Domain-RAG:
Retrieval-Guided Compositional Image Generation for Cross-Domain Few-Shot Object Detection

Yu Li[1*] **Xingyu Qiu**[1*] **Yuqian Fu**[2*†] **Jie Chen**[3] **Tianwen Qian**[4] **Xu Zheng**[2,5]
Danda Pani Paudel[2] **Yanwei Fu**[1] **Xuanjing Huang**[1] **Luc Van Gool**[2] **Yu-Gang Jiang**[1]
[1]Fudan University [2]INSAIT, Sofia University "St. Kliment Ohridski"
[3]Fuzhou University [4]East China Normal University [5]HKUST(GZ)

Abstract

Cross-Domain Few-Shot Object Detection (CD-FSOD) aims to detect novel objects with only a handful of labeled samples from previously unseen domains. While data augmentation and generative methods have shown promise in few-shot learning, their effectiveness for CD-FSOD remains unclear due to the need for both visual realism and domain alignment. Existing strategies, such as copy-paste augmentation and text-to-image generation, often fail to preserve the correct object category or produce backgrounds coherent with the target domain, making them non-trivial to apply directly to CD-FSOD. To address these challenges, we propose **Domain-RAG**, a training-free, retrieval-guided compositional image generation framework tailored for CD-FSOD. Domain-RAG consists of three stages: domain-aware background retrieval, domain-guided background generation, and foreground-background composition. Specifically, the input image is first decomposed into foreground and background regions. We then retrieve semantically and stylistically similar images to guide a generative model in synthesizing a new background, conditioned on both the original and retrieved contexts. Finally, the preserved foreground is composed with the newly generated domain-aligned background to form the generated image. Without requiring any additional supervision or training, Domain-RAG produces high-quality, domain-consistent samples across diverse tasks, including CD-FSOD, remote sensing FSOD, and camouflaged FSOD. Extensive experiments show consistent improvements over strong baselines and establish new state-of-the-art results. The source code and instructions are available at https://github.com/LiYu0524/Domain-RAG.

1 Introduction

Cross-Domain Few-Shot Object Detection (CD-FSOD) [10], an emerging task derived from cross-domain few-shot learning (CD-FSL) [13], aims to tackle few-shot object detection (FSOD) across different domains. Unlike conventional FSOD [20], which assumes source and target data share similar distributions, CD-FSOD considers more realistic scenarios with significant domain shifts, for example, transferring from natural images to industrial anomaly images, remote sensing imagery, or underwater environments. By simultaneously involving the challenges of few-shot learning and domain shift, CD-FSOD poses significant challenges for existing detectors.

[*]These authors have equal contributions.
[†]Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

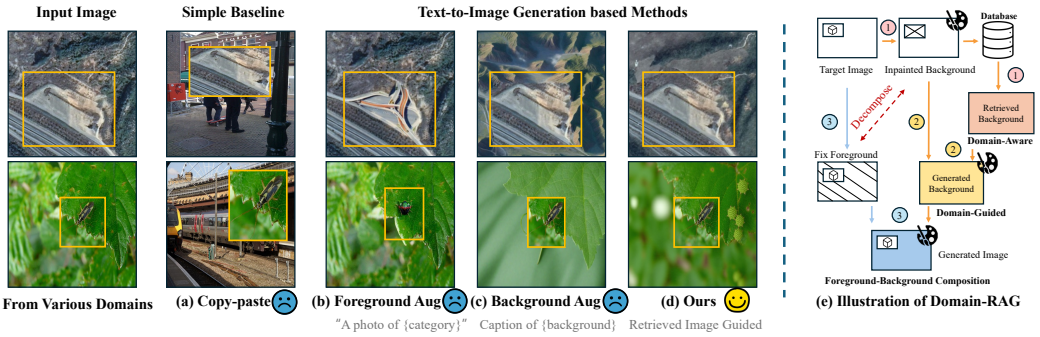


Figure 1: Given images from distinct novel domains, we compare generation results of baseline methods (a–c) and our approach (d), and illustrate the main pipeline of our Domain-RAG (e).

Due to the extreme scarcity of labeled data, e.g., as few as 1 or 5 annotated samples per category, a natural and intuitive solution is to leverage data augmentation to alleviate the data bottleneck. Although image augmentation and generation techniques have been extensively studied and shown effective in other few-shot learning tasks [61, 27, 29, 30], it remains unclear whether they can produce high-quality training samples for CD-FSOD. Different from the few-shot classification or in-domain FSOD, this setting requires not only accurate object annotation but also strong domain consistency, which has not been tackled in prior data augmentation-based few-shot learning methods.

To generate training images for CD-FSOD, the most straightforward approach is copy-paste (Fig. 1(a)). While easy to implement, such images often lack realism and domain coherence. A more advanced strategy is to build on recent generative models, particularly the trending text-to-image generation, such as SDXL [43], FLUX [21]. Most existing methods in this area focus on synthesizing foreground objects via text prompts, such as "a photo of category" (Fig. 1(b)). However, they might fail to preserve the object semantics when applied to novel categories and domains. Such a category shift is problematic for CD-FSOD, which has to tackle fine-grained objects and the domain gap. Other approaches generate diverse backgrounds (Fig. 1(c)), guided by text descriptions of the image. While this better preserves the foreground, purely textual descriptions often fall short of capturing precise domain characteristics and struggle to ensure semantic and visual consistency between foreground and background. These limitations motivate us to develop a new image generation framework capable of synthesizing visually coherent, domain-aligned training samples for CD-FSOD. Specifically, we aim to: ① preserve the original foreground object, ② generate diverse backgrounds that are both semantically and stylistically aligned with the query image and its domain, and ③ produce visually realistic images with valid annotations suitable for downstream detection training.

To that end, we propose **Domain-RAG**, a *retrieval-guided compositional image generation framework* built upon the principle of *fix the foreground, adapt the background*. Leveraging the nature of object detection, Domain-RAG begins by decomposing the target image into its foreground object and background, where the background is recovered by applying an inpainting model [49] to the object-masked region. Although simple in principle, this step is critical for preserving the original object and its annotations, laying the foundation for controllable compositional generation. The core challenge then lies in generating a new background that is semantically and stylistically compatible with the foreground. Inspired by the paradigm of retrieval-augmented generation (RAG) [24], we inject structured visual priors into the generative process to guide background synthesis. As illustrated in Fig. 1(e), Domain-RAG consists of the following three stages: **1) Domain-Aware Background Retrieval.** We introduce an image database (e.g., COCO [31]) containing diverse natural scenes, from which we retrieve candidate backgrounds that are semantically and stylistically similar to the inpainted background of the target image. Semantic similarity is computed using high-level visual features, while style similarity is measured via style-based descriptors [16]. **2) Domain-Guided Background Generation.** Rather than using the retrieved backgrounds directly, we feed them along with the target's inpainted background into a generative model to synthesize a new background that better reflects the visual characteristics of the target domain. To ensure compatibility with modern diffusion models, Redux [23] is applied to convert visual image cues into descriptive text prompts, enabling direct use of text-to-image generation models. **3) Foreground-Background Composition.** Finally, the preserved foreground is seamlessly composed onto the synthesized, domain-aligned background using a mask-guided generative model. The resulting image maintains the original object while embedding it in a realistic, domain-consistent context (Fig. 1(d)). The entire Domain-RAG pipeline is

training-free and can be directly integrated with existing detectors without any additional supervision or retraining, making it particularly suitable for low-shot scenarios such as 1-shot CD-FSOD.

