Free-Lunch Augmentation by Revisiting Diffusion-Based Data Generation for Cross-Domain Few-Shot Object Detection

Zijian Zhuang†, Yixiong Zou†⋆, Yuhua Li, and Ruixuan Li

School of Computer Science and Technology, Huazhong University of Science and Technology, China
{zhuangzijian, yixiongz, idcliyuhua, rxli}@hust.edu.cn

Abstract. Cross-Domain Few-Shot Object Detection (CDFSOD) aims to transfer knowledge from data-rich upstream generic domains to downstream expert domains using scarce training data, where the significant domain gap and data scarcity make it an unsolved challenge. To address this problem, we revisit a natural yet underexplored approach in CDFSOD: data augmentation, by directly synthesizing data through diffusion models to supplement limited training samples. However, due to large domain gaps, we find that current diffusion methods cannot produce good results, leading to performance even lower than using the original images. To address these limitations, we divide the domain gaps into visual gaps and semantic gaps for separate analysis. For the visual gap, we find that the diffusion model cannot distinguish noise from useful information on expert domains, which can be mitigated by adding weakened noise. For the semantic gap, we find that the background semantics shows much smaller gaps between domains than foreground semantics, and we can bridge this gap by background inpainting. Based on the above analysis, we propose a method (Selective Inpainting with Tailored Noise, SITN) to dynamically take different strategies for downstream data synthesis based on their different gaps from the general domain, including a Generation Module for adding tailored noise and a Selection Module to dynamically select the inpainting regions. Extensive experiments on 6 datasets of CDFSOD and 4 datasets of cross-domain few-shot segmentation (CDFSS) validate that we can synthesize helpful data, achieving new state-of-the-art performance. Our codes is available at https://github.com/zzzzj311-droid/Free-Lunch-SITN.

Keywords: Cross-domain · Generative Model · Objection-detection

1 Introduction

In real-world scenarios, obtaining sufficient data is often challenging for downstream expert-domain tasks [52] [44] [54], such as medical analysis [51] [19] [53]. To handle this problem, Cross-Domain Few-Shot Object Detection (CDFSOD) [43]

⋆ Corresponding author, † Equal contribution

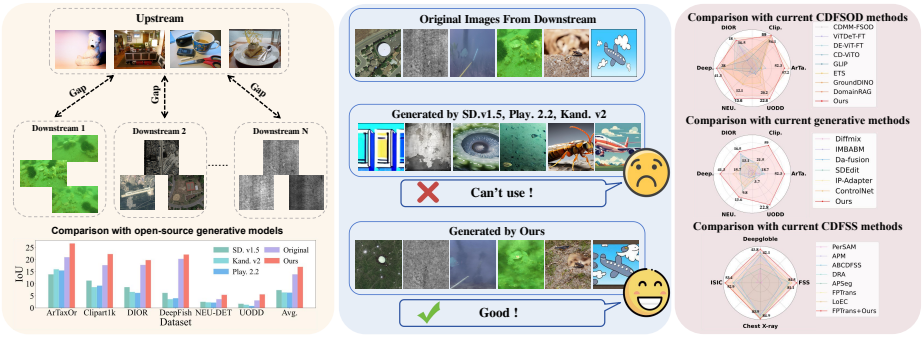

Fig. 1: (Left-top) Cross-Domain Few-Shot Object Detection learns from scarce training data on downstream expert domains, with knowledge generalized from upstream general domains, where domain gaps and data scarcity make it challenging. (Left-bottom & Mid) To address it, we revisit a natural but ignored approach in CDFSOD (i.e., data augmentation) to directly supplement the scarce training data with diffusion models. However, due to large domain gaps, we find that current methods can hardly synthesize satisfactory samples, leading to a performance even lower than merely using the original data, which we aim to address in this paper. (Right) While existing diffusion-based methods are not directly applicable to CDFSOD, our method not only addresses this task effectively but also seamlessly extends to CDFSS. It achieves new state-of-the-art results on both benchmarks, demonstrating its versatility and superiority.

[22] has been proposed to transfer models pretrained on large-scale upstream datasets to downstream data-scarce expert datasets for object detection, where the main challenge lies in both the downstream data scarcity and domain gaps between downstream and upstream datasets.

To handle this problem, current state-of-the-art methods primarily focus on transfer learning or efficient fine-tuning. However, driven by recent progress in AIGC [9], a straightforward method for directly addressing the data scarcity is still under-explored: **directly synthesizing the scarce training data**. However, it also faces great challenges. As in Fig. 1, we find that directly applying standard image-to-image or text-to-image generation cannot produce suitable images due to large domain gaps. Take Stable Diffusion v1.5 [35] for example, although its training data covers a wide range of images, it still cannot contain sufficient downstream expert-domain data, such as medical data, which are expensive to obtain and may not be open to society due to privacy issues. Therefore, the generated data proves unsatisfactory, thereby degrading the performance.

In this paper, we aim to address this data synthesis problem in diffusion models caused by huge domain gaps. We divide the domain gap into the visual gap (e.g., general images vs. X-Ray images) and the semantic gap (e.g., COCO classes vs. medical classes), and analyze them separately.

For the visual gap, given expert-domain datasets, we find that the difference between synthesized data and the original data is much larger than that on the general-domain datasets, which we explain as the model's inability to distinguish noise and useful information on expert domains. Then, we find that adding just

weak noises can effectively maintain the useful information in the given expert-domain images, mitigating the negative effect of visual gaps.

For the semantic gap, as most categories do not appear in the diffusion model, the model can hardly comprehend the unknown semantics. However, we then find these categories are mostly in the foreground of the image, but the background is more likely to be shared across domains. For example, a general domain may contain a dog on a grassland, and a target domain may contain an alien against a green background. Indeed, the foreground object (dog vs. alien) shows large semantic gaps, but their backgrounds (grassland vs. green background) are more likely to share similar patterns. Therefore, it is possible to utilize background inpainting to circumvent the synthesis of unknown semantics.

