. 2018), DIOR (Inoue et al. 2018), DeepFish (Saleh et al. 2020), NEUDET (Song and Yan 2013), and UODD (Jiang et al. 2021). The experimental pipeline follows the standard "pretraining, finetuning, testing" paradigm. The base DE-ViT (Zhang et al. 2023) model pretrained on COCO is chosen as the result of the pretraining stage, and then the model is finetuned on novel class support sets from the target domains. SI-ViTO is developed with the same configuration with the base CD-ViTO: the learnable parameters are tuned around 80 epochs on 1-shot, and around 200 epochs on 5/10-shot. The SGD with a learning rate 0.002 is used as the optimizer. Through extensive experimental validation, the hyper-parameter ζ in the instance

Method	M_{IS}	M_{QS}	ArTaxOr	Clipart1K	DIOR	DeepFish	NEU-DET	UODD	Avg.
CD-ViTO	✗	✗	47.9	41.1	26.9	**22.3**	11.4	6.8	26.1
SI-ViTO	✓	✗	48.3	42.7	26.7	21.6	11.0	6.9	26.2
	✗	✓	51.3	41.9	26.3	20.8	11.4	6.3	26.3
	✓	✓	**52.1**	**43.2**	**27.6**	22.0	**11.8**	**7.2**	**27.3**

Table 2: Performance in mAP of different modules in the dual-sparsity mechanism in 5-shot setting. M_{IS} and M_{QS} respectively stand for instance sparsity and query sparsity.

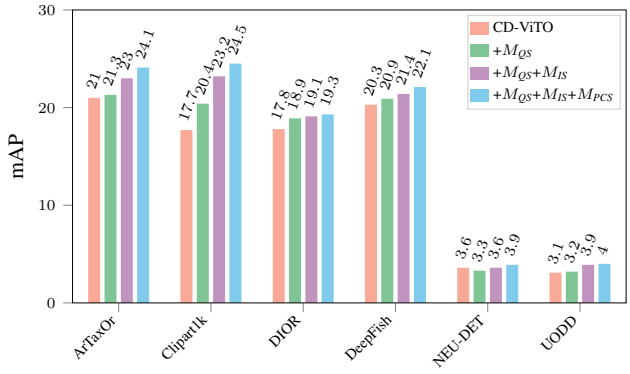

Figure 4: Performance comparison of the proposed method with different modules on CD-FSOD benchmarks in 1-shot setting. M_{QS}, M_{IS} and M_{PCS} respectively stand for query sparsity, instance sparsity and prototype calibration module.

sparsity, η in the query sparsity and temperature τ in prototype calibration are empirically set to be 0.2, 0.5 and 0.05 respectively. The performance is measured in AP and mAP. All the experiments are running on four RTX4090 GPUs.

Main Results on CD-FSOD Benchmarks

The results on six challenging datasets under the 1-shot, 5-shot, and 10-shot settings are shown in Tab. 1. In all these experiments, we take the average performance across 10 random runs. It can be clearly observed that our proposed SI-ViTO obtained state-of-the-art performance across nearly all the datasets and shot scenarios. In the extremely challenging 1-shot setting, SI-ViTO achieves an average mAP of 16.3, outperforming ViT-based methods such as CD-ViTO (13.9) and IFC (14.3). In particular, a performance gain of 6.3 is obtained on the Clipart1k dataset compared to IFC (Huang et al. 2025), where the image style differs significantly from the source domain, demonstrating its robustness in highly domain-shifted environments. In both the 5-shot and 10-shot settings, SI-ViTO also maintains stable and consistent improvements when more support samples are available, for instance, on the ArTaxOr dataset, SI-ViTO improves mAP from 47.9 to 52.1 in the 5-shot setting and from 60.5 to 63.0 in the 10-shot setting compared to the ones of CD-ViTO. The performance indicate that the proposed SI-ViTO not only exhibits stronger cross-domain generalization, but also obtains more gains from increasing support information, validating its effectiveness for cross-domain few-shot object detection task.

