

## A Technical Appendices and Supplementary Material

We first provide the implementation details in Sec. A.1. Then, in Sec. A.2, we present additional ablation studies of our proposed Domain-RAG method. Finally, Sec. A.3 includes more visualization results and further analysis.

### A.1 Implementation Details.

#### A.1.1 More Details on Proposed Method.

**Domain-Aware Background Retrieval.** Considering the large domain gap, we additionally include the inpainted target support set in the background retrieval pool  $\mathcal{D}_{base}$ . All the images contained in  $\mathcal{D}_{base}$  are encoded with a CLIP vision encoder. For the retrieval process, we set  $m = 100$  and  $n = 5$ . In other words, using the inpainted source image as the query, we select the 100 images with the highest cosine similarity in the CLIP embedding space as semantically aligned candidates. From these, we extract style descriptors using the first four layers of a ResNet-50 and choose the 5 images with the smallest  $L_2$  distance to the query, yielding the final set of retrieved backgrounds.

**Domain-Guided Background Generation.** In this phase, the target image and each retrieved image are processed by Redux [23]; their embeddings are combined by a weighted sum (1.0 for the target, 0.8 for the retrieval) without any additional textual prompt. The resulting embedding is fed into the FLUX [21] diffusion model with a guidance scale of 2.5 and 50 sampling steps to synthesize a  $1024 \times 1024$  background image.

**Foreground-Background Composition.** In the generation stage, we employ the FLUX-Fill [22] model. Given a target image and its background, we supply FLUX-Fill with the source image, an object-mask that excludes the out-painting bounding box, and a prompt embedding extracted from the second-stage background image via the Redux encoder; no additional textual cue is provided and both weights are kept at 1. Because FLUX-Fill struggles with very small inputs, we introduce an adaptive rescaling strategy. For UODD [18], where bounding boxes are tiny, we preserve the aspect ratio and iteratively upsample the image until its longer side exceeds 2048 pixels, then generate a corresponding upsampled mask. Conversely, in ArTaxOr [6], some images exceed  $4000 \times 3000$  pixels and would exhaust GPU memory; whenever either edge is larger than 2800 pixels, we downsample by an integer factor and create a matching mask. The (up- or down-)sampled target image, its mask, and the Redux embedding are then passed to FLUX-Fill. We keep the guidance scale at 30.0 but modulate the overall noise strength to suit each target, with 0.8 for FISH [46] and DIOR [26], 0.9 for ArTaxOr and clipart1k [17], 0.3 for NEU-DET [47] 0.8 for NWPU VHR-10 [41], 0.6 for Camouflage [39] FSOD benchmark, and 0.4 for UODD.

#### A.1.2 More Details on Training.

During training, following Grounding DINO [33] and ETS [42], we apply diverse data augmentations, including Mosaic, MixUp, color jittering, random flipping, multi-scale resizing, and cropping, to improve few-shot generalization of both the base GroundingDINO and our Domain-RAG. For evaluation, we follow COCO metrics and disable all augmentations.

### A.2 More Ablation Studies.

In Sec. 4.3, we reported the ablation study on each proposed module using NEU-DET under the 1-shot setting as a representative example. To provide a more comprehensive analysis, in Tab. 6, we report the same ablation experiments but conducted on all six CD-FSOD target domains. This includes evaluations of the impact of removing domain-aware background retrieval, domain-guided background generation, and compositional generation stages.

As shown in Tab. 6, starting from the baseline GroundingDINO (average 26.3), incorporating our modules makes our full Domain-RAG model attain the highest average mAP (33.6), validating the complementary benefits of domain-aware background retrieval, generation, and compositional synthesis for cross-domain few-shot detection. More specifically, 1) removing the background retrieval module causes a clear drop in average performance (33.6 to 31.8), demonstrating its essential role in capturing domain-relevant context. 2) Without background generation, the model achieves a slightly lower average (31.1) and notably fails on UODD, indicating less stable generalization.

3) Replacing the compositional synthesis stage with a simple copy-paste strategy further degrades performance to 29.2, confirming the advantage of our generative approach.

Table 6: Ablation study results under the 1-shot setting on different datasets.