Based on this analysis, we propose our method (Selective Inpainting with Tailored Noise, SITN) to tailor current pretrained diffusion models for data synthesis of expert-domain data. Specifically, our method contains a Generation Module, which synthesizes images with noises tailored (e.g., weakened) for mitigating the negative effect of domain gaps, and a Selection Module, which dynamically selects whether the background inpainting or the foreground inpainting strategy should be adopted. As shown in Fig. 1, our method successfully supplements data for the CDFSOD task and achieves new state-of-the-art performance on six benchmark datasets. Our contribution can be listed as

- To the best of our knowledge, we are the first to study the data synthesis problem under large domain gaps for CDFSOD. *Experiments show that our conclusion also fits the cross-domain few-shot segmentation (CDFSS) task.*
- By extensive experiments, we propose to take weakened noises to reduce the negative effect of visual gaps, and utilize background inpainting to handle semantic gaps.
- We propose a method (Selective Inpainting with Tailored Noise, SITN) to dynamically take different strategies for downstream data synthesis based on their different gaps from the general domain.
- Extensive experiments validate that we can successfully synthesize helpful data for the CDFSOD task and even the CDFSS task, achieving new state-of-the-art methods on 6 CDFSOD datasets and 4 CDFSS datasets.

2 Related Work

Cross-Domain Few-Shot Object Detection (CDFSOD) aims to effectively learn the target domain using only a small amount of training samples [10], where the domain gap and scarce training data make it challenging. [2] proposes a prototype-based few-shot object detection method specifically designed to address the scarcity of data in satellite imagery. [13] proposes the UPPR method to address the problems of base class forgetting and poor novel class detection in Generalized Few-Shot Object Detection. FM-FSOD [14] integrates the DINOv2 visual backbone with the contextual learning capabilities of large language models. However, data synthesis is still under-explored in CDFSOD.

Diffusion Models are to learn the distribution of the training dataset via denoising, thereby generating images of the same distribution [4]. Current diffusion-based methods relevant to our task focus on source-domain training for transfer learning [21], Multi-sample semantic segmentation [38], and Multi-sample image inpainting [8]. For generation fields, most methods are based on image-to-image generation [13, 27] with whole-noised information, text-to-image generation [12, 36, 48]. However, applying diffusion models to handle the data scarcity problem for CDFSOD is still underexplored, and our method is the first to hold the cross-domain few-shot data synthesis under large domain gaps.

Data Augmentation is widely adopted by current works. Recent advances in diffusion models have significantly enhanced data augmentation techniques for image generation and semantic editing. DA-Fusion [42] employs a pre-trained text-to-image diffusion model to perform semantic editing on original images. IMBABM [15] systematically investigates the applicability of synthetic data generated by text-to-image models in image recognition tasks. DIFFUSEMIX [20] employs diffusion models to create augmented images through generation, splicing, and fractal mixing processes. IP-Adapter [45] is a lightweight adapter that adds image prompt capabilities to pre-trained text-to-image diffusion models. ControlNET [47] locks the parameters of the original model and trains a trainable copy of its encoding layers. **SDEdit** [32] formulates the image editing task via a stochastic differential equation (SDE) to simulate the reverse-time stochastic process. However, its generation process is inherently constrained by the training data distribution, struggling to generalize to cross-domain tasks (e.g., editing real-world photos using medical image priors), and its backbone network is limited in scale. Even if retrained on cross-domain datasets, the stochastic gradient descent optimization within the SDE paradigm remains computationally prohibitive. In contrast, our method is training-free and plug-and-play. It is not only compatible with existing open-source generative priors to mitigate domain gaps but also significantly outperforms SDEdit in inference efficiency, overcoming the dual limitations of SDE-based methods in domain generalization and computational cost. **DomainRAG** [26] adopts a retrieval-augmented pipeline that heavily relies on large-scale source-domain datasets such as COCO, requiring background image screening and ResNet-50 feature extraction, which significantly increases inference time. In contrast, our method is based on the purely generative DDIM paradigm, eliminating the need for external retrieval and feature matching, thereby achieving substantially higher generation efficiency and making it more suitable for real-time or large-scale applications.

3 Why diffusion models fail in cross-domains?

3.1 Negative Effect of Domain Gaps

Fig. 1 shows that images synthesized by current methods are largely different from the original ones. To delve into this phenomenon, we first plot images synthesized on source domains and target domains in Fig. 2 (Left). We can see that the difference between the synthesized and original images is much smaller than that in the target domains.

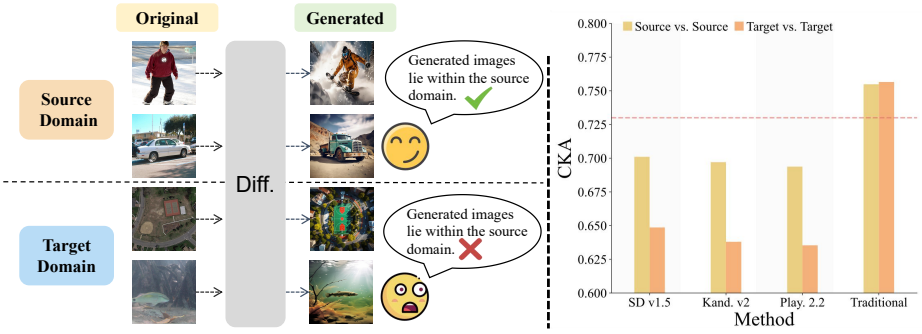

Fig. 2: (Left) Synthesized images on source domains well preserve the semantics of the original images, but largely destroy the semantics on the target domains. (Right) To quantify such a difference between domains, we use the CKA similarity to measure the semantic similarity between synthesized and original images. We can see the traditional method (e.g., flip, rotation) shows consistent CKA similarities across domains, verifying its robustness to domains, but diffusion-based ones suffer from a large decline in target domains, showing the difficulty brought by domain gaps.

To quantify this difference, we follow [6] to take the CKA (Centered Kernel Alignment) similarity to measure the similarity between synthesized and original images. We use a vision encoder to extract features from images and calculate the CKA similarity between them. Here, we compare mainstream generative models (including Stable-Diffusion v1.5, Playground v2, and Kandinsky 2.2) and the traditional image augmentation methods (i.e., horizontal/vertical flipping). As in Fig. 2 (Right), for the traditional augmentation methods, there is a consistency between the source and target domains, which naturally serves as an "oracle" for other synthesis methods due to its wide adoption[1].

However, for other diffusion-based models, there exists a huge gap between the CKA on source domains and target domains, indicating the challenges in data generation. That is, we hypothesize that the model can hardly distinguish noise from useful visual information in target-domain images; therefore, it just "randomly" guesses what to synthesize based on its pre-training information. However, the pre-training information also largely differs from the target-domain information. As a result, it cannot successfully supplement the scarce training data in Fig. 1.

In the following subsections, we will divide the domain gap into the visual gap (e.g., natural images vs. underwater images) and the semantic gap (e.g., ImageNet classes vs. medical classes), and analyze them separately.

3.2 Visual Gaps

To verify our hypothesis about the noisy information, we then try to control the strength of noise added to the diffusion model and evaluate the CKA similarity between the synthesized and original images.