Ablation Study

Ablation on Dual-Sparsity Mechanism Tab. 2 presents the ablation results about instance sparsity and query sparsity in the dual-sparsity mechanism on the 5-shot setting. It can be seen that each module is able to obtain performance gains over the baseline CD-ViTO on the most target domains. When both instance and query sparsity are combined together, the best overall performance is achieved across almost all datasets. Notably, the model achieves the highest performance on ArTaxOr (52.1), Clipart1k (43.2), and NEU-DET (11.8), which clearly substantiate the complementary advantages of the dual sparsity mechanism: by suppressing high-variance cross-domain noise and spurious textures, it reduces the incidence of gradient spikes during training, yielding a flatter loss landscape and thereby enhancing stability under domain perturbations and few-shot fluctuations; concurrently, it compresses the discriminative subspace toward structurally and shape-centric, cross-domain robust patterns, attenuating color and fine-grained texture components that are prone to drift, and reducing cross-domain prototype deviation.

Module Ablation We further evaluate the impacts of three modules (i.e. instance sparsity, query sparsity and prototype calibration strategy). The ablation results on six representative cross-domain datasets are shown in Fig. 4. Firstly, incorporating the query sparsity yields consistent performance improvements across several datasets, indicating that sparse informative regions from query images enable to suppress redundant background noise. Secondly, integrating instance sparsity in support sets further enhances performance, demonstrating that sparse representative support features construct cleaner and more discriminative class prototypes, effectively mitigating overfitting under few-shot conditions. Finally, the prototype calibration strategy yields additional gains by dynamically optimizing query-prototype matching via the InfoNCE loss, serving as a fine-tuning-driven excitation mechanism that improves prototype adaptability to the target domain. In summary, the three modules contribute complementarily at different stages of the detection pipeline, ensuring SI-ViTO effectively enhances prototype quality and generalization for cross-domain few-shot object detection.

ζ	0	0.1	0.2	0.3	0.5	0.7	0.9
Sparsity ratio	0%	7.7%	15.3%	23.0%	37.6%	51.2%	63.3%
AP	23.6	23.8	**24.3**	23.7	23.9	23.5	23.3

Table 3: Sparsity ratio and AP of different thresholds ζ with η=0.5 in the proposed dual-sparsity mechanism on the Clipart1k dataset in the 1-shot setting.

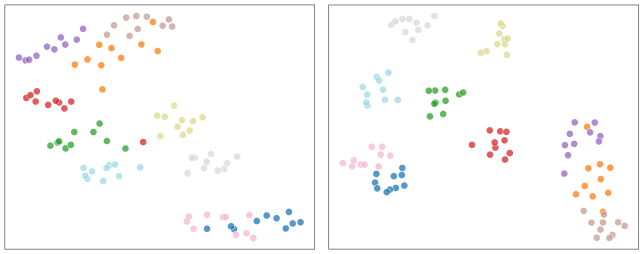

Figure 5: t-SNE visualization of CD-ViTO (left) and SI-ViTO (right). Each point denotes an instance, and different colors indicate different object categories.

Analysis

Sparsity Analysis To investigate the impact of sparsity thresholds in the dual-sparsity mechanism, we reports the overall sparsity rate and detection performance when varying instance sparsity thresholds ζ with the query sparsity parameter η=0.5 in Tab. 3. The sparsity ratio is calculated as the number of zero-value features divided by the total number of features. As the threshold increases from 0 to 0.9, the instance sparsity ratio rises progressively, reaching a maximum of 63.3%. However, the optimal performance is observed at a threshold of 0.2, corresponding to a sparsity ratio of 15.3%, where AP reaches its peak at 24.3. These results indicates that a moderate level of sparsity is better to remove redundant instance features while retaining essential discriminative information. In contrast, excessive sparsity may lead to the loss of valuable features, resulting in performance degradation. The results highlight the importance of balancing sparsity and information preservation in the dual-sparsity mechanism.

Visualization To further demonstrate the effectiveness of the proposed SI-ViTO, we first visualize the learned instance feature distributions of both the base CD-ViTO and our SI-ViTO on Clipart1k dataset in the 10-shot setting. As illustrated in Fig. 5, SI-ViTO produces noticeably more compact intra-class clusters and clearer inter-class separations, highlighting its superior capability in suppressing noise and calibrating feature prototypes. We also visualize several instances of our proposed SI-ViTO detector in 10-shot setting. As shown in Fig. 6, we compare the localization and classification results of SI-ViTO against the base CD-ViTO and groundtruth. From the qualitative comparisons, we observe that CD-ViTO often tends to detect irrelevant regions and produce comparable large and meaningless boxes. In contrast, our SI-ViTO significantly improves both localization precision and semantic accuracy. By leveraging the dual-sparsity mechanism, it is able to capture discriminative features while suppressing noise, leading to better predictions that are more consistent with the groundtruth.

