Overcoming Shortcut Problem in VLM for Robust Out-of-Distribution Detection

Supplementary Material

1. The proposed ImageNet-bg

To further evaluate the robustness of out-of-distribution (OOD) detectors against background interference, we propose a new OOD test set ImageNet-Bg, derived from the original ImageNet validation set, containing 48,285 images. This dataset is created by removing all the in-distribution (ID)-related content from the samples in the ImageNet validation set, by Eq.(2) from the main paper. This method ensures that the resulting images primarily feature background elements rather than identifiable objects. Fig. 1 presents several representative examples from ImageNet-Bg.

However, the images with ID regions removed using only Eq.(2) from the main paper still present several issues: (1) Some images appear unnatural after the ID-related areas are removed, and (2) ID-related regions are not entirely eliminated, leading to the retention of some residual ID information. Several problematic examples from ImageNet-Bg are illustrated in Fig. 2. To address these issues, we apply additional filtering criteria based on Eq.(3) from the main paper to further refine the dataset. This filtering process creates the ImageNet-Bg(S) test set, which contains 24,863 images that primarily feature clearer background information, thereby reducing any residual ID-related elements. Consequently, ImageNet-Bg(S) includes fewer problematic cases shown in Fig. 2 and includes more images like those illustrated in Fig. 1.

Nonetheless, since we filter images based on CLIP responses, some cases that are not problematic but prone to shortcuts have also been excluded from ImageNet-Bg(S). Therefore, we recommend using both ImageNet-Bg(S) and ImageNet-Bg to thoroughly evaluate the model's robustness against background interference.

2. Experimental Setting

2.1. Descriptions of datasets

2.1.1 ID Dataset

We utilize ImageNet-1k [3] as ID dataset, which comprises 1,000 categories with 1,281,167 training images and 50,000 images for validation. For few-shot settings, we employ 1, 2, 4, 8, and 16 shots per class for training. For evaluation, we use the ImageNet validation dataset, consisting of 50,000 images across 1,000 classes.

2.1.2 OOD Dataset

Following the setting from [5], we use the commonly used iNaturalist [15], Places [21], SUN [16] and Texture [2] as OOD datasets. Besides, we also use our proposed ImageNet-Bg and ImageNet-Bg(S) as OOD datasets. Additional details about these datasets are provided below.

iNaturalist. iNaturalist [15] is a large-scale dataset of real-world nature images, including 859,000 images of plants and animals across more than 5,000 species. For our evaluation, we use the subset of 10,000 images from 110 classes that do not overlap with ImageNet-1K.

SUN. SUN [16] is a large-scale scene dataset with various scene images across 397 categories. For our evaluation, we utilize a subset of 10,000 images from 50 classes that do not overlap those classes in ImageNet-1K.

Places. Similar to SUN dataset, Places [21] is also a comprehensive scene dataset. For our evaluation, we use a subset of 10,000 images from 50 categories, ensuring no overlap with ImageNet-1K.

Texture. The Describable Textures Dataset [2] includes 5,640 texture images across 47 distinct classes. The entire dataset is utilized for evaluation.

ImageNet-Bg. ImageNet-Bg is a synthetic background interference OOD test set containing 48,285 background images. The images in this dataset are generated by inpainting models. This process involves removing ID-related information from the images in the ImageNet validation set and inpainting the ID regions with background information.

ImageNet-Bg(S). ImageNet-Bg(S) is a synthetic background interference OOD test set sampled from ImageNet-Bg, containing 24,863 images. Compared to ImageNet-Bg, the images in ImageNet-Bg(S) feature cleaner and more natural backgrounds with less ID information.