[1] There is a trade-off between maintaining original information and introducing augmented information. Traditional methods are effective at the former but less effective at the latter; therefore are not the optimal choice.

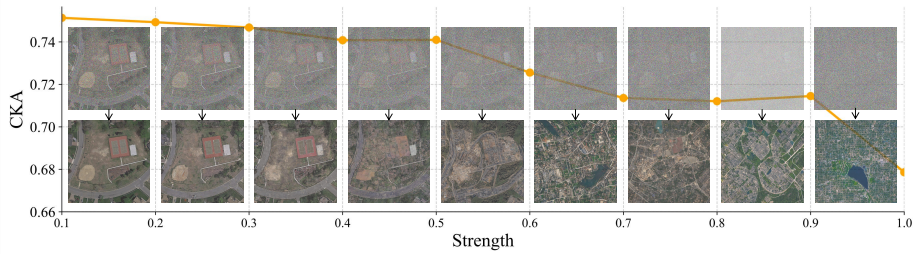

Fig. 3: The relationship between noise strength and CKA similarity shows that low noise strength effectively preserves the semantics of the original image (i.e., similarity only slightly decreases), while high noise levels severely disrupt semantics (i.e., generated images differ greatly from the original one), indicating diffusion models can hardly recover target-domain information from the noise.

Specifically, diffusion models generate images in two stages. In the forward stage, noise is gradually added to the image, progressively obscuring the information in the image. In the backward stage, the diffusion model gradually removes the noise and utilizes the knowledge learned from the source domain to infer the obscured information. Therefore, we can control the noise added in the forward stage and use the backward stage to generate images.

Fig. 3 shows the generated images and the corresponding CKA. By gradually increasing the strength of noise, the generated images gradually deviate from the original images, with more information synthesized by the diffusion model. Simultaneously, the CKA value consistently decreases. In contrast, by reducing the strength of noise, more semantic information is preserved, and the CKA gradually increases towards that of the traditional augmentation methods in Fig. 3. By controlling noise strength, we could directly guide the diffusion model to preserve content and bridge visual gaps, verifying our hypothesis.

3.3 Semantic Gaps

For the semantic gap, since the target-domain category may not appear in the diffusion model's training data, it is difficult for the diffusion model to synthesize such an unknown category. However, we find that an intuition naturally holds that the category information is mostly relevant to the foreground object, but for the background part, it is easier to be shared across domains. For example, a general domain may contain a dog on a grassland, and a target domain may contain an alien against a green background. Indeed, the foreground object (dog vs. alien) shows large semantic gaps, but their backgrounds (grassland vs. green background) are more likely to share similar patterns. Therefore, it is possible to utilize background inpainting to circumvent the synthesis of unknown semantics.

To verify this intuition, in Fig. 4 (Left), we try to use the bounding box annotation to crop the foreground and background parts in the target-domain images, and use diffusion models to synthesize the foreground and background images. Then, we also evaluate the CKA similarity between the synthesized and original images. We can see that for different diffusion models, the background CKA is consistently much higher than that of the foreground, verifying our

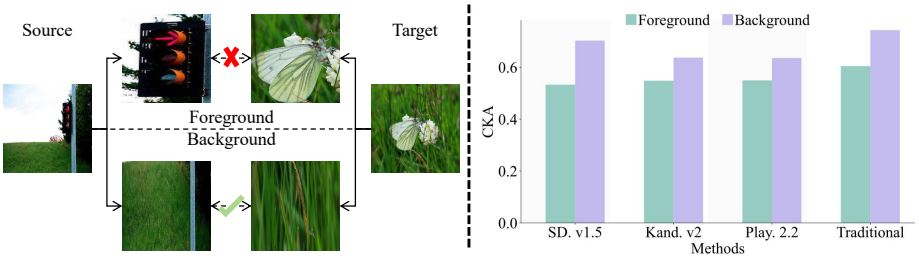

Fig. 4: (Left) Compared with the foreground relevant to unseen classes, the background is easier to transfer across domains. (Right) To verify this intuition, we crop the foreground and background for data synthesis and find that the background CKA is consistently higher than that of the foreground, indicating it is possible to circumvent the semantic gap by background inpainting.

intuition that the background inpainting could be much easier on the target domains, even when the semantic gap is large.

4 Method

Based on the above analysis, we can see that both the visual gap and the semantic gap make diffusion models fail to synthesize for the CDFSOD task. Moreover, we find that controlling the noise strength in diffusion models can effectively reduce the visual gap, while employing background generation can minimize the semantic gap between synthesized images and original ones. Therefore, we further propose a novel image generation method capable of producing images that closely resemble the original CDFSOD dataset in both visual and semantic features, which effectively supplements training data for the CDFSOD task.

4.1 Problem Definition

Cross-Domain Few-Shot Object Detection There exists a source domain dataset D_{source} and a target domain dataset D_{target}, which have classes C_{source} and C_{target} respectively, and we have $C_{source} \cap C_{target} = \phi$. Typically, the model is initialized from a model pre-trained on $D_{source} = \{x_i^s, y_i^s\}_{i=1}^{N_S}$ which is used for acquiring source-domain knowledge where $y_i^s \in C_{source}$. Then, the object detection model obtained after training on the source-domain dataset is transferred to the target-domain dataset $D_{target} = \{x_i^t, y_i^t\}_{i=1}^{N_T}$where $y_i^t \in C_{target}$. For each target-domain class, there are only 1, 5, or 10 training samples, which makes the fine-tuning training in the target domain challenging. For fairness, in the target-domain fine-tuning stage (our main focus), we adopt the setting of [10], with the support set $S = \{x_S, y_S\}$ and the query set fixed. The number of classes in each target-domain dataset is fixed, and each class has only 1, 5, or 10 samples. After fine-tuning on the support set (our main focus), the model is evaluated on the query set $Q = \{x_Q, y_Q\}$.

Generative Models The core objective of generative models is to learn the underlying probability distribution $p_\theta(x)$ of real data $x \sim p_{data}(x)$, and to be able

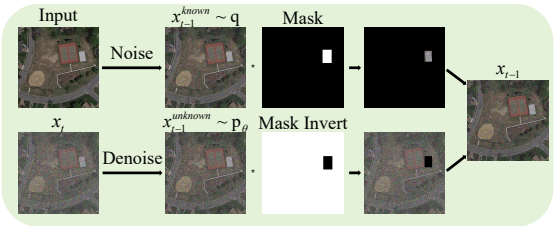

Fig. 5: Inpainting at timestep t, where the intermediate result x_{t-1}^{known} from the forward process and the prediction $x_{t-1}^{unknown}$ from the reverse process are jointly synthesized to generate the image x_{t-1} in the backward process.

to sample new data x_{new} from it [31]. Meanwhile, it ensures that the generated samples are both of high quality (realistic) and can cover the diverse patterns of the data distribution. In this work, firstly, we input the original image into a large language model (LLM) to obtain a textual description of the image, and then use existing generative models to generate images via a text-and-image-to-image process.