Method	ArTaxOr	Clipart	DIOR	DeepFish	NEU-DET	UODD	Avg.
Baseline (GroundingDINO)	26.3	55.3	14.8	36.4	9.3	15.9	26.3
w/o Background Retrieval	51.7	57.3	17.0	37.2	10.9	16.4	31.8
w/o Background Generation	48.4	54.6	17.2	37.9	10.2	18.0	31.1
Copy-Paste as Composition	46.6	56.2	16.7	35.0	9.1	11.7	29.2
<b>Domain-RAG (Ours)</b>	<b>57.2</b>	<b>56.1</b>	<b>18.0</b>	<b>38.0</b>	<b>12.1</b>	<b>20.2</b>	<b>33.6</b>

### A.2.1 Ablation on Domain-Aware Background Retrieval.

**Ablation on Different Database.** we set COCO as our database in our main exp setting, while we are also interested in validate our idea on other dataset. Therefore, we conduct additional experiments where we replace COCO with MiniImageNet as the retrieval source. As shown in Tab. 7, COCO achieves better performance on most domains, likely due to its richer scene diversity and natural statistics. Performance on DeepFish remains similar across both databases. MiniImageNet is originally designed for classification tasks, and compared to the COCO dataset, it typically contains larger foreground objects and less informative background content. As a result, using MiniImageNet as the retrieval database leads to weaker performance than using COCO. Nevertheless, it still outperforms the baseline, indicating the effectiveness of retrieval-based background generation.

Table 7: Ablation on different background retrieval databases for 1-shot CD-FSOD across six datasets.

Database	ArTaxOr	Clipart	DIOR	DeepFish	NEU-DET	UODD	Avg.
Baseline	26.3	55.3	14.8	36.4	9.3	15.9	26.3
MiniImagenet	55.6	53.2	15.6	<b>38.0</b>	<b>14.0</b>	16.2	32.1
<b>COCO (Ours)</b>	<b>57.2</b>	<b>56.1</b>	<b>18.0</b>	<b>38.0</b>	12.1	<b>20.2</b>	<b>33.6</b>

**Ablation on Retrieval Strategy.** As stated in Sec. 3, we use both of the CLIP-semantic and ResNet-style to retrieve the background images from  $\mathcal{D}_{base}$ ; thus, we compare our method with only CLIP-semantic, ResNet-style. Compared to using only CLIP features or only style features, our method achieves higher average performance, as shown in Tab. 8. This demonstrates the effectiveness of combining both CLIP and style features in our framework.

Table 8: Ablation on retrieval strategy under the 1-shot setting. We compare CLIP-based, ResNet-style, and our combined retrieval strategy. Results are reported on six CD-FSOD benchmarks.

Retrieval Strategy	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Avg.
Baseline	26.3	55.3	14.8	36.4	9.3	15.9	26.3
CLIP Only	50.0	<b>56.2</b>	16.3	36.0	<b>13.0</b>	17.7	31.5
Style Only	48.8	54.5	16.7	35.5	12.1	16.5	30.7
<b>CLIP + Style (Ours)</b>	<b>57.2</b>	56.1	<b>18.0</b>	<b>38.0</b>	12.1	<b>20.2</b>	<b>33.6</b>

**Ablation on Number of Retrieved Images.** We set  $n = 5$  in our main results. Here, we also conduct different choices of  $n$ , such as  $n = 1$ ,  $n = 3$ , and  $n = 10$ , to better investigate our methods. As summarized in Tab. 9, we observe that retrieving  $n = 5$  images yields the best performance under the 1-shot setting. Smaller values ( $n = 1$  or  $n = 3$ ) provide limited diversity and result in weaker generalization. In contrast, setting  $n = 10$  often leads to a performance drop, suggesting that more retrieved samples do not always bring further gains. This trend is largely due to our data generation process, which composes retrieved foregrounds with background scenes. Too few retrieved images limit visual variation, while too many increase the chance of unrealistic compositions (e.g., a sofa on a grass field). These unnatural contexts may confuse the model and reduce its ability to align with real-world test distributions.

Table 9: Ablation on the number of retrieved images ( $n$ ) under the 1-shot setting. Results are reported on six CD-FSOD benchmarks.