We validate Domain-RAG on three various tasks that address few-shot object detection with domain shifts: CD-FSOD, remote sensing FSOD (RS-FSOD), and camouflaged FSOD. In all tasks, our method consistently improves a strong baseline by an average of +7.3, +1.1, and +2.1 mAP under the lowest-shot setting, achieving new state-of-the-art (SOTA) performance. These results demonstrate its broad applicability and effectiveness across diverse domains.

Our main contributions are as follows: 1) We propose Domain-RAG, a training-free, model-agnostic, retrieval-guided compositional image generation framework for boosting cross-domain few-shot object detection. 2) Domain-RAG enables image generation that preserves the original foreground while synthesizing domain-aligned backgrounds, guided by semantically and stylistically similar retrieved examples. 3) We achieve consistent performance improvements and new state-of-the-art results across a broad range of CD-FSOD, remote sensing FSOD, and camouflaged FSOD tasks.

2 Related Works

Cross-Domain Few-Shot Tasks. Few-shot learning across domains has been widely studied [13, 50, 8, 29, 63, 54, 11, 64], but most works focus only on classification. The more realistic task of cross-domain few-shot object detection (CD-FSOD) [10, 9], which involves both recognizing and localizing objects, remains underexplored. Recent methods like CD-ViTO [10] and ETS [42] address CD-FSOD. CD-ViTO introduces the task with a closed-source setting (COCO as the only source), while ETS uses a more practical open-source setting [9] and leverages data augmentation via pretrained GroundingDINO [33]. In this paper, we adopt the open-source setting and further improve augmentation using retrieval-guided compositional generation.

FSOD Beyond Domains. Beyond classic CD-FSOD tasks, many FSOD or detection problems also involve domain shifts, even if not explicitly labeled as cross-domain. Two notable examples are Remote Sensing FSOD (RS-FSOD) [34] and Camouflaged FSOD [39]. RS-FSOD uses remote sensing images, which differ from natural scenes in color, perspective, and resolution, creating a clear domain gap. Camouflaged FSOD involves detecting objects that blend into their backgrounds—like fish underwater or animals in the wild—posing challenges for generalization. We include both tasks to assess our method under diverse and difficult cross-domain scenarios.

Data Augmentation. Data augmentation is a key technique for the vision community. Traditional methods for object detection, like copy-paste [12], cropping, and color jittering [2], are simple but offer limited semantic variety. Recently, generative models—especially text-to-image models like ControlNet [55], SDXL [43], FLUX [21], and FLUX-Fill [22] have enabled more advanced augmentations. Methods such as X-Paste [59], Lin et al. [30], and Zhang et al. [57] generate new foregrounds to paste on diverse backgrounds, while others [40, 56, 55, 3] use text prompts to jointly create foregrounds and backgrounds. However, these methods typically rely on large amounts of in-domain data for training, which limits their adaptability to novel categories or unseen domains. In contrast, our Domain-RAG is training-free and leverages retrieved real-world images as visual priors to generate domain-consistent samples, making it well-suited for CD-FSOD.

Retrieval-Augmented Generation in Vision. First introduced in NLP [24], retrieval-augmented generation (RAG) enhances outputs by incorporating relevant retrieved content as external knowledge. Its strong performance has led to applications in vision tasks such as image captioning [45, 25], visual question answering [32, 15], and image generation [1, 38, 60], and pose estimation. However, current RAG-based image generation methods are aimed at open-ended synthesis and are not suited for object detection, particularly in cross-domain few-shot settings, where both domain alignment and object fidelity are crucial. To the best of our knowledge, we are the first to introduce a RAG-inspired, training-free image generation framework specifically designed for CD-FSOD.

3 Proposed Method

Problem Setup. The CD-FSOD task aims to adapt an object detector from a source domain $\mathcal{D}_S$ to a target domain $\mathcal{D}_T$, where the data distributions $\mathcal{P}_S$ and $\mathcal{P}_T$ differ. We use the few-shot setting, i.e., N-way K-shot protocol to evaluate detection results in $\mathcal{D}_T$. Specifically, a support set $\mathcal{S}^{N\times K} \subset \mathcal{D}_T$

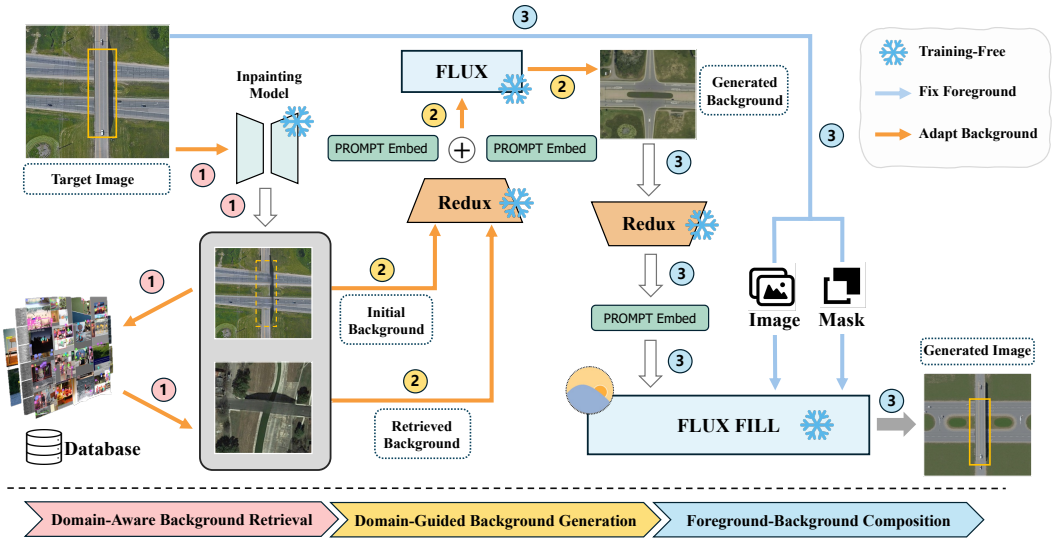


Figure 2: **Illustration of our Domain-RAG.** Built on our principle of "fix the foreground, adapt the background", we first decompose image and process it with three key modules: domain-aware background retrieval, domain-guided background generation, and foreground-background composition.

provides K labeled examples per novel class, and a query set $\mathcal{Q}$ is used for evaluation. We use the open-source setting introduced in the 1st CD-FSOD Challenge [9], which allows foundation models pretrained on large-scale data, to explore the potential of foundation models in CD-FSOD. Particularly, instead of pretraining on $\mathcal{D}_S$, we directly finetune a pretrained detector (e.g., GroundingDINO [33]) on the support set $\mathcal{S}$ and evaluate the results on the query set $\mathcal{Q}$. To mitigate the limited size of $\mathcal{S}$, we augment each support instance with n synthetic images, effectively expanding each class from K to $K \times (n+1)$ examples.