Finetuning In the fine-tuning phase, we specifically optimize the parameters of the detection head, classification head, and newly added modules for the target domain support set while keeping the backbone network frozen. Finally, the model performance is evaluated on the target domain test set. In this work, we use CD-ViTO [10] for fine-tuning in experiments. Our method is general and also compatible with alternative fine-tuning schemes.

4.2 Inpainting in Diffusion Models

Inpainting in diffusion models fills in missing, occluded, or damaged areas in an image [30], which involves progressively reconstructing the missing parts while preserving known regions through the backward process of diffusion models (As shown in Fig. 5). Specifically, the model first corrupts the intact image by adding noise. During the reverse denoising process, it treats the undamaged areas as conditional constraints and guides the generation of missing regions through mechanisms like cross-attention. At each denoising step, the system enforces the retention of known pixels while predicting exclusively within the masked areas, ensuring semantic, structural, and textural consistency between the generated content and its surroundings.

4.3 Selective Inpainting with Tailored Noises

SITN operates through three sequential stages: initially obtaining generative samples via the Generation Module, subsequently filtering the generated samples through the Selection Module to construct an expanded support set as shown in Fig. 6, and ultimately applying it to downstream object detection tasks for both training and inference purposes.

Generation Module As in Section 2, the style of target-domain images significantly differs from the source domain, with backgrounds being more transferable

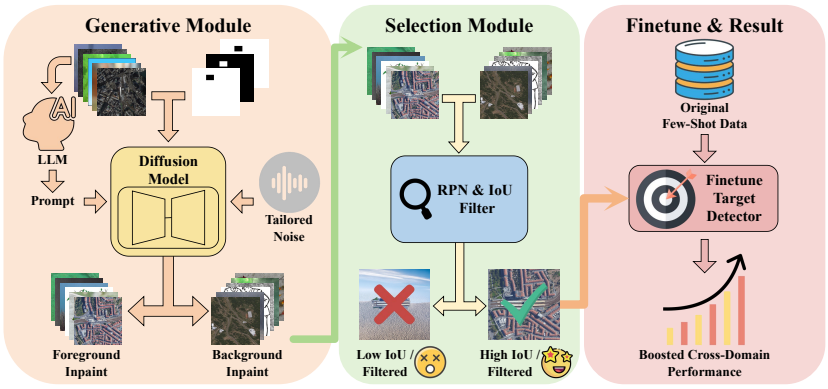

Fig. 6: Our method operates solely on the target domain and consists of three stages. In the first stage, we feed images into an LLM to obtain prompts, extract bounding boxes from annotations, and compute their masks. In the second stage, the prompts, original images, and bounding box information are input into a diffusion model to perform inpainting on both foreground and background regions of the images. During the third stage, the generated images are fed into an encoder to calculate Intersection over Union (IoU) metrics. Through a selection module, we select higher-quality samples to be incorporated into the target detection model for target-domain finetuning.

than foregrounds. Therefore, apart from semantic regions of images, we also mask the background regions with tailored noise and use diffusion models to inpaint these regions, which is used as the main augmentation. The specific approach is as follows:

1. Semantic Description Generation: An LLM extracts a global description of original support-set images x_{sup} as the input *prompt* for the diffusion model.

$$prompt = M_{LLM}(x_{sup}) \tag{1}$$

2. Target Region Localization: Based on ground truth annotations from the dataset, the object detection bounding box (BBox) is extracted as the mask for inpainting.

$$mask_{sup} = \begin{cases} 1, & \text{if the pixel is in } y_{sup} \\ 0, & \text{if the pixel is not in } y_{sup} \end{cases} \tag{2}$$

3. Diffusion Model Synthesis: The prompt, original image, and mask are sent into a frozen diffusion model to synthesize new images that are visually and semantically similar to the original. For noise control, we set the maximum step $T_{max} = 1000$ and **adjust the actual noise step** $T = \epsilon T_{max}$ using a noise strength scalar $\epsilon < 1$, preventing x_T from degenerating entirely into Gaussian noise.

$$x_{sup}^{low} = AddNoise(x_{sup}, \epsilon_{low}) \tag{3}$$

During this process, the diffusion model leverages knowledge pre-trained on the source domain to infer semantic information obscured by noise, effectively restoring key content. Thus, the denoised foreground x_{sup}^{fore} and background x_{sup}^{back} remain category-consistent with the original image x_{sup}. Moreover, since the

Fig. 7: Images generated by foreground and background inpainting, with the bounding box (yellow).

mask is determined by the BBox, the inpainted images retain the original object positions without requiring re-annotation(As shown in Fig. 7).

$$x_{sup}^{back} = M_{diff}(x_{sup}^{low}, \text{prompt}, mask_{sup}) \tag{4}$$

$$x_{sup}^{fore} = M_{diff}(x_{sup}^{low}, \text{prompt} + \text{category}, (1 - mask_{sup})) \tag{5}$$

Finally, the augmented image-label pairs $\{x_{sup}^{back}, y_{sup}\}$ and $\{x_{sup}^{fore}, y_{sup}\}$ are combined with the original support set S to form two extended support sets.

Selection Module However, we also observe that not all generated samples are effective for augmentation, which is possibly caused by the ineffective LLM-generated descriptions or non-jointly trained LLM and diffusion models. Moreover, we also observe that some foreground-inpainted images are also helpful for augmentation, although the ratio of useful samples is much smaller than the proposed background-inpainted ones. Therefore, to further improve the quality of generated samples for CDFSOD, we propose a selection module to further filter out ineffective samples in the extended support sets. Finally, the two extended sets are simultaneously sent into the object detection model.

Specifically, for each candidate image, we compute the box IoU (Intersection over Union) values between the enhanced images x_{sup}^{back} and x_{sup}^{fore}, and then select the positive samples based on IoU values. These samples are then used for subsequent tasks in object detection, including feature extraction, instance reweighting, domain prompting, fine-tuning, and inference.