Retrieved Images	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Avg.
Baseline	26.3	55.3	14.8	36.4	9.3	15.9	26.3
$n = 1$	43.2	54.3	16.6	37.1	9.4	14.7	29.2
$n = 3$	49.8	54.5	16.3	36.7	12.7	18.1	31.4
$n = 5$ (Ours)	<b>57.2</b>	56.1	<b>18.0</b>	38.0	12.1	<b>20.2</b>	<b>33.6</b>
$n = 10$	49.6	<b>56.8</b>	16.6	<b>39.3</b>	<b>13.2</b>	17.3	32.1

### A.2.2 Ablation on Domain-Guided Background Generation.

**Ablation on with/without Initial Background Guidance.** For the Ablation on with/without initial background guidance, we keep all other components of the framework unchanged. During background generation, we do not use the initial background. Instead, we only use the retrieval image to obtain the prompt embedding via Redux, and then feed  $0.8 \times$  the retrieval image’s prompt embedding into the Flux model to generate the background. As shown in Tab. 10, incorporating initial background guidance consistently improves performance under the 1-shot setting across six CD-FSOD datasets. Removing this guidance leads to a noticeable drop in average accuracy, especially on challenging domains like UODD. This demonstrates that initial background information plays an important role in stabilizing and enhancing background generation quality.

Table 10: Ablation on initial background guidance for 1-shot CD-FSOD across six datasets.

Method	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Avg.
Baseline	26.3	55.3	14.8	36.4	9.3	15.9	26.3
W/o Bg Guidance	49.8	56.0	16.3	<b>38.0</b>	<b>12.7</b>	15.4	31.4
<b>With Bg Guidance (Ours)</b>	<b>57.2</b>	<b>56.1</b>	<b>18.0</b>	<b>38.0</b>	12.1	<b>20.2</b>	<b>33.6</b>

**Ablation on Text-to-Image Backbones.** In this section, we conduct ablation experiments on the background generation module. In the proposed framework, we use the Redux module to fuse features from the retrieval image and the inpainted image, and convert them into a prompt embedding. For the ablation setting, we replace Redux with InstructBLIP and use Diffusion XL to generate backgrounds. Specifically, we use InstructBLIP to extract captions for both the retrieval image and the inpainted image. The resulting captions are then concatenated in the form of " $< caption_1 > . < caption_2 >$ ". Since Diffusion XL uses a CLIP text encoder that cannot handle excessively long texts, we constrain InstructBLIP to generate captions of no more than 20 words, and truncate any prompt exceeding 40 words before feeding it into the model. In this ablation, the background generation strategy is only applied to the domain-guided background generation, while all other components of the pipeline remain unchanged. As shown in Tab. 11, generating richer backgrounds contributes positively to the final generation quality. Our adopted Flux + Redux approach provides more effective supervision, guiding the generation process to produce data that better aligns with the target domain.

Table 11: Ablation on text-to-image backbones for 1-shot CD-FSOD across six datasets.

Method	ArTaxOr	Clipart	DIOR	DeepFish	NEU-DET	UODD	Avg.
Baseline	26.3	55.3	14.8	36.4	9.3	15.9	26.3
InstructBLIP + Diffusion XL	49.0	54.2	15.8	<b>38.6</b>	11.8	14.3	30.6
<b>Redux + Flux (Ours)</b>	<b>57.2</b>	<b>56.1</b>	<b>18.0</b>	38.0	<b>12.1</b>	<b>20.2</b>	<b>33.6</b>

### A.2.3 Ablation on Foreground-Background Composition.

**Ablation on Text-to-Image Backbones.** In our framework, we utilize the semantic information provided by Redux in the third stage to guide image generation. In the ablation experiments for this stage, we replace the Redux module with InstructBLIP, and substitute the Flux-Fill model with the Diffusion-XL-Inpaint model to generate new backgrounds. Specifically, for each background retrieved during the background retrieval stage, we use InstructBLIP to extract a caption prompt describing the image. After obtaining the prompt, we feed the Target image, the corresponding mask, and the caption (as an instruction to modify the background) into the Diffusion-XL-Inpaint model to

synthesize a new background. As shown in Tab. 12, compared to the baseline using InstructBLIP and Diffusion-XL-Inpaint, our method (Redux + Flux) achieves consistently better performance, especially on ArTaxOr and UODD, demonstrating the effectiveness of Redux guidance and Flux in generating semantically coherent and diverse images.

Table 12: Ablation on text-to-image backbones under the 1-shot setting. Results are reported on six CD-FSOD benchmarks.