2.2. Implementation Details

To obtain the mask for the ID-related areas, which will be inpainted with background information, we utilize the Grounded SAM model [12]. For the inpainting process, we use the LaMa model [14]. For the learnable modules, we incorporate 16 learnable parameters following CoOp [22] to fine-tune CLIP for the OOD detection task. All learnable modules are trained for 50 epochs using the SGD optimizer with a learning rate of 2e-3. We utilize a cosine learning rate scheduler with a constant warmup period, where the warmup epoch is set to 1 and the initial warmup learning rate is 1e-5. The batch size is set to 32. For the hyper-parameters, we set the similarity threshold ϵ to 5, which filters out 50% of the samples. λ_{out}^{r} to 1.5 and λ_{out}^{g} and 0.5 respectively. Additionally, for all experiments involving ID augmentation, we set the number of epochs to 20. The main experiments were implemented using PyTorch on an NVIDIA RTX 4090 with 24 GB of memory.

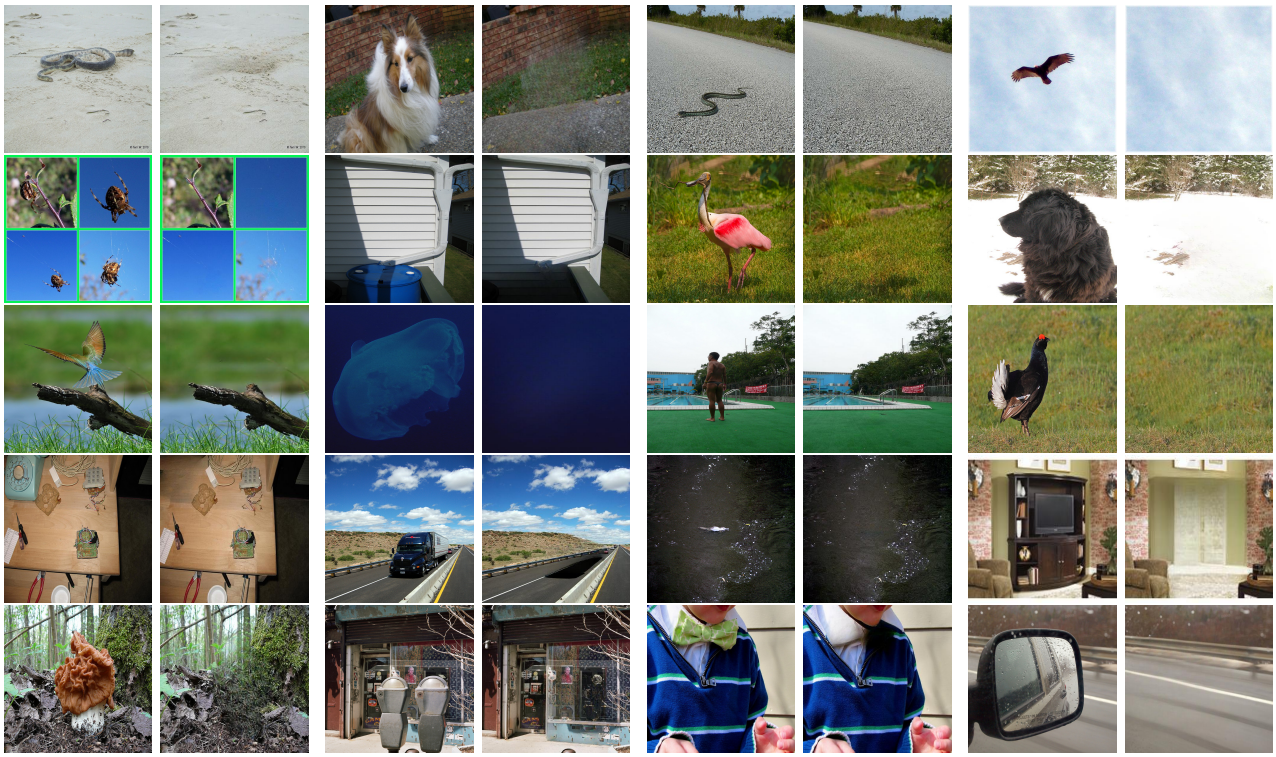

Figure 1. Representative cases from ImageNet-Bg (right) and the ImageNet validation set (left).