We employ the Region Proposal Network (RPN) pre-trained on the CD-ViTO model (Baseline) to generate candidate bounding boxes (A) for the extended support set. We calculate the average IoU between predicted boxes and ground truth ones (Y), and only synthesized images with high IoUs are selected.

$$x_{select}^{k} = M_{RPN}^{k} \Sigma(x_{sup}^{back}, x_{sup}^{fore}) \tag{6}$$

We employ the top-k selection algorithm to filter the generated images, with the value of k tailored to each target domain dataset. The specific choices of k are detailed in the supplementary material.

5 Experiments

5.1 Dataset and evaluation setup

We use the benchmark of [10] to conduct experimental evaluations for Cross-Domain Few-Shot Object Detection. The model is trained on the COCO dataset and evaluated on six datasets, namely ArTaxOr, DIOR, UODD, Clipart1k, DeepFish, and NEU-DET.

Table 1: The main results including the state-of-the-art works and works with ours method under 1-shot, 5-shot, 10-shot settings (mAP). † means results are produced by us. Avg. means average result. Values in the table are highlighted as follows: bold for the best and underlined for the second best, respectively.

Setting	Method	Backbone	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Avg.
1-shot	CDMM-FSOD	DETR-R101	15.1	-	14.3	-	6.9	-	/
	CDMM-FSOD + **Ours**†	DETR-R101	16.3	-	15.1	-	7.4	-	/
	ViTDeT-FT	ViT-B/14	5.9	6.1	12.9	0.9	2.4	4.0	5.4
	ViTDeT-FT + **Ours**†	ViT-B/14	5.4	7.2	9.6	5.1	6.3	5.6	6.5
	DE-ViT-FT	ViT-L/14	10.5	13.0	14.7	19.3	0.6	2.4	10.1
	DE-ViT-FT + **Ours**†	ViT-L/14	12.6	14.1	17.1	20.2	2.5	3.2	11.6
	CD-ViTO	ViT-L/14	21.0	17.7	17.8	20.3	3.6	3.1	13.9
	CD-ViTO + **Ours**†	ViT-L/14	26.7	22.3	**19.8**	22.1	5.4	5.6	17.0
	GLIP†	Swin-B	12.6	52.1	4.8	40.1	1.6	4.5	19.3
	GLIP + **Ours**†	Swin-B	38.4	54.7	16.8	41.3	3.2	4.6	26.3
	ETS†	Swin-B	13.7	51.0	7.4	35.0	7.0	11.0	21.9
	ETS + **Ours**†	Swin-B	51.0	52.7	15.6	**43.2**	**13.9**	19.3	32.1
	GroundDINO†	Swin-B	20.0	57.6	8.0	34.3	7.2	17.1	24.0
	DomainRAG	Swin-B	**57.2**	56.1	18.0	38.0	12.1	20.2	33.6
	GroundDINO + **Ours**†	Swin-B	52.3	**59.0**	16.5	41.3	13.6	**22.8**	**34.3**
5-shot	CDMM-FSOD	DETR-R101	48.7	-	26.9	-	12.5	-	/
	CDMM-FSOD + **Ours**†	DETR-R101	55.0	-	29.1	-	17.5	-	/
	ViTDeT-FT	ViT-B/14	20.9	23.3	23.3	9.0	13.5	11.1	16.9
	ViTDeT-FT† + **Ours**	ViT-B/14	18.0	25.1	21.1	18.8	17.1	13.5	19.2
	DE-ViT-FT	ViT-L/14	38.0	38.1	23.4	21.2	7.8	5.0	22.3
	DE-ViT-FT + **Ours**†	ViT-L/14	39.5	41.2	22.9	22.9	8.1	6.1	23.5
	CD-ViTO	ViT-L/14	47.9	41.1	26.9	22.3	11.4	6.8	26.1
	CD-ViTO + **Ours**†	ViT-L/14	52.3	42.8	27.2	22.6	13.4	7.6	27.7
	GLIP†	Swin-B	15.0	57.5	11.8	41.5	9.4	4.4	23.3
	GLIP + **Ours**†	Swin-B	16.9	58.9	13.1	41.4	16.0	4.4	25.1
	ETS†	Swin-B	33.0	58.4	12.8	42.1	9.2	19.3	29.1
	ETS + **Ours**†	Swin-B	63.2	**61.1**	22.2	**51.1**	15.1	27.4	40.0
	GroundDINO†	Swin-B	29.0	60.1	16.0	42.3	12.5	25.5	30.9
	DomainRAG	Swin-B	**70.0**	59.8	**31.5**	43.8	**24.2**	26.8	42.7
	GroundDINO + **Ours**†	Swin-B	65.6	**61.1**	29.8	50.5	23.4	**31.7**	**43.7**
10-shot	CDMM-FSOD	DETR-R101	61.4	-	31.4	-	17.5	-	/
	CDMM-FSOD + **Ours**†	DETR-R101	63.1	-	33.2	-	18.6	-	/
	ViTDeT-FT	ViT-B/14	23.4	25.6	29.4	6.5	15.8	15.6	19.4
	ViTDeT-FT + **Ours**†	ViT-B/14	20.7	27.2	27.2	21.2	20.1	20.1	22.7
	DE-ViT-FT	ViT-L/14	49.2	40.8	25.6	21.3	8.8	5.4	25.2
	DE-ViT-FT + **Ours**†	ViT-L/14	**51.1**	42.2	26.3	22.5	9.6	7.2	26.5
	CD-ViTO	ViT-L/14	60.5	44.3	30.8	22.3	12.8	7.0	29.6
	CD-ViTO + **Ours**†	ViT-L/14	61.4	46.5	31.9	23.6	15.2	9.3	31.3
	GLIP†	Swin-B	26.3	55.3	14.8	36.4	9.3	15.9	26.3
	GLIP + **Ours**†	Swin-B	49.9	60.5	24.2	42.0	8.7	6.5	32.0
	ETS†	Swin-B	63.1	61.2	30.8	40.1	18.9	21.5	39.3
	ETS + **Ours**†	Swin-B	**74.1**	62.7	37.8	47.9	22.5	25.8	45.1
	GroundDINO†	Swin-B	62.8	62.3	24.1	39.3	16.3	25.1	38.1
	DomainRAG	Swin-B	73.4	61.1	**39.0**	41.3	**26.3**	31.2	45.4
	GroundDINO + **Ours**†	Swin-B	69.4	**65.2**	32.6	**53.5**	24.6	**34.8**	**46.9**

5.2 Implementation Details

We evaluate model performance using mean Average Precision (mAP) under 1-shot, 5-shot, and 10-shot settings. For all generated datasets, the experimental pipeline adheres to the three-phase “pretrain, fine-tune, test" protocol. The implementation first initializes with a DE-ViT backbone pre-trained on the COCO dataset, followed by module-specific optimization leveraging the expanded support set. Comprehensive finetuning configurations—hyperparameter settings, optimizer selection, learning rate schedules, training loss, and training epochs—are in the appendix.