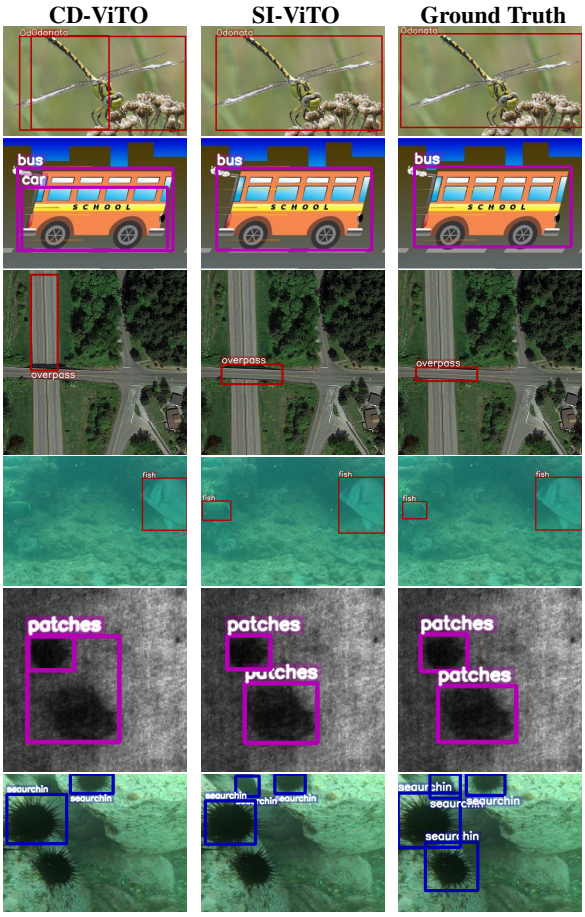

Figure 6: Visualization comparison of detection results of image instances on the six CD-FSOD benchmarks in the 10-shot setting (from top to bottom: ArTaxOr, Clipart1K, DIOR, DeepFish, NEU-DET and UODD dataset).

Conclusion

This paper presents SI-ViTO, a sparse instance learning framework for cross-domain few-shot object detection. By applying dual sparsity to support and query instances, the method effectively suppresses cross-domain noise, while dynamic prototype calibration refines class representations to improve adaptability under domain shifts and sample scarcity. Experiments on multiple benchmarks demonstrate consistent gains over strong baselines, confirming the robustness and stability of the approach. Ablation studies highlight the complementary roles of dual sparsity and prototype calibration. Despite its strong overall performance, SI-ViTO still faces challenges on extremely degraded domains such as UODD due to low-quality imagery and weak cues. Future work will investigate uncertainty-aware prototype aggregation, structure-aware sparsity for tiny or blurred objects, and continual cross-domain adaptation.

Acknowlegements

This work is supported by the National Natural Science Foundation of China (No.s 62272422, U22B2051, 62502461), and also partially by the Natural Science Foundation of Henan Province (No. 252300421225) and Organized Young Scientific Research Team Cultivation Foundation of Zhengzhou University (No. 35220549).

References

Candès, E. J.; Romberg, J.; and Tao, T. 2006. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2): 489–509.

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35: 16664–16678.

Cheng, H.; Wang, Y.; Li, H.; Kot, A. C.; and Wen, B. 2023. Disentangled feature representation for few-shot image classification. *IEEE transactions on neural networks and learning systems*, 35(8): 10422–10435.

Cheng, Y.; Wei, F.; Bao, J.; Chen, D.; Wen, F.; and Zhang, W. 2021. Dual path learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9082–9091.

Drange, G. 2019. Arthropod taxonomy orders object detection dataset. In *https://doi.org/10.34740/kaggle/dsv/1240192*.

Elad, M. 2010. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media.

Fan, Q.; Zhuo, W.; Tang, C.-K.; and Tai, Y.-W. 2020. Few-shot object detection with attention-RPN and multi-relation detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4013–4022.

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. Springer.

Gao, Z.; Tong, Z.; Wang, L.; and Shou, M. Z. 2023. Sparseformer: Sparse visual recognition via limited latent tokens. *arXiv preprint arXiv:2304.03768*.

Han, J.; Ren, Y.; Ding, J.; Yan, K.; and Xia, G.-S. 2023. Few-shot object detection via variational feature aggregation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 755–763.

Huang, Y.; Mei, J.; Yang, Y.; Guo, M.; Jiu, M.; and Xu, M. 2025. Instance Feature Caching for Cross-Domain Few-Shot Object Detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, 1567–1575.