Overview. We propose Domain-RAG—a novel, training-free, retrieval-guided compositional generation framework that enhances support diversity by generating domain-aligned samples. To enable retrieval, we use COCO [31] as the database $\mathcal{D}_{base}$, serving as a gallery of candidate backgrounds. Following the core principle of "*fix the foreground, adapt the background*", Domain-RAG processes each support image $x \in \mathcal{S}$ by first decomposing it into foreground object(s) and background. As shown in Fig. 2, the framework then proceeds through three key stages: *1) domain-aware background retrieval* first obtains the inpainted background b_{init} from x and then retrieves G candidate backgrounds b_{re} from $\mathcal{D}_{base}$ that are semantically and stylistically similar. *2) domain-guided background generation* feeds each $\{b_{init}, b_{re}\}$ pair into a generative model to synthesize a new domain-aligned background b_{dom}. *3) foreground-background composition* finally produces n new images x^+ by compositing the preserved foreground onto each b_{dom} using a mask-guided generative model.

3.1 Domain-Aware Background Retrieval

We propose a two-stage retrieval strategy that combines CLIP's high-level semantic features with ResNet's low-level style descriptors to search an existing image database. The method retrieves images whose semantics and appearance are most similar to the target domain, providing background candidates that better match the target-domain distribution and thus enrich the support set S.

In practice, given a support image x, we remove the ground-truth bounding box with LaMa inpainting [49] to obtain a background without foreground b_{init}. We use the CLIP encoder to extract embeddings from the initial background b_{init} and the database $\mathcal{D}_{base}$, which we refer to as F_{bg} and F_{base}, respectively. We compute cosine similarity between the visual feature of the current background query F_{bg} and the CLIP embedding of each sample in the database. The top m most similar images are selected based on this similarity ranking, forming the candidate set $B_{<clip,m>}$, which contains m images of the form b_{clip}. The subscript notation indicates that the set is constructed using CLIP vision encoder and contains m elements.

Building on this step, we re-rank the b_{clip} by extracting low-level style descriptors using shallow-layer ResNet features. For each background image b_{clip} retrieved by CLIP, we extract its low-level feature

map F using the early layers of a ResNet encoder. We further compute the per-channel mean μ_c and standard deviation σ_c by averaging over the spatial dimensions of the feature map F. Concatenating the means and standard deviations over all channels yields a 128-D style vector as,

$$\mathbf{s}(b_{clip}) = [\mu_1, \ldots, \mu_C, \sigma_1, \ldots, \sigma_C] \in \mathbb{R}^{2C}. \tag{1}$$

For each retrieval candidate b_{clip} in the set $B_{<clip,m>}$, we compute the style distance as the L2 norm between the style features of the original image b_{init} and the candidate:

$$d = \|\mathbf{s}(b_{init}) - \mathbf{s}(b_{clip})\|_2 . \tag{2}$$

Here, each d corresponds to a candidate in $B_{<clip,\ m>}$, and we use the distance to rank and select the most stylistically similar backgrounds. We then re-rank the m CLIP-retrieved candidates based on their style distances and retain the top n images that are most similar in style. The resulting set of selected images is denoted as $B_{<re,\ n>}$. The images indexed by $B_{<re,\ n>}$ serve as style-matched references for the subsequent background generation stage.

3.2 Domain-Guided Background Generation

To fully leverage the retrieved images while keeping the generation process training-free, we adopt the Flux-Redux model[23] to encode each image into a prompt embedding. Given our domain-aware retrieval results $B_{<re,\ n>}$, let redux$(\cdot)$ denote FLUX-Redux encoder and FLUX$(\cdot)$ denote the FLUX generator. For each support image, we extract its clean background embedding $F_{bg} = \text{redux}(b_{init})$ and the embedding of the top retrieved image $b_{re} \in B_{<re,\ n>}$ as $F_{re} = \text{redux}(b_{re})$. We then fuse them as $F_{dom} = \lambda_1 F_{bg} + \lambda_2 F_{re}$, where λ_1 and λ_2 are hyper parameters.

Finally, the FLUX generator produces diverse background images at 1024×1024 resolution by applying a generative function FLUX to the domain embedding F_{dom} i.e., $b_{dom} = \text{FLUX}(F_{dom})$. We sample this process n times to generate a set of diverse images $\{b^{(1)}, b^{(2)}, \ldots, b^{(n)}\}$.

3.3 Foreground-Background Composition

Based on the diverse backgrounds generated in the previous stage, we aim to seamlessly integrate new backgrounds into the original images while preserving foreground pixels and maintaining the target-domain distribution. To achieve this, we employ Flux-Fill for outpainting. Specifically, for each corresponding support image x, we construct a binary mask $M \in \{0,1\}^{H \times W}$. The mask is computed based on the ground-truth bounding box bbox(x) as:

$$M(p) = \begin{cases} 0, & \text{if } p \in \text{bbox}(x), \\ 1, & \text{otherwise.} \end{cases} \quad \text{for each } p \in \Omega_x, \tag{3}$$

where Ω_x denotes the spatial domain of image x, and p indexes a pixel location. We then extract the prompt embedding F_{gen} by $F_{gen} = \texttt{redux}(b_{dom})$ and feed $\{x, M, F_{gen}\}$ into Flux-Fill. To preserve foreground details, Flux-Fill encodes the input x using a VAE and blends the encoded latent features with the initial noise. However, due to the VAE-based downsampling, it struggles to retain fine-grained structures such as small objects. To mitigate this issue, before generation, we denote the up-sampling method s_{up} on each image as,

$$s_{up}(x) = \begin{cases} 0, & \text{if width}(x) > 1024 \text{ and height}(x) > 1024, \\ 1, & \text{otherwise.} \end{cases} \tag{4}$$

After generation, we denote a corresponding down-sampling method s_{down} as,

$$s_{down}(x) = \begin{cases} 0, & \text{if } s_{up}(x) = 0, \\ 1, & \text{otherwise.} \end{cases} \tag{5}$$

The model then repaints only the masked regions, merging the style of b_{dom} while keeping the foreground object's appearance and position unchanged. The final output of Domain-RAG, denoted x^+ , is given by,

$$x^+ = s_{down}\left(\text{Flux-Fill}\left(s_{up}(x),\ s_{up}(M),\ F_{gen}\right)\right). \tag{6}$$

This completes the foreground-background composition, yielding an augmented support image with a domain-aligned background and unchanged foreground objects.