5.3 Comparison with state-of-the-art works

Tab. 1 presents a comprehensive comparison with state-of-the-art methods, evaluating three backbone architectures: Swin-B, Vision Transformer-L (ViT-L), Vi-

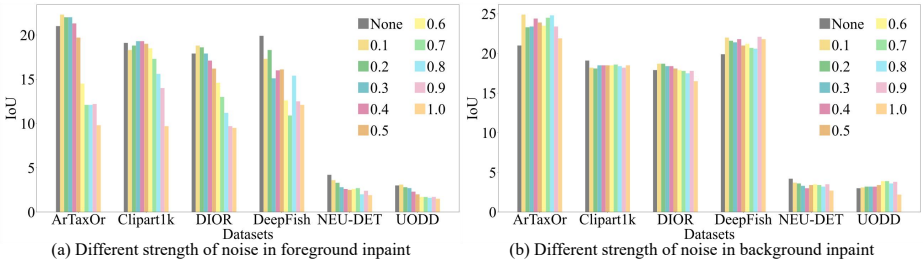

Fig. 8: Ablation study of tailored noise. The legends in different colors indicate different noise strengths.

sion Transformer-B (ViT-B), and DETR-R101, under 1-shot, 5-shot, and 10-shot configurations. The benchmark approaches include CDMM-FSOD [37], ViTDeT-FT [25], DE-ViT-FT [50], CD-ViTO (baseline) [10], GLIP [24], ETS [34], Ground-DINO [28], DomainRAG [26], series. We consistently achieve state-of-the-art average performance in all settings.

5.4 Ablation Study

Generative Module (1) Foreground inpainting: As analyzed in Section 3, within specific noise strength (ϵ) parameter ranges, the Generation Module can effectively synthesize qualified foreground samples. Fig. 8 (a) demonstrates the existence of a definable parameter interval that can produce foreground inpainting samples suitable for the CDFSOD task.

(2) Background inpainting: Fig. 8 (b) shows that even without relying on the Section Module's filtering mechanism, a significant proportion of samples independently generated by the background inpainting could still enhance the performance in the CDFSOD task. This phenomenon indicates that the background inpainting itself possesses intrinsic effectiveness, with its performance improvement not entirely dependent on subsequent selection strategies.

(3) Other diffusion models: Tab. 2 indicates that different versions of the Stable-Diffusion (including SDv2 and SDv3) achieve comparable performance in terms of visual fidelity of the generated images. However, there exists a noticeable quality gap between these versions and other existing diffusion-based methods (including Diffmix [20], IMBABM [15], DA-fusion [42], SDEdit [32], ControlNET [47], and IP-Adapter [45]).

Selection Module The effectiveness of Selection: As shown in Tab. 3, the performance improvement achieved solely by the baseline model without the Selection Module is relatively limited. This is possibly caused by ineffective image descriptions and non-jointly trained LLM and diffusion models, resulting in significant quality variations in the images directly generated by the Generation Module—some samples are difficult for object detection models to identify accurately. Therefore, our proposed Selection Module effectively filters high-quality

Table 2: Comparison with other diffusion models. † means results are produced by us.

Method	ArTa.	Clip.	DIOR	Deep.	NEU.	UODD	Avg.
Diffmix†	11.9	6.7	7.3	0.1	2.4	0.2	4.8
IMBABM†	5.9	3.5	3.9	5.2	0.9	0.9	3.4
Da-fusion†	18.7	18.5	14.7	15.8	2.4	1.9	12.0
SDEdit†	35.5	41.5	13.1	35.7	9.8	13.7	13.2
IP-Adapter†	2.5	15.4	7.1	1.1	3.5	1.0	5.1
ControlNet†	3.1	23.3	7.5	24.5	7.2	3.0	11.4
Ours†	**52.3**	**59.0**	**16.5**	**41.3**	**13.6**	**22.8**	**34.3**

Table 3: Ablation study of Selection Module (mAP). Fore. means that only the foreground is synthesized. Back. means that only the background is synthesized. Sele. means Selection Module.

Dataset	Baseline	Fore.	Back.	Fore. + Back.	Fore. + Back. + Sele.
ArTa.	21.0	22.3	24.9	22.5	**26.7**
Clip.	17.7	19.3	18.6	19.3	**22.3**
DIOR	17.8	18.8	18.7	16.6	**19.8**
Deep.	20.3	18.3	22.1	13.5	**22.1**
NEU.	3.6	3.6	3.7	3.2	**5.4**
UODD	3.1	3.1	3.9	2.1	**5.6**

generated images. These carefully selected samples, combined with the original training data for object detection model finetuning, lead to a substantial overall performance enhancement.

Migrating to CDFSS tasks As shown in Tab. 4, our method has been successfully extended to the cross-domain few-shot segmentation (CDFSS) task, demonstrating competitive performance and validating its generalizability. The benchmark approaches include PerSAM [49], RePRI [3], HSNet [33], PATNet [23], SSP [7], ABCDFSS [17], DRA [40], APSeg [16], FPTrans [46], LoEC [29] series.

Table 4: Migrating SITN to CDFSS tasks

Method	PerSAM	RePRI	HSNet	PATNet	SSP	APM	ABCDFSS	DRA	APSeg	FPTrans	LoEC	FPTrans+Ours
FSS-1000	60.9	71	77.5	78.6	78.9	79.3	74.6	79.1	79.7	80.7	81.1	**84.5**
Deepgloble	36	25	29.7	37.9	40	40.9	42.6	41.3	35.9	38.4	42.1	**43.8**
ISIC	23.3	23.3	31.2	41.2	35.5	41.7	45.7	40.8	45.4	48.6	52.9	**53.4**
Chest X-ray	30	65.1	51.9	66.6	74.4	78.3	79.8	82.4	84.1	80.9	83.9	**84.9**
Average	37.6	46.1	47.6	56.1	57.2	60	60.7	60.9	61.3	62.2	65	**66.7**

Comparison with data augmentations We compare our method with other data augmentations under the 1-shot setting, including Mixup, flipping, and brightness enhancement (Bright+), Simple Copy-Paste [11] (SCP), Entropy Maximization [5] (EM), and Dense outlier detection [1] (Dod). Notably, Tab. 5 shows that most augmentation methods cannot improve the performance against using only the original images (CD-ViTO), verifying the difficulty in augmentation under large domain gaps. The results show that our method achieves significant performance gains compared to other methods. This means that SITN is different from traditional augmentation methods, effectively solving the scarce data problem in CDFSOD.