Inoue, N.; Furuta, R.; Yamasaki, T.; and Aizawa, K. 2018. Cross-Domain Weakly-Supervised Object Detection through Progressive Domain Adaptation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ji, F.; Chen, Y.; Liu, L.; and Yuan, X.-T. 2023. Cross-domain few-shot classification via dense-sparse-dense regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3): 1352–1363.

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 29th ACM International Conference on Multimedia*.

Li, Y.; Liu, Q.; Jin, Z.; Wei, J.; Nie, J.; and Fu, Y. 2025. MAFE R-CNN: Selecting More Samples to Learn Category-aware Features for Small Object Detection. *arXiv preprint arXiv:2505.16442*.

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. Springer.

Li, Y.; Ren, J.; Liu, J.; and Chang, Y. 2021. Deep sparse autoencoder prediction model based on adversarial learning for cross-domain recommendations. *Knowledge-Based Systems*, 220: 106948.

Liang, K. J.; Rangrej, S. B.; Petrovic, V.; and Hassner, T. 2022. Few-shot learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9089–9098.

Liu, Z.; Li, J.; Shen, Z.; Huang, G.; Yan, S.; and Zhang, C. 2017. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE international conference on computer vision*, 2736–2744.

Ma, R.; Fang, P.; Avraham, G.; Zuo, Y.; Zhu, T.; Drummond, T.; and Harandi, M. 2022. Learning instance and task-aware dynamic kernels for few-shot learning. In *European Conference on Computer Vision*, 257–274. Springer.

Oreshkin, B.; Rodríguez López, P.; and Lacoste, A. 2018. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems*, 31.

Pan, J.; Liu, Y.; He, X.; Peng, L.; Li, J.; Sun, Y.; and Huang, X. 2025. Enhance then search: An augmentation-search strategy with foundation models for cross-domain few-shot object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1548–1556.

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.

Roh, B.; Shin, J.; Shin, W.; and Kim, S. 2021. Sparse detr: Efficient end-to-end object detection with learnable sparsity. *arXiv preprint arXiv:2111.14330*.

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*.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

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*, 858–864.

Song, L.; Xue, R.; Wang, H.; Sun, H.; Ge, Y.; Shan, Y.; et al. 2023. Meta-adapter: An online few-shot learner for vision-language model. *Advances in Neural Information Processing Systems*, 36: 55361–55374.

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.

Wang, H.; Lv, D.; Lin, T.; Han, C.; and Song, L. 2025. Task-adaptive unbiased regularization meta-learning for few-shot cross-domain fault diagnosis. *Engineering Applications of Artificial Intelligence*, 144: 110200.

Wang, S.; Tan, Z.; Liu, H.; and Li, J. 2023. Contrastive meta-learning for few-shot node classification. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, 2386–2397.

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*.

Xiao, K.; Wang, Z.; and Li, J. 2024. Semantic-guided robustness tuning for few-shot transfer across extreme domain shift. In *European Conference on Computer Vision*, 303–320. Springer.

Xin, Z.; Chen, S.; Wu, T.; Shao, Y.; Ding, W.; and You, X. 2024. Few-shot object detection: Research advances and challenges. *Information Fusion*, 107: 102307.

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. IEEE.

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.

Yin, Y.; Yang, J.; An, N.; and Li, L. 2025. CDC-FSL: A Causal De-Confounding framework for Few-Shot Learning. *Knowledge-Based Systems*, 113732.

Yue, X.; Zheng, Z.; Zhang, S.; Gao, Y.; Darrell, T.; Keutzer, K.; and Vincentelli, A. S. 2021. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13834–13844.

Zhang, D.; Yan, H.; Chen, Y.; Li, D.; and Hao, C. 2024. Cross-domain few-shot learning based on feature adaptive distillation. *Neural Computing and Applications*, 36(8): 4451–4465.

Zhang, G.; Luo, Z.; Cui, K.; and Lu, S. 2021. Meta-detr: Few-shot object detection via unified image-level meta-learning. *arXiv preprint arXiv:2103.11731*, 2(6): 2.

Zhang, X.; Liu, Y.; Wang, Y.; and Boularias, A. 2023. Detect everything with few examples. *arXiv preprint arXiv:2309.12969*.

Zhao, Y.; Zhang, T.; Li, J.; and Tian, Y. 2023. Dual adaptive representation alignment for cross-domain few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 11720–11732.

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. Springer.