3.4 Applying Domain-RAG to CD-FSOD

In principle, our proposed **Domain-RAG** framework can be seamlessly integrated with any existing detector to enhance its performance in cross-domain scenarios. As a training-free, plug-and-play data augmentation module, Domain-RAG requires no modification to the detection architecture or training pipeline. Once the augmented support images are generated, the model is fine-tuned on the combination of the original support set $\mathcal{S}$ and the generated samples. At inference time, Domain-RAG is not involved; the detector is evaluated directly on the original query set $\mathcal{Q}$.

4 Experiments

Setups. We conduct experiments on three FSOD tasks with domain shifts: **1) CD-FSOD:** Following the CD-ViTO benchmark [10], we evaluate on six diverse target domains: ArTaxOr [6] (photorealistic), Clipart1k [17] (cartoon), DIOR [26] (aerial), DeepFish [46] (underwater), NEU-DET [47] (industrial), and UODD [18] (underwater). **2) Remote Sensing FSOD (RS-FSOD):** In addition to DIOR, we include NWPU VHR-10 [41], a popular remote sensing dataset for FSOD. **3) Camouflaged FSOD:** We also test on CAMO-FS [39], a recent dataset with 47 categories where objects are deliberately camouflaged into the background. For each task, we follow the standard dataset splits and evaluation protocols: 1/5/10-shot for CD-FSOD, 3/5/10/20-shot for RS-FSOD, and 1/2/3/5-shot for Camouflaged FSOD. Results are reported using mean Average Precision (mAP).

Table 1: **Main results (mAP) on the CD-FSOD benchmark** under the 1/5/10-shot setting. † marks methods implemented or reproduced by us. Best results are highlighted in pink.

	Method	Backbone	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Average
1-shot	Meta-RCNN [53]	ResNet50	2.8	-	7.8	-	-	3.6	/
	TFA w/cos [51]	ResNet50	3.1	-	8.0	-	-	4.4	/
	FSCE [48]	ResNet50	3.7	-	8.6	-	-	3.9	/
	DeFRCN [44]	ResNet50	3.6	-	9.3	-	-	4.5	/
	Distill-cdfsod [52]	ResNet50	5.1	7.6	10.5	NaN	NaN	5.9	/
	ViTDeT-FT [28]	ViT-B/14	5.9	6.1	12.9	0.9	2.4	4.0	5.4
	Detic [62]	ViT-L/14	0.6	11.4	0.1	0.9	0.0	0.0	2.2
	Detic-FT [62]	ViT-L/14	3.2	15.1	4.1	9.0	3.8	4.2	6.6
	DE-ViT [58]	ViT-L/14	0.4	0.5	2.7	0.4	0.4	1.5	1.0
	CD-ViTO [10]	ViT-L/14	21.0	17.7	17.8	20.3	3.6	3.1	13.9
	GroundingDINO† [33]	Swin-B	26.3	55.3	14.8	36.4	9.3	15.9	26.3
	ETS† [42]	Swin-B	28.1	55.8	12.7	**39.3**	11.7	18.9	27.8
	Domain-RAG (Ours)	Swin-B	**57.2**	**56.1**	**18.0**	38.0	**12.1**	**20.2**	**33.6**
5-shot	Meta-RCNN [53]	ResNet50	8.5	-	17.7	-	-	8.8	/
	TFA w/cos [51]	ResNet50	8.8	-	18.1	-	-	8.7	/
	FSCE [48]	ResNet50	10.2	-	18.7	-	-	9.6	/
	DeFRCN [44]	ResNet50	9.9	-	18.9	-	-	9.9	/
	Distill-cdfsod [52]	ResNet50	12.5	23.3	19.1	15.5	16.0	12.2	16.4
	ViTDeT-FT [28]	ViT-B/14	20.9	23.3	23.3	9.0	13.5	11.1	16.9
	Detic [62]	ViT-L/14	0.6	11.4	0.1	0.9	0.0	0.0	2.2
	Detic-FT [62]	ViT-L/14	8.7	20.2	12.1	14.3	14.1	10.4	13.3
	DE-ViT [58]	ViT-L/14	10.1	5.5	7.8	2.5	1.5	3.1	5.1
	CD-ViTO [10]	ViT-L/14	47.9	41.1	26.9	22.3	11.4	6.8	26.1
	GroundingDINO† [33]	Swin-B	68.4	57.6	29.6	41.6	19.7	25.6	40.4
	ETS† [42]	Swin-B	64.5	59.7	29.3	42.1	23.5	**27.7**	41.1
	Domain-RAG (Ours)	Swin-B	**70.0**	**59.8**	**31.5**	**43.8**	**24.2**	26.8	**42.7**
10-shot	Meta-RCNN [53]	ResNet50	14.0	-	20.6	-	-	11.2	/
	TFA w/cos [51]	ResNet50	14.8	-	20.5	-	-	11.8	/
	FSCE [48]	ResNet50	15.9	-	21.9	-	-	12.0	/
	DeFRCN [44]	ResNet50	15.5	-	22.9	-	-	12.1	/
	Distill-cdfsod [52]	ResNet50	18.1	27.3	26.5	15.5	21.1	14.5	20.5
	ViTDeT-FT [28]	ViT-B/14	23.4	25.6	29.4	6.5	15.8	15.6	19.4
	Detic [62]	ViT-L/14	0.6	11.4	0.1	0.9	0.0	0.0	2.2
	Detic-FT [62]	ViT-L/14	12.0	22.3	15.4	17.9	16.8	14.4	16.5
	DE-ViT [58]	ViT-L/14	9.2	11.0	8.4	2.1	1.8	3.1	5.9
	CD-ViTO [10]	ViT-L/14	60.5	44.3	30.8	22.3	12.8	7.0	29.6
	GroundingDINO† [33]	Swin-B	73.0	58.6	37.2	38.5	25.5	30.3	43.9
	ETS† [42]	Swin-B	70.6	60.8	37.5	**42.8**	26.1	28.3	44.4
	Domain-RAG (Ours)	Swin-B	**73.4**	**61.1**	**39.0**	41.3	**26.3**	**31.2**	**45.4**

Implementation Details. We use pretrained GroundingDINO [33] with Swin-Transformer [35] Base (Swin-B) as backbone as our baseline. For the retrieval stage, the hyper parameters are set to $m = 100$, $n = 5$, throughout all experiments. For the background generation stage, fusion hyper parameters λ_1 and λ_2 are set to 1.0 and 0.8 respectively. We fine-tune the model for 30 epochs by default, but reduce to 5 for faster-converging datasets like Clipart1k and DeepFish. We use AdamW [36] with learning rate and weight decay set to 1×10^{-4}, and we scale the backbone's learning rate by 0.1. All experiments are run on four Tesla V100 GPUs or eight 5880 Ada GPUs, or a single A800 GPU. Further details are in the Appendix.