Table 5: Comparison with existing data augmentation (mAP). Bright+ means Bright Enhance, SCP means Simple Copy-Paste, EM means Entropy Maximization, and Dod means Dense outlier detection.

Method	Mixup	Flip	Bright+	SCP	EME	Dod	Baseline	Ours
ArTa.	13.4	20.5	20.5	24.2	15.6	25.8	21.0	**26.7**
Clip.	12.3	17.7	17.9	17.5	7.8	15.4	17.7	**22.3**
DIOR	13	17.3	18.66	16.7	7.7	16.8	17.8	**19.8**
Deep.	10.9	20.3	17.5	15.5	11	21.8	20.3	**22.1**
NEU.	1.2	4.1	3.1	1.0	0.6	2.0	3.6	**5.4**
UODD	1.5	3.2	2.9	2.3	0.5	3.2	3.1	**5.6**
Avg.	8.7	13.9	13.4	12.9	7.2	14.8	13.9	**17.0**

Object Detection under Occlusion As shown in Fig.9 and Fig. 10, we tested our method on a constructed 1-shot occlusion dataset and achieved promising results. Occlusion remains a general challenge in object detection, and our method offers a generic solution by supplementing data. Moreover, by adding controlled noise during generation, our approach preserves semantic information while improving diversity, making it effective for handling location variation, occlusion, and labeling errors.

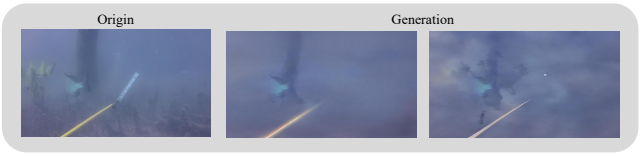

Fig. 9: Application of the our method under occlusion dataset.

Dataset	Deep.
Baseline	35.5
w/o Ours	38.5
with Ours	**41.2**

Fig. 10: Ablation study of occlusion scenes.

Comparision with other open-source diffusion model As shown in Tab. 6, we use Nucleus-Image [41] for generation. Directly applying Nucleus-Image yields poor performance. However, Nucleus-Image with our method performs comparably to SDv1.5 with our method. This shows our method remains effective on stronger diffusion models, confirming its long-term relevance.

Table 6: Comparison with other open-source diffusion models. † means results are produced by us.

Model	ArT.	Clip.	DIOR	Deep.	NEU.	UODD	Avg.
SDv1.5	13.9	11.3	8.6	6.2	2.5	1.7	7.4
Nucleus-Image†	16.8	14.2	5.1	18.8	3.5	9.0	11.2
Ours + SDv1.5	52.3	59	16.5	**41.3**	**13.6**	**22.8**	**34.3**
Ours + Nucleus-Image†	**55.4**	**58.8**	16.3	36.9	8.4	16.8	32.1

Computational Efficiency Due to the diffusion model only applying partial noise to the image, both the forward and backward processes avoid the com-

plete diffusion steps; the image generation is constrained to within 50 steps by leveraging DDIM [39]. Moreover, our method is training-free. Fig. 12 shows the generation time of 1 image between the original diffusion model (DDPM) [18] and our method.

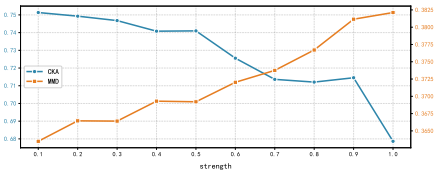

Fig. 11: The variations of CKA and MMD.

Method	Generation Time of 1 image
DDPM	460s
SDEdit	300s-450s
DomainRAG	5s-8s
Ours	0.02s-1.5s

Fig. 12: The computational efficiency of the original diffusion model and our method.

CKA vs. MMD Moreover, the CKA similarity comparison is shown in Fig. 2, where our method achieves much higher CKA scores than other diffusion-based methods, even approaching those of traditional augmentations. The comparison with diffusion-based methods in performance and generated samples is in Fig. 1, where our method consistently achieves much higher performance. We reproduce the CKA experiments by a widely-adopted metric, the MMD distance, where larger MMDs indicate larger domain distances and smaller CKAs. We can see the trend is perfectly symmetric in Fig. 11.

Additional experiments (e.g., comparisons between augmented and real samples, generation quantities) will be provided in the Appendix.

6 Conclusion

We revisit a neglected method for the CDFSOD task: diffusion-based data augmentation, finding that domain gaps make it difficult for current diffusion-based methods to synthesize useful samples for supplementing the scarce training data. We then analyze the domain gap from both the visual and the semantic side, and propose our method, SITN, achieving state-of-the-art results.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grants 62206102; the National Key Research and Development Program of China under grant 2024YFC3307900; the National Natural Science Foundation of China under grants 62436003, 62376103 and 62302184; Major Science and Technology Project of Hubei Province under grant 2025BAB011 and 2024BAA008; Hubei Science and Technology Talent Service Project under grant 2024DJC078; and Ant Group through CCF-Ant Research Fund. The computation is completed in the HPC Platform of Huazhong University of Science and Technology.