Backbone	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Avg.
Baseline	26.3	55.3	14.8	36.4	9.3	15.9	26.3
Text2Image	44.4	54.5	14.7	35.1	11.9	15.6	29.4
<b>Redux + Flux (Ours)</b>	<b>57.2</b>	<b>56.1</b>	<b>18.0</b>	<b>38.0</b>	<b>12.1</b>	<b>20.2</b>	<b>33.6</b>

### A.3 More Visualization and Analysis.

#### A.3.1 More Analysis on Generated Image Quantity.

We evaluate the quality of generated images using CLIP-I similarity and Fréchet Inception Distance (FID). CLIP-I measures the average cosine similarity between the target image and generated samples using the CLIP image encoder, while FID assesses visual fidelity based on InceptionV3 features. Due to its reliance on distribution statistics, FID is not computed for DeepFish in the 1-shot setting, where only one target image is available. Higher CLIP-I and lower FID indicate better semantic alignment and image realism, respectively.

Table 13: Evaluation of different augmentation methods on the CD-FSOD 1-shot benchmark using CLIP-I and FID.

Method	ArTAXOr		Clipart		DIOR		DeepFish		NEU-DET		UODD	
	CLIP-I	FID	CLIP-I	FID	CLIP-I	FID	CLIP-I	FID	CLIP-I	FID	CLIP-I	FID
GroundingDINO	-	-	-	-	-	-	-	-	-	-	-	-
Copy-Paste	62.9	321.4	60.1	312.8	52.8	353.6	60.9	-	49.7	476.2	50.4	396.2
Foreground Aug	88.1	165.0	89.3	107.7	95.1	131.1	93.7	-	89.3	129.3	90.9	27.0
Background Aug	89.7	78.7	85.0	128.0	83.7	265.4	68.9	-	72.8	453.7	78.1	287.8
<b>Domain-RAG (Ours)</b>	<b>92.6</b>	<b>70.5</b>	<b>88.5</b>	<b>117.6</b>	<b>79.8</b>	<b>288.8</b>	<b>77.1</b>	<b>-</b>	<b>93.3</b>	<b>127.4</b>	<b>79.3</b>	<b>289.6</b>

As shown in Tab. 13 and visualized in Fig. 4, the "Copy-Paste" method performs poorly on both CLIP-I and FID due to its use of randomly selected COCO backgrounds, resulting in a distribution mismatch with the target domain. "Foreground Aug" yields very high CLIP-I and low FID, as only small local regions are modified, making the overall image similar to the original image. In contrast, "Background Aug" alters large portions of the image, leading to lower CLIP-I and higher FID. Our Domain-RAG achieves a balanced trade-off, maintaining domain relevance while introducing sufficient visual diversity.

In addition, we would like to argue that though CLIP and InceptionV3 are widely used for image similarity evaluation, their general-purpose nature can lead to unreliable assessments in cross-domain settings. Our goal is not to produce images that are distributionally identical to the originals, but to enrich semantic diversity while preserving domain-specific features. Therefore, higher CLIP-I or lower FID does not necessarily indicate better generation quality in our case. Instead, the quantitative results demonstrate the effectiveness of our approach.

#### A.3.2 Visualization Results from Each Stage.

To better illustrate the effectiveness and progression of our data generation pipeline, we provide visualizations of intermediate outputs from each stage, as shown in Fig.5. From left to right, Fig.5 presents: (1) the target query image, (2) the retrieved support images from the domain-aware background retrieval stage, (3) the generated background images from the domain-guided background generation stage, (4) the final synthesized image from the foreground-background composition stage, which is ultimately used for training. This progressive visualization highlights how each stage contributes to generating diverse and semantically meaningful training samples that align with the target domain.