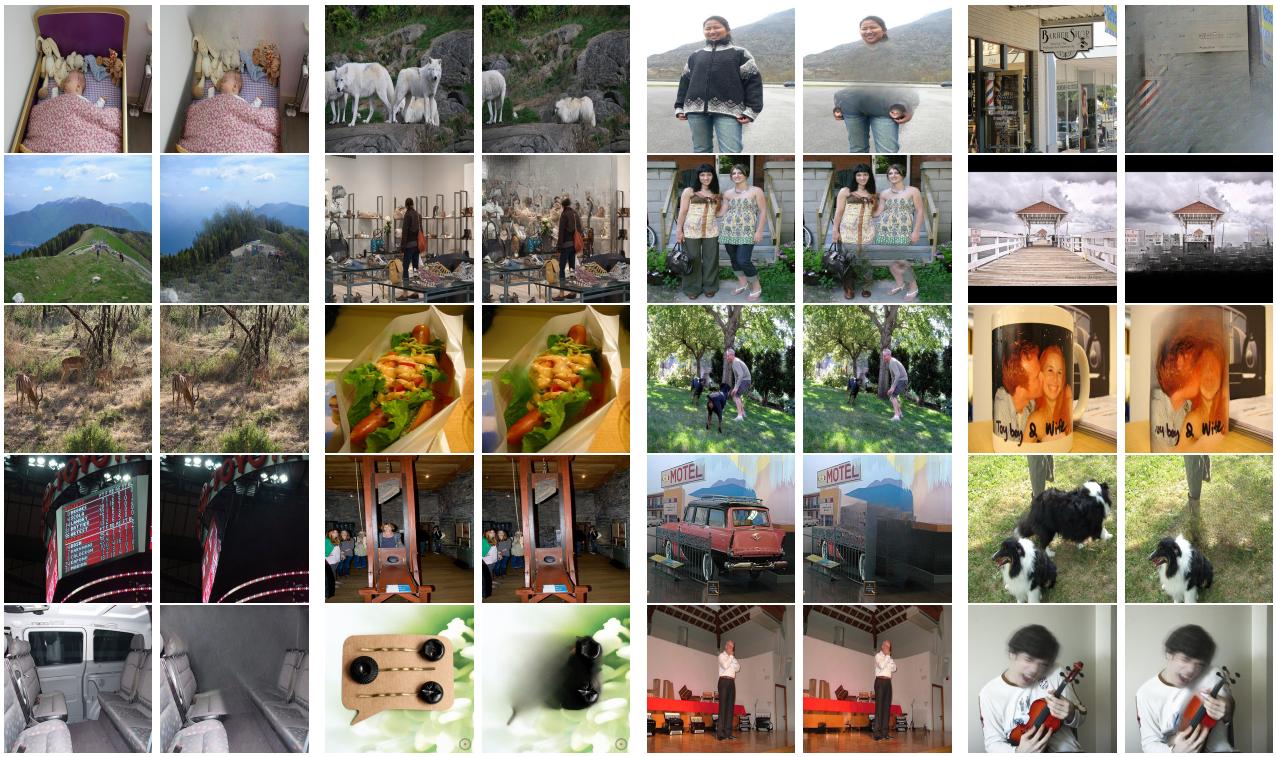

Figure 2. Bad cases from ImageNet-Bg (right) and the ImageNet test set (left).