4.1 Main Comparison Results

CD-FSOD Results. Tab. 1 summarizes the main comparison results on CD-FSOD under 1/5/10 shots across six novel targets. Particularly, we include several competitors: Meta-RCNN [53], TFA w/cos [51], FSCE [48], DeFRCN [44], Distill-cdfsod [52], ViTDeT-FT [28], Detic/Detic-FT [62], DE-ViT [58] as reported in CD-ViTO [10]. In addition, we also report the results of fine-tuned GroundingDINO [33], ETS [42] to compare with our Domain-RAG. Note that both ETS and Domain-RAG build on GroundingDINO but with different augmentation strategies.

We highlight that Domain-RAG consistently outperforms existing competitors across most target domains, achieving new state-of-the-art (SOTA) results. Compared to the GroundingDINO baseline, our method improves mAP by 7.3, 2.3, and 1.5 points under the 1, 5, and 10 shots, respectively. These results not only show the effectiveness of Domain-RAG, but also reveal its superiority over other proposed augmentation strategies such as ETS. Beyond the average gains, we also notice: 1) *Significant gains on ArTaxOr.* Domain-RAG achieves a 117.5% relative improvement in the 1-shot setting. We attribute this to the strong semantic and visual compatibility between ArTaxOr and the retrieved COCO-style backgrounds, where ArTaxOr features the fine-grained foreground but with a relatively close visual domain to COCO regarding background. 2) *Robustness under low-shot settings.* The largest gains are observed in the 1-shot scenario, which is the most challenging FSOD scenario. This shows our benefits under severe data scarcity. 3) *Strong generalization to severe domain shift.* On NEU-DET, an industrial defect detection dataset characterized by uncommon objects and background styles, Domain-RAG consistently improves all shot settings, demonstrating its capability to handle the most challenging cross-domain FSOD cases.

RS-FSOD Results. Tab. 2 summarizes the results on the NWPU VHR-10 remote sensing dataset [41] under the 3/5/10/20-shot settings. The dataset is divided into 7 base classes and 3 novel classes. The table is split into two parts. In the *upper part*, we follow the standard RS-FSOD protocol: models are first trained on the base classes and then fine-tuned and evaluated on the novel classes. Under this setting, the base classes contain a sufficient number of annotated samples, and we apply our augmentation strategy on top of the previous state-of-the-art method SEA-FSDet [34], and report the mean Average Precision (mAP) over the 3 novel classes. In the *lower part*, we explore the dataset in a CD-FSOD setting, where the pretrained model is directly fine-tuned on all 10 classes (both base and novel), each with only a few labeled samples. To ensure comparability with the upper setting, the reported mAP here reflects performance exclusively on the three novel categories.

Table 2: **Main results (mAP) on the NWPU VHR-10 benchmark** under the 3/5/10/20-shot settings. The upper part follows the standard **RS-FSOD** problem setup, while the lower part adapts **CD-FSOD** setup, with the best results highlighted in pink. † means results are produced by us.

Method	Training Setting	Backbone	3-shot	5-shot	10-shot	20-shot	Average
Meta-RCNN [53]	RS-FSOD	ResNet-50	20.51	21.77	26.98	28.24	24.38
FsDetView [19]	RS-FSOD	ResNet-50	24.56	29.55	31.77	32.73	29.65
TFA w/cos [51]	RS-FSOD	ResNet-50	16.17	20.49	21.22	21.57	19.86
P-CNN [4]	RS-FSOD	ResNet-50	41.80	49.17	63.29	66.83	55.27
FSOD [7]	RS-FSOD	ResNet-50	10.95	15.13	16.23	17.11	14.86
FSCE [48]	RS-FSOD	ResNet-50	41.63	48.80	59.97	79.60	57.50
ICPE [37]	RS-FSOD	ResNet-50	6.10	9.10	12.00	12.20	9.85
VFA [14]	RS-FSOD	ResNet-50	13.14	15.08	13.89	20.18	15.57
SAE-FSDet [34]	RS-FSOD	ResNet-50	57.96	59.40	71.02	**85.08**	68.36
Domain-RAG (Ours)	RS-FSOD	ResNet-50	**59.99**	**65.78**	**72.87**	84.05	**70.67**
GroundingDINO† [33]	CD-FSOD	Swin-B	57.1	61.3	65.1	69.5	63.3
Domain-RAG (Ours)	CD-FSOD	Swin-B	**58.2**	**62.1**	**66.6**	**69.7**	**64.2**

Notably, from the upper standard RS-FSOD results, we observe the following findings: 1) Our Domain-RAG achieves the best result via improving the strong SEA-FSDet, achieving 2.31 mAP improvement across all shots on average. This indicates that our plug-and-play augmented method is compatible with existing methods. 2) Minor decrease is observed for 20-shot, from 85.08 to 84.05. We speculate that it is due to the base training being sufficient. Further augmentation in this regime may lead to overfitting on synthetic data patterns rather than benefiting novel-class generalization. From the lower part of the CD-FSOD setting results, we highlight that our method again improves the strong GroundingDINO baseline, indicating its effectiveness.

Camouflaged FSOD Results. Tab. 3 presents the results on the CAMO-FS [39] under 1/2/3/5 shots. All categories in this dataset are treated as novel classes and are further split into a support set and a query set, naturally aligning with the formulation of CD-FSOD. The first two rows in the table report the results of "FS-CDIS-ITL" and "FS-CDIS-IMS", two methods developed from the original CAMO-FS paper. Below that, we include our reproduced baseline using GroundingDINO as the detector, along with the results of our proposed Domain-RAG method built on top of GroundingDINO.

As shown by the results, the large-scale pretrained model, i.e., GroundingDINO, brings a substantial performance boost to this task, improving results from around 7 to over 65 mAP.

We believe this remarkable advancement will advance the frontier of this field. Moreover, the performance gains introduced by our proposed method over the GroundingDINO baseline remain consistently clear across all shot settings. The consistent success across CD-FSOD, RS-FSOD, and camouflaged FSOD, covering eight challenging and diverse domains, demonstrates that our method serves as a general and effective solution for addressing the gap issue in few-shot object detection.

Table 3: **Main results (mAP) on the Camouflage FSOD** under the 1/2/3/5-shot settings. † means the results are produced by us, the best results are highlighted in pink.