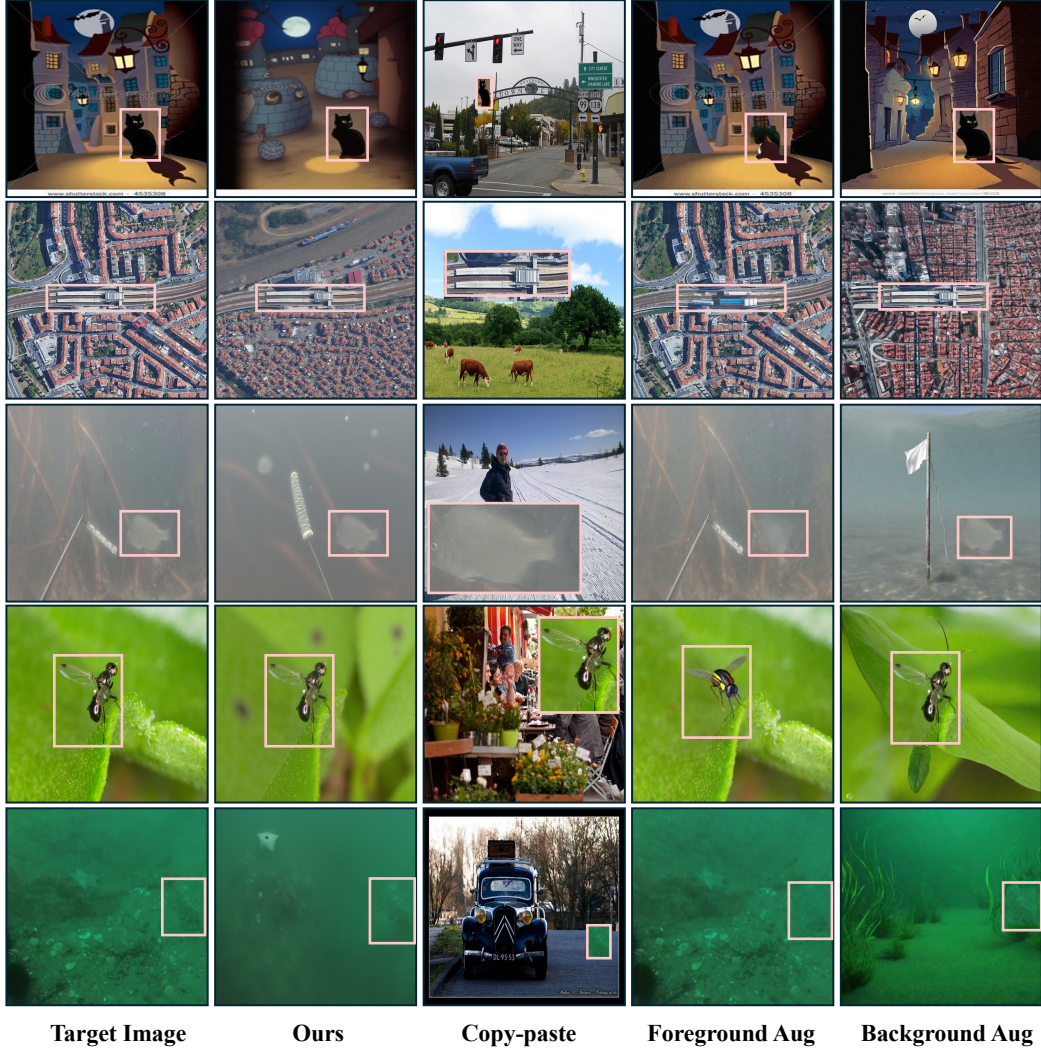


Figure 4: Visualization comparison between DomainRAG and other augmentation methods.

### A.3.3 More Visualization Results of our Domain-RAG.

Fig.6 presents additional qualitative results of our Domain-RAG module across multiple domains. Each row shows two examples, arranged from left to right as: (a) target image, (b) generated image, (c) target image, and (d) generated image. The results demonstrate that our method generates visually consistent and domain-aware backgrounds across diverse visual styles, including artistic, aerial, underwater, and industrial scenes.

### A.3.4 More Analysis on Limitations and Future Work.

As stated in Sec. 4.3, our model exhibits foreground information leakage. We attribute this issue to the limited generation quality of Simple LaMa Inpaint—for target images with large foregrounds, the inpainting results are often suboptimal, frequently showing blurriness or patch artifacts after foreground removal. The pretrained Redux module typically struggles to handle these artifacts effectively, often preserving them in subsequent generation steps, which in turn degrades the overall generation quality. Enhancing the capability of the Redux module to better support such cases and mitigate foreground leakage remains an important direction for future work. In addition, we currently lack an effective filtering mechanism. Mostly commonly used and general-purpose vision backbones, e.g., CLIP, fail to deliver ideal filtering results in our cross-domain scenarios. To address

this limitation, future work could explore the integration of more powerful vision-language models to enable more precise and background-aware filtering strategies.