Method	iNaturalist		SUN		Places		Texture		Avg	
	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓
One-shot										
CoOp [22]	91.41	42.61	91.39	39.70	88.43	47.34	88.37	48.04	89.90	44.42
LoCoOp [10]	94.97	25.20	94.17	29.10	90.90	37.67	87.45	51.87	91.87	35.96
OSPCoOp	95.54	25.03	94.82	29.85	91.93	35.99	90.73	41.25	93.25	33.03
Two-shot										
CoOp [22]	94.47	28.52	92.53	34.94	89.60	42.83	88.78	46.78	91.35	38.27
LoCoOp [10]	94.49	25.69	93.89	30.11	90.61	38.86	89.49	45.30	92.13	34.99
OSPCoOp	94.06	33.26	95.51	23.35	92.39	31.61	90.88	41.63	93.21	32.46
Four-shot										
CoOp [22]	92.61	36.32	91.72	39.92	88.69	46.73	89.18	44.74	90.55	41.93
LoCoOp [10]	93.76	29.83	93.04	34.03	90.50	39.83	89.30	46.72	91.65	37.60
OSPCoOp	94.49	30.72	95.06	27.39	92.17	33.86	91.20	41.25	93.23	33.31
Eight-shot										
CoOp [22]	92.90	32.85	92.11	37.55	89.14	44.76	89.58	43.91	90.94	39.77
LoCoOp [10]	94.09	27.98	92.86	36.58	89.82	43.91	89.78	45.43	91.64	38.47
OSPCoOp	94.68	28.41	95.29	24.68	92.45	31.51	91.15	39.53	93.40	31.04
Sixteen-shot										
CoOp [22]	90.63	47.42	90.96	44.5	88.3	51.85	89.02	46.63	89.73	47.6
LoCoOp [10]	94.98	24.42	92.73	34.76	90.54	38.95	91.5	39.49	92.44	34.41
OSPCoOp	95.91	22.44	95.8	24.05	93.26	28.72	91.74	39.47	94.18	28.67

Table 1. Results on ImageNet-1K as ID datasets with different few-shot settings. OOD score: MCM.

Method	iNaturalist		SUN		Places		Texture		Avg	
	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓
One-shot										
CoOp [22]	93.23	30.44	91.93	34.28	88.99	42.39	87.42	48.12	90.39	38.81
LoCoOp [10]	96.33	17.47	95.12	23.04	91.70	32.67	86.69	51.64	92.46	31.21
OSPCoOp	96.57	18.05	95.83	21.59	92.72	30.17	90.12	41.10	93.81	27.73
Two-shot										
CoOp [22]	95.89	16.04	93.01	29.13	90.06	37.91	87.83	46.08	91.70	32.29
LoCoOp [10]	95.83	19.22	94.98	20.61	91.63	33.85	89.27	45.21	92.93	29.72
OSPCoOp	95.58	23.81	95.83	21.10	92.52	30.76	90.10	43.93	93.51	29.90
Four-shot										
CoOp [22]	95.50	20.84	92.56	32.89	89.44	41.31	87.91	45.61	91.35	35.16
LoCoOp [10]	95.31	20.77	94.56	24.70	91.75	33.08	88.86	45.88	92.62	31.11
OSPCoOp	95.86	21.54	96.01	20.81	92.99	29.20	90.56	41.51	93.86	28.26
Eight-shot										
CoOp [22]	95.34	20.37	91.73	35.63	89.78	41.09	89.16	42.56	91.50	34.92
LoCoOp [10]	95.60	20.33	94.53	26.94	91.28	36.59	89.41	45.17	92.71	32.26
OSPCoOp	96.04	21.26	96.19	19.72	93.20	28.15	90.54	41.26	93.99	27.60
Sixteen-shot										
CoOp [22]	94.10	29.23	91.92	35.74	89.3	43.81	87.25	47.54	90.64	39.08
LoCoOp [10]	96.45	17.07	94.44	25.19	91.18	32.74	91.24	38.46	93.33	28.37
OSPCoOp	97.13	15.25	96.74	18.26	94.01	25.74	91.13	41.26	94.75	25.13

Table 2. Results on ImageNet-1K as ID datasets with different few-shot settings. OOD score: GL-MCM.

λ_{out}^{r}	λ_{out}^{g}	iNaturalist		SUN		Places