Method	Backbone	1-shot	2-shot	3-shot	5-shot	Average
FS-CDIS-ITL [39]	ResNet-101	4.0	7.3	7.5	9.8	7.1
FS-CDIS-IMS [39]	ResNet-101	4.5	7.0	7.6	10.4	7.4
GroundingDINO† [33]	Swin-B	63.4	66.8	67.1	69.1	66.6
Domain-RAG (Ours)	Swin-B	**65.5**	**67.7**	**68.3**	**70.3**	**68.0**

4.2 Comparison with Other Augmentation Methods

To assess the effectiveness of our Domain-RAG framework, we compare it with several strong baselines that are designed for augmenting data for CD-FSOD. Specifically, 1) "Copy-Paste" directly overlays foreground objects onto random COCO backgrounds without considering semantic relevance or compositional integrity. 2) "Foreground Augmentation" attempts to diversify object appearances by inpainting new foregrounds after object removal. This is done by using the category label of each bounding box as a text prompt and applying SDXL-inpaint to generate a new foreground after removing the original object. 3) "Background Augmentation", which we use InstructBLIP [5] to caption the remaining background, and guide SDXL to generate a new background based on this caption after removing the foreground from a target image. To ensure fair comparison, all the augmentation methods use $G = 5$. Comparison results are summarized in Tab. 4. The results are reported on CD-FSOD under the 1-shot setting.

Table 4: Comparison of augmentation methods (mAP) on the CD-FSOD benchmark under 1-shot.

Method	ArTaxOr	Clipart	DIOR	DeepFish	NEU-DET	UODD	Average
GroundingDINO	26.3	55.3	14.8	36.4	9.3	15.9	26.3
Copy-Paste	38.8	55.0	15.0	36.4	8.4	14.2	27.9
Foreground Augmentation	32.4	56.1	13.9	**41.4**	9.6	14.9	28.1
Background Augmentation	52.3	53.7	16.9	34.2	8.9	10.8	29.5
Domain-RAG (Ours)	**57.2**	**56.1**	**18.0**	38.0	**12.1**	**20.2**	**33.6**

We observe that: 1) copy-paste methods can work reasonably well on relatively simple datasets such as ArTaxOr. However, due to a lack of semantic consistency and domain alignment, they tend to fail on most target domains. 2) Foreground-augmentation baseline performs well when the foreground is visually simple and isolated, for example, in datasets like DeepFish, where only a single object is present. However, due to the potential semantic shift issue, it failed on more complex datasets such as DIOR and UODD. 3) Background-augmentation baseline also suffers in CD-FSOD, often failing on datasets with distinctive domain characteristics, such as NEU-DET. 4) In contrast, our method consistently improves upon the baseline across all datasets, demonstrating its robustness and effectively addressing the limitations of prior approaches.

4.3 More Analysis

Ablation Study on Proposed Modules. To evaluate each module's effectiveness, we conduct ablation studies by removing or replacing it with naive alternatives. As a typical challenging case, the NEU-DET under a 1-shot setting is demonstrated as an example. Results are shown in Fig. 5 (a). Specifically, 1) the grey bar marks the "baseline", i.e., vanilla fine-tuned GroundingDINO. 2) The pink bar ("w/o background retrieval") disables the domain-aware retrieval module and replaces the backgrounds with random COCO images while keeping the rest of the pipeline unchanged. 3) The yellow bar ("w/o background generation") skips the domain-guided background generation and directly performs the foreground-background composition with the raw retrieved images from COCO. 4) The blue bar ("copy-paste as compositional") removes the last foreground–background composition stage and simply pastes the foreground onto the domain-aligned generated background. 5) The last colorful bar represents our full Domain-RAG.

Results show that our full model outperforms all ablated variants, achieving the best overall performance. Furthermore, we observe the following: 1) By comparing our method with the pink bar, we verify that the domain-aware background retrieval stage provides backgrounds that are better aligned with the target domain. 2) The comparison between the gray and yellow bars indicates that simply augmenting backgrounds using COCO images offers limited benefits. In contrast, the domain-guided background generation stage significantly improves performance by producing backgrounds that are both semantically and stylistically aligned, as evidenced by the gap between the yellow and final colorful bars. 3) The performance drop seen with the blue bar underscores the importance of the foreground-background composition stage, which enables seamless integration of foreground objects into the generated backgrounds. Together, these observations confirm that each component of Domain-RAG is both indispensable and complementary for achieving robust CD-FSOD performance.

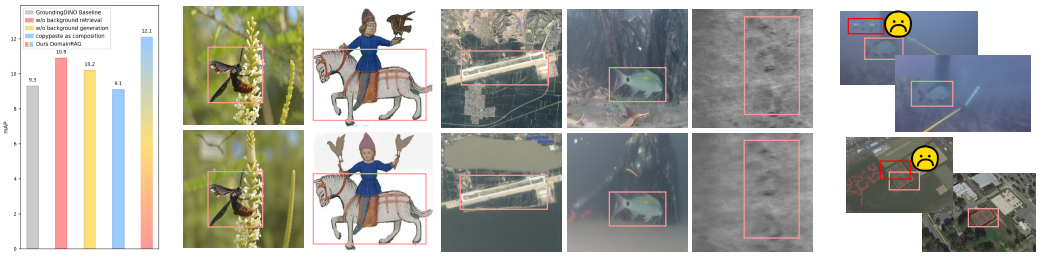
The Construction of RAG Database In the defined (closed-source) CD-FSOD setting as proposed in CD-ViTO [10], COCO serves as the only single-source dataset for training, while other datasets (ArTaxOr, Clipart1k, DIOR, DeepFish, NEU-DET, UODD) are treated as unseen targets. Using COCO as the RAG database brings two key advantages: (1) It does not introduce any extra data beyond the default setting, ensuring the fairness of comparison; (2) COCO provides diverse and general-domain backgrounds that better cover novel domain scenarios.

Furthermore, we conducted additional experiments using different database options, including COCO with reduced category numbers, NEU-DET (non-general-domain), and miniImageNet.

Table 5: Effect of different database choices on Domain-RAG performance.

DataBase	ArTaxOr	Clipart1k	DIOR	FISH	NEU-DET	UODD	Avg
Base (GroundingDINO)	26.3	55.3	14.8	36.4	9.3	15.9	26.3
COCO-1class	50.1	55.0	15.7	36.6	12.0	16.0	30.9
COCO-5classes	51.0	55.1	16.6	36.7	11.9	17.5	31.5
COCO-20classes	53.0	56.2	16.2	37.0	12.2	18.9	32.3
COCO-80classes (Ours)	**57.2**	**56.1**	**18.0**	**38.0**	12.1	**20.2**	**33.6**
NEU-DET	49.8	55.2	16.4	37.0	12.0	16.1	31.1
miniImageNet	55.6	53.2	15.6	38.0	**14.0**	16.2	32.1

From the results summarized in Table 5, we observe that: (1) broader category coverage consistently improves performance; (2) general-domain databases such as COCO outperform specific-domain



(a) Ablation Study On Modules (b) Visualization of Target Image & Generated Image (c) Failure Cases

Figure 3: (a) Ablation study on modules, results reported on NEU-DET, 1 shot. (b) Visualization of target image (top row) and generated image (second row). (c) Failure Cases.

ones like NEU-DET; and (3) although miniImageNet can serve as an alternative database, it performs slightly worse than COCO due to its larger foreground regions and less diverse backgrounds. These findings demonstrate that our Domain-RAG consistently enhances the base GroundingDINO across all benchmarks, validating the robustness and effectiveness of our approach.