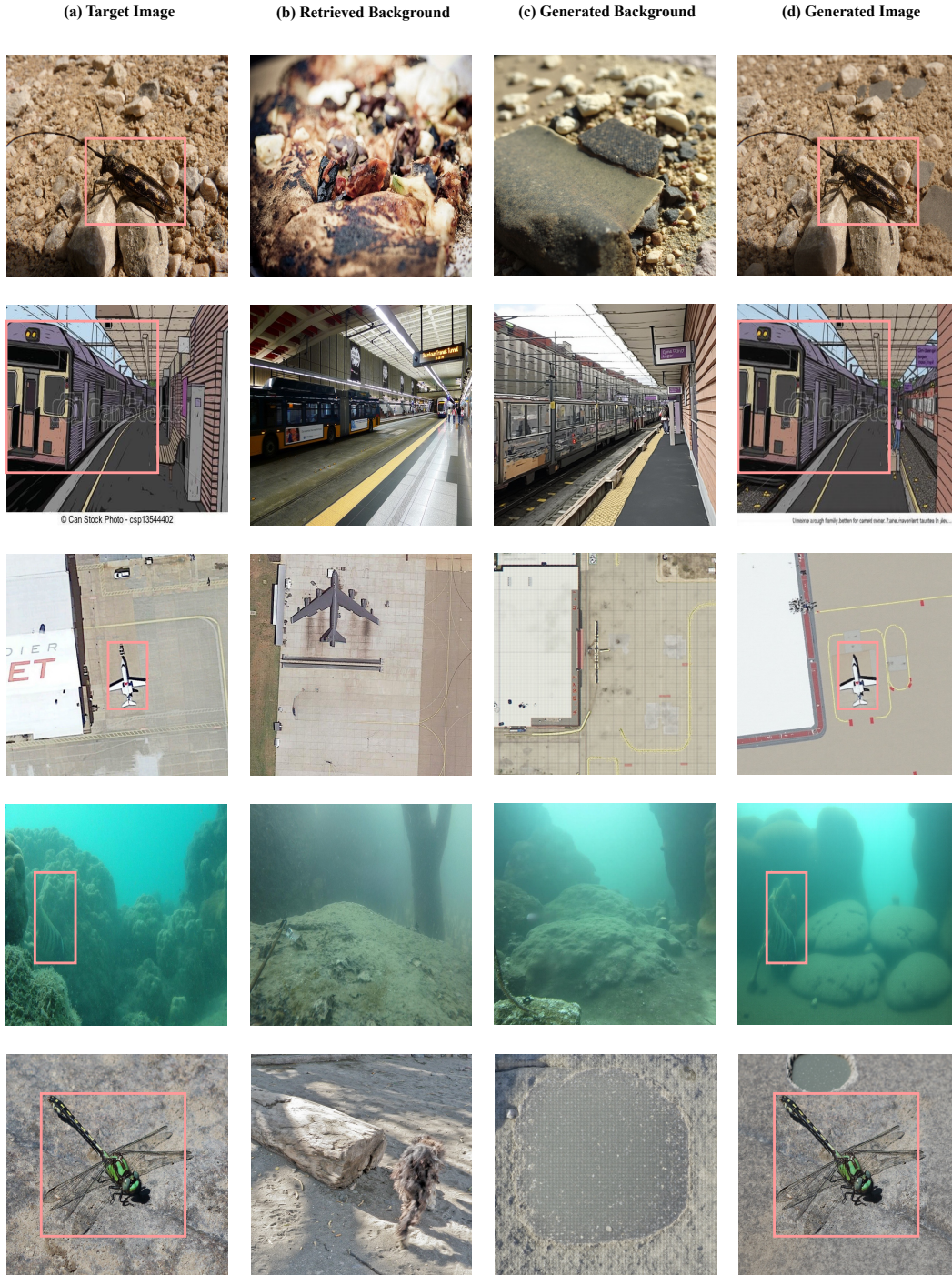


Figure 5: Visualization of our data generation pipeline. From left to right: (1) the target query image, (2) retrieved backgrounds, (3) generated new backgrounds, and (4) the final synthesized image used for further model finetuning.