References

1. Bevandic, P., Krešo, I., Oršic, M., Šegvic, S.: Dense outlier detection and open-set recognition based on training with noisy negative images. arXiv preprint arXiv:2101.09193 **2**(3) (2021)
2. Bou, X., Facciolo, G., Von Gioi, R.G., Morel, J.M., Ehret, T.: Exploring robust features for few-shot object detection in satellite imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 430–439 (2024)
3. Boudiaf, M., Kervadec, H., Masud, Z.I., Piantanida, P., Ben Ayed, I., Dolz, J.: Few-shot segmentation without meta-learning: A good transductive inference is all you need? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13979–13988 (2021)
4. Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P.A., Li, S.Z.: A survey on generative diffusion models. IEEE Transactions on Knowledge and Data Engineering (2024)
5. Chan, R., Rottmann, M., Gottschalk, H.: Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In: Proceedings of the ieee/cvf international conference on computer vision. pp. 5128–5137 (2021)
6. Davari, M., Horoi, S., Natik, A., Lajoie, G., Wolf, G., Belilovsky, E.: Reliability of cka as a similarity measure in deep learning (Oct 2022)
7. Fan, Q., Pei, W., Tai, Y.W., Tang, C.K.: Self-support few-shot semantic segmentation. In: European conference on computer vision. pp. 701–719. Springer (2022)
8. Fei, B., Lyu, Z., Pan, L., Zhang, J., Yang, W., Luo, T., Zhang, B., Dai, B.: Generative diffusion prior for unified image restoration and enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9935–9946 (2023)
9. Foo, L.G., Rahmani, H., Liu, J.: Ai-generated content (aigc) for various data modalities: A survey. ACM Computing Surveys **57**(9), 1–66 (2025)
10. Fu, Y., Wang, Y., Pan, Y., Huai, L., Qiu, X., Shangguan, Z., Liu, T., Fu, Y., Van Gool, L., Jiang, X.: Cross-domain few-shot object detection via enhanced open-set object detector. In: European Conference on Computer Vision. pp. 247–264. Springer (2024)
11. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2918–2928 (2021)
12. Go, H., Lee, Y., Kim, J.Y., Lee, S., Jeong, M., Lee, H.S., Choi, S.: Towards practical plug-and-play diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1962–1971 (2023)
13. Gui, G., Gao, B.B., Liu, J., Wang, C., Wu, Y.: Few-shot anomaly-driven generation for anomaly classification and segmentation. In: European Conference on Computer Vision. pp. 210–226 (2024)
14. Han, G., Lim, S.N.: Few-shot object detection with foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28608–28618 (2024)
15. He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., Qi, X.: Is synthetic data from generative models ready for image recognition? arXiv preprint arXiv:2210.07574 (2022)

16. He, W., Zhang, Y., Zhuo, W., Shen, L., Yang, J., Deng, S., Sun, L.: Apseg: Auto-prompt network for cross-domain few-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23762–23772 (2024)
17. Herzog, J.: 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. pp. 23605–23615 (2024)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Neural Information Processing Systems,Neural Information Processing Systems (Jan 2020)
19. Huang, J., Wu, Q., Ren, Y., Yang, F., Yang, A., Yang, Q., Pu, X.: Sparse bayesian deep learning for cross domain medical image reconstruction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2339–2347 (2024)
20. Islam, K., Zaheer, M.Z., Mahmood, A., Nandakumar, K.: Diffusemix: Label-preserving data augmentation with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27621–27630 (2024)
21. Jabbour, S., Kondas, G., Kazerooni, E., Sjoding, M., Fouhey, D., Wiens, J.: Depict: Diffusion-enabled permutation importance for image classification tasks. In: European Conference on Computer Vision. pp. 35–51. Springer (2025)
22. Jiang, Y., Zou, Y., Li, Y., Li, R.: Remedying target-domain astigmatism for cross-domain few-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19580–19590 (2026)
23. Lei, S., Zhang, X., He, J., Chen, F., Du, B., Lu, C.T.: Cross-domain few-shot semantic segmentation. In: European conference on computer vision. pp. 73–90. Springer (2022)
24. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10965–10975 (2022)
25. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: European conference on computer vision. pp. 280–296. Springer (2022)
26. Li, Y., Qiu, X., Fu, Y., Chen, J., Qian, T., Zheng, X., Paudel, D.P., Fu, Y., Huang, X., Van Gool, L., et al.: Domain-rag: Retrieval-guided compositional image generation for cross-domain few-shot object detection. arXiv preprint arXiv:2506.05872 (2025)
27. Liu, J., Wang, Q., Fan, H., Wang, Y., Tang, Y., Qu, L.: Residual denoising diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2773–2783 (2024)
28. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
29. Liu, Y., Zou, Y., Li, Y., Li, R.: The devil is in low-level features for cross-domain few-shot segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4618–4627 (2025)
30. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11461–11471 (2022)
31. Luo, C.: Understanding diffusion models: A unified perspective. arXiv preprint arXiv:2208.11970 (2022)

32. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
33. Min, J., Kang, D., Cho, M.: Hypercorrelation squeeze for few-shot segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6941–6952 (2021)
34. Pan, J., Liu, Y., He, X., Peng, L., Li, J., Sun, Y., Huang, X.: 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. pp. 1548–1556 (2025)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
36. Sain, A., Bhunia, A.K., Chowdhury, P.N., Koley, S., Xiang, T., Song, Y.Z.: Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2765–2775 (2023)
37. Shangguan, Z., Seita, D., Rostami, M.: Cross-domain multi-modal few-shot object detection via rich text. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6570–6580. IEEE (2025)
38. Shen, J., Kuang, K., Wang, J., Wang, X., Feng, T., Zhang, W.: Cgmgm: A cross-gaussian mixture generative model for few-shot semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4784–4792 (2024)
39. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv: Learning,arXiv: Learning (Oct 2020)
40. Su, J., Fan, Q., Pei, W., Lu, G., Chen, F.: Domain-rectifying adapter for cross-domain few-shot segmentation. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 24036–24045 (2024)
41. Team, N.A.: Nucleus-image: Sparse moe for image generation (2026)
42. Trabucco, B., Doherty, K., Gurinas, M., Salakhutdinov, R.: Effective data augmentation with diffusion models. arXiv preprint arXiv:2302.07944 (2023)
43. Xiong, W.: 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). pp. 1–5. IEEE (2023)
44. Yan, Q., Chen, G., Zou, Y.: Start small, think big: Curriculum-based relative policy optimization for visual grounding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 11550–11558 (2026)
45. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
46. Zhang, J.W., Sun, Y., Yang, Y., Chen, W.: Feature-proxy transformer for few-shot segmentation. Advances in neural information processing systems **35**, 6575–6588 (2022)
47. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
48. Zhang, R., Hu, X., Li, B., Huang, S., Deng, H., Qiao, Y., Gao, P., Li, H.: Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15211–15222 (2023)

49. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot. arXiv preprint arXiv:2305.03048 (2023)
50. Zhang, X., Liu, Y., Wang, Y., Boularias, A.: Detect everything with few examples. arXiv preprint arXiv:2309.12969 (2023)
51. Zhao, Y., Zou, Y., Li, Y., Li, R.: Interpretable cross-domain few-shot learning with rectified target-domain local alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 41605–41615 (June 2026)
52. Zou, Y., Liu, Y., Hu, Y., Li, Y., Li, R.: Flatten long-range loss landscapes for cross-domain few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23575–23584 (2024)
53. Zou, Y., Ma, R., Li, Y., Li, R.: Attention temperature matters in vit-based cross-domain few-shot learning. Advances in Neural Information Processing Systems **37**, 116332–116354 (2024)
54. Zou, Y., Zhang, S., Zhou, H., Li, Y., Li, R.: Compositional few-shot class-incremental learning. arXiv preprint arXiv:2405.17022 (2024)