Visualization of Generation Images. To provide a more intuitive illustration of our method's effectiveness, we present qualitative results in Fig. 5 (b). Each example shows the original target image from a different domain in the first row and the corresponding generated image in the second row, with annotated object bounding boxes. From the results, we observe the following: 1) Our method effectively preserves the foreground object without introducing noticeable changes, even in challenging domains such as remote sensing, underwater, and industrial defect scenarios. 2) The generated images successfully introduce new backgrounds while maintaining overall semantic coherence and visual consistency with the original domain. Also, the outputs appear natural and realistic. These two observations align well with our goals and further validate the effectiveness of the proposed Domain-RAG framework.

Failure Cases and Limitations. We further examine the quality of the generated images and observe that, in a few cases, our model exhibits foreground information leakage. Parts of the foreground object are unintentionally regenerated within the background, as illustrated in the area highlighted with red boxes of Fig. 5 (c). Since these regenerated foregrounds are not explicitly controlled and lack corresponding annotations, they may introduce noise into model fine-tuning and potentially harm detection performance.

Additional results, including ablation studies of our proposed modules on other targets, more detailed analyses, and extended visualizations, are provided in the Appendix.

5 Conclusion

In this paper, we investigate few-shot object detection (FSOD) across domains—a more realistic yet significantly more challenging scenario than conventional FSOD. We focus on three representative tasks: cross-domain FSOD (CD-FSOD), remote sensing FSOD (RS-FSOD), and camouflaged FSOD. To improve performance under these settings, we propose **Domain-RAG**, a training-free compositional image generation framework designed to produce domain-aligned and detection-friendly samples. Unlike existing text-to-image generation approaches that rely solely on textual prompts, Domain-RAG retrieves semantically and stylistically similar images as structured priors to guide the generation process. To the best of our knowledge, this is the first application of retrieval-augmented generation to cross-domain object detection, particularly in a training-free way suitable for low-shot scenarios. Domain-RAG achieves new state-of-the-art results across all three tasks, demonstrating its generalization ability and opening new directions for training-free data synthesis.

6 Acknowledgment

This work was supported by the Science and Technology Commission of Shanghai Municipality (No. 24511103100).The authors gratefully thank the organization for their support and resources.

References

[1] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022.

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[3] Kai Chen, Enze Xie, Zhe Chen, Yibo Wang, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Geodiffusion: Text-prompted geometric control for object detection data generation. *arXiv preprint arXiv:2306.04607*, 2023.

[4] Gong Cheng, Bowei Yan, Peizhen Shi, Ke Li, Xiwen Yao, Lei Guo, and Junwei Han. Prototype-cnn for few-shot object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–10, 2021.

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

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

[7] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4013–4022, 2020.

[8] Yuqian Fu, Yanwei Fu, and Yu-Gang Jiang. Meta-fdmixup: Cross-domain few-shot learning guided by labeled target data. In *ACM MM*, 2021.

[9] Yuqian Fu, Xingyu Qiu, Bin Ren, Yanwei Fu, Radu Timofte, Nicu Sebe, Ming-Hsuan Yang, Luc Van Gool, Kaijin Zhang, Qingpeng Nong, et al. Ntire 2025 challenge on cross-domain few-shot object detection: Methods and results. *CVPRW*, 2025.

[10] Yuqian Fu, Yu Wang, Yixuan Pan, Lian Huai, Xingyu Qiu, Zeyu Shangguan, Tong Liu, Yanwei Fu, Luc Van Gool, and Xingqun Jiang. Cross-domain few-shot object detection via enhanced open-set object detector. In *European Conference on Computer Vision*, 2024.

[11] Yuqian Fu, Yu Xie, Yanwei Fu, and Yu-Gang Jiang. Styleadv: Meta style adversarial training for cross-domain few-shot learning. In *CVPR*, 2023.

[12] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. 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*, pages 2918–2928, 2021.

[13] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *European Conference on Computer Vision*, 2020.

[14] Jiaming Han, Yuqiang Ren, Jian Ding, Ke Yan, and Gui-Song Xia. Few-shot object detection via variational feature aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 755–763, 2023.

[15] Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *Advances in Neural Information Processing Systems*, 37:132876–132907, 2024.

[16] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[17] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, 2018.

[18] Lihao Jiang, Yi Wang, Qi Jia, Shengwei Xu, Yu Liu, Xin Fan, Haojie Li, Risheng Liu, Xinwei Xue, and Ruili Wang. Underwater species detection using channel sharpening attention. In *ACM MM*, 2021.

[19] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8420–8429, 2019.

[20] Mona Köhler, Markus Eisenbach, and Horst-Michael Gross. Few-shot object detection: A comprehensive survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[21] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[22] Black Forest Labs. Flux.fill. `https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev`, 2024.

[23] Black Forest Labs. Flux.redux. `https://huggingface.co/black-forest-labs/FLUX.1-Redux-dev`, 2024.

[24] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 2020.

[25] Jiaxuan Li, Duc Minh Vo, Akihiro Sugimoto, and Hideki Nakayama. Evcap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13733–13742, 2024.

[26] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS*, 2020.

[27] Pengfang Li, Fang Liu, Licheng Jiao, Shuo Li, Lingling Li, Xu Liu, and Xinyan Huang. Knowledge transduction for cross-domain few-shot learning. *Pattern Recognition*, 141:109652, 2023.

[28] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, 2022.

[29] Hanwen Liang, Qiong Zhang, Peng Dai, and Juwei Lu. Boosting the generalization capability in cross-domain few-shot learning via noise-enhanced supervised autoencoder. In *ICCV*, 2021.

[30] Shaobo Lin, Kun Wang, Xingyu Zeng, and Rui Zhao. Explore the power of synthetic data on few-shot object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 638–647, 2023.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*. Springer, 2014.

[32] Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36:22820–22840, 2023.

[33] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.

[34] Yanxing Liu, Zongxu Pan, Jianwei Yang, Bingchen Zhang, Guangyao Zhou, Yuxin Hu, and Qixiang Ye. Few-shot object detection in remote sensing images via label-consistent classifier and gradual regression. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[37] Xiaonan Lu, Wenhui Diao, Yongqiang Mao, Junxi Li, Peijin Wang, Xian Sun, and Kun Fu. Breaking immutable: Information-coupled prototype elaboration for few-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1844–1852, 2023.

[38] Yuanhuiyi Lyu, Xu Zheng, Lutao Jiang, Yibo Yan, Xin Zou, Huiyu Zhou, Linfeng Zhang, and Xuming Hu. Realrag: Retrieval-augmented realistic image generation via self-reflective contrastive learning. *arXiv preprint arXiv:2502.00848*, 2025.

[39] Thanh-Danh Nguyen, Anh-Khoa Nguyen Vu, Nhat-Duy Nguyen, Vinh-Tiep Nguyen, Thanh Duc Ngo, Thanh-Toan Do, Minh-Triet Tran, and Tam V Nguyen. The art of camouflage: Few-shot learning for animal detection and segmentation. *IEEE Access*, 2024.

[40] Minheng Ni, Zitong Huang, Kailai Feng, and Wangmeng Zuo. Imaginarynet: Learning object detectors without real images and annotations. *arXiv preprint arXiv:2210.06886*, 2022.

[41] Joachim Niemeyer, Franz Rottensteiner, and Uwe Soergel. Contextual classification of lidar data and building object detection in urban areas. *ISPRS journal of photogrammetry and remote sensing*, 87:152–165, 2014.

[42] Jiancheng Pan, Yanxing Liu, Xiao He, Long Peng, Jiahao Li, Yuze Sun, and Xiaomeng Huang. Enhance then search: An augmentation-search strategy with foundation models for cross-domain few-shot object detection. In *CVPRW*, 2025.

[43] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[44] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection. *arXiv preprint arXiv:2108.09017*, 2021.

[45] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849, 2023.

[46] Alzayat Saleh, Issam H Laradji, Dmitry A Konovalov, Michael Bradley, David Vazquez, and Marcus Sheaves. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific Reports*, 2020.

[47] Kechen Song and Yunhui Yan. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Applied Surface Science*, 2013.

[48] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *computer vision and pattern recognition*, 2021.

[49] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022.

[50] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. In *ICLR*, 2020.

[51] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*, 2020.

[52] Wuti Xiong. 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)*, pages 1–5. IEEE, 2023.

[53] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9577–9586, 2019.

[54] Ji Zhang, Jingkuan Song, Lianli Gao, and Hengtao Shen. Free-lunch for cross-domain few-shot learning: Style-aware episodic training with robust contrastive learning. In *Proceedings of the 30th ACM international conference on multimedia*, pages 2586–2594, 2022.

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[56] Manlin Zhang, Jie Wu, Yuxi Ren, Ming Li, Jie Qin, Xuefeng Xiao, Wei Liu, Rui Wang, Min Zheng, and Andy J Ma. Diffusionengine: Diffusion model is scalable data engine for object detection. *arXiv preprint arXiv:2309.03893*, 2023.

[57] Tong Zhang, Yin Zhuang, Xinyi Zhang, Guanqun Wang, He Chen, and Fukun Bi. Advancing controllable diffusion model for few-shot object detection in optical remote sensing imagery. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 7600–7603. IEEE, 2024.

[58] Xinyu Zhang, Yuhan Liu, Yuting Wang, and Abdeslam Boularias. Detect everything with few examples. *arXiv preprint arXiv:2309.12969*, 2023.

[59] Hanqing Zhao, Dianmo Sheng, Jianmin Bao, Dongdong Chen, Dong Chen, Fang Wen, Lu Yuan, Ce Liu, Wenbo Zhou, Qi Chu, et al. X-paste: Revisiting scalable copy-paste for instance segmentation using clip and stablediffusion. In *International Conference on Machine Learning*, pages 42098–42109. PMLR, 2023.

[60] Xu Zheng, Ziqiao Weng, Yuanhuiyi Lyu, Lutao Jiang, Haiwei Xue, Bin Ren, Danda Paudel, Nicu Sebe, Luc Van Gool, and Xuming Hu. Retrieval augmented generation and understanding in vision: A survey and new outlook. *arXiv preprint arXiv:2503.18016*, 2025.

[61] Fei Zhou, Peng Wang, Lei Zhang, Wei Wei, and Yanning Zhang. Revisiting prototypical network for cross domain few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20061–20070, 2023.

[62] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022.

[63] Linhai Zhuo, Yuqian Fu, Jingjing Chen, Yixin Cao, and Yu-Gang Jiang. Tgdm: Target guided dynamic mixup for cross-domain few-shot learning. In *ACM MM*, 2022.

[64] Yixiong Zou, Shuai Yi, Yuhua Li, and Ruixuan Li. A closer look at the cls token for cross-domain few-shot learning. *Advances in Neural Information Processing Systems*, 37:85523–85545, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state that the paper proposes DomainRAG, a training-free, plug-and-play background augmentation framework designed for cross-domain few-shot object detection. We explicitly frame the task as CD-FSOD, describe the retrieval-augmented generation pipeline, and clarify that improvements are demonstrated across multiple datasets.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses a specific limitation of the proposed image generation method, namely the unintended foreground information leakage into the background. This phenomenon, highlighted in Fig. 5(c), is identified as a source of noise that may negatively impact model fine-tuning and downstream detection performance. The authors acknowledge this issue and explain its implications, demonstrating awareness of their method's boundaries.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper presents an empirical augmentation pipeline with no formal theorems or proofs; therefore this question is not applicable.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides sufficient details to reproduce the main experimental results. It includes comprehensive descriptions of the datasets used, model architectures, training settings, evaluation metrics, and implementation details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We plan to release all code and data used in the experiments. The final version will include detailed instructions in the supplemental material to ensure faithful reproduction of the main results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all necessary experimental details, including training and test splits, hyperparameters, optimizer types, learning rate schedules, and other implementation specifics.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We validate our method on three tasks, and each dataset follows the settings used in previous works.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Implementation Details (Sec. 4) specifies the exact hardware—1× A800 (80 GB) for training and 4× V100 (32 GB) for inference—as well as the wall-time (12 GPU·h for full training) and peak memory usage (48 GB per GPU).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm that our study complies with it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our paper includes a "Broader Impact" paragraph in Sec. 7. **Positive:** Domain-RAG can improve object detection in low-data domains such as environmental monitoring and medical or remote-sensing imagery, enabling faster deployment where manual annotation is costly. **Negative:** Like other generative augmentations, it could be misused to create synthetic datasets that bias detectors or aid surveillance. We discuss these risks and note that our method does not release new high-capacity generators; it reuses publicly licensed assets, and safeguards follow the original licenses.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work releases no new model or dataset with elevated misuse risk; it only employs publicly available assets (e.g., Flux diffusion model and public vision datasets) under their original licenses. Therefore no additional safeguards are required or applicable.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All third-party assets are properly credited and their licenses are listed in Appendix A, Table A.1. Specifically, the LaMa model (MIT), Flux diffusion model (Apache 2.0), Grounding DINO (Apache 2.0), CLIP (MIT), and ResNet-50 (MIT). For each asset we state the version or commit hash and provide a URL.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We plan to release the code and implementation details upon acceptance. The released assets will be accompanied by clear documentation, including usage instructions, model configurations, and license terms.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our study relies solely on publicly available computer-vision datasets and automated evaluation; it involves no crowdsourcing tasks or human-subject experiments.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work uses the publicly released Flux diffusion model but does not release any new high-risk model or dataset; safeguards are governed by the original Flux license.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Our work does not employ any LLM in its core methodology, experiments, or analysis; hence no LLM usage needs to be declared.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.