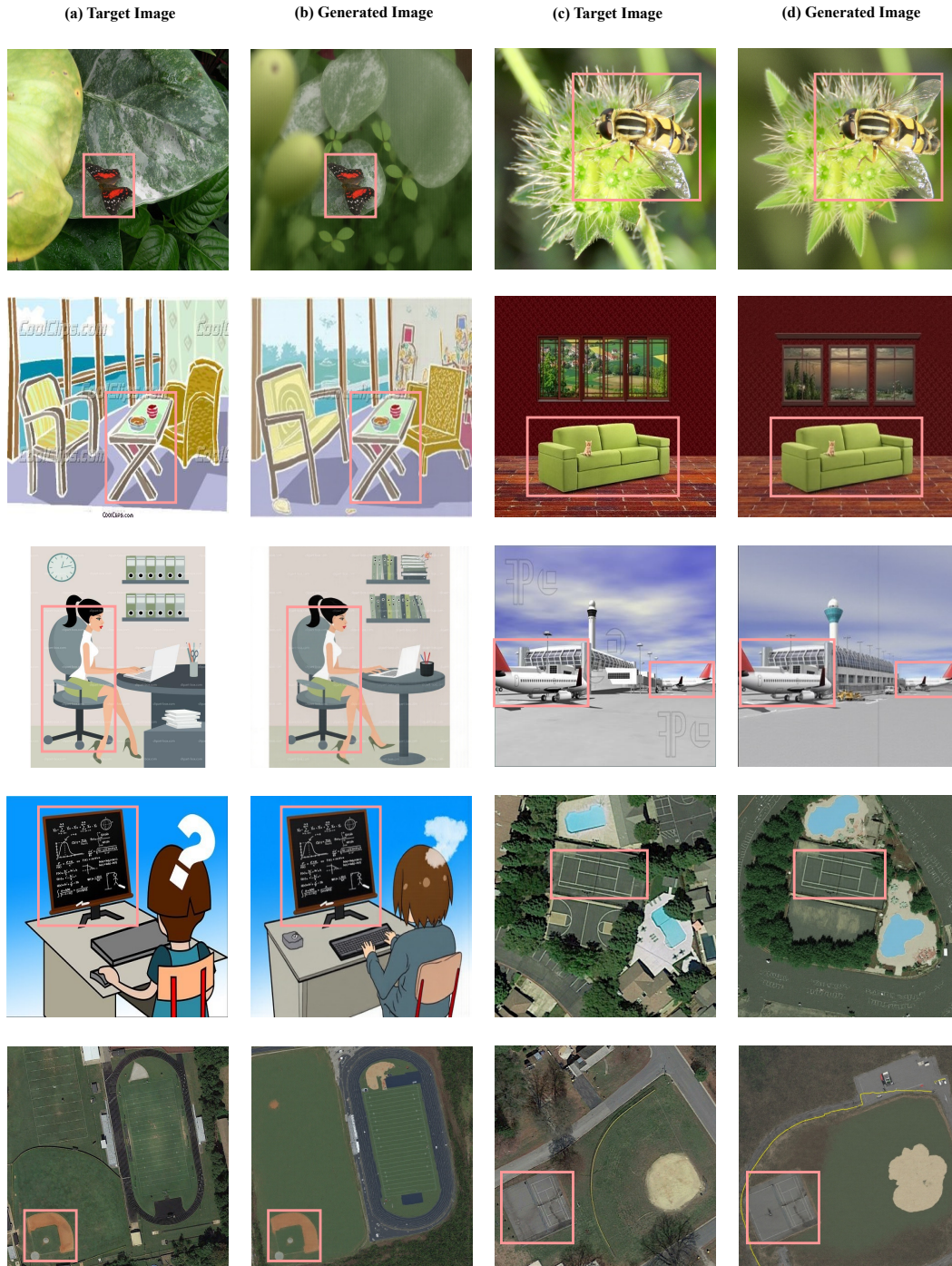


Figure 6: More visualization results of our Domain-RAG module across different domains. Each row shows two pairs of query images and their corresponding generated backgrounds. Our method demonstrates the ability to generate semantically aligned and domain-aware backgrounds across diverse visual domains such as artistic, aerial, underwater, and industrial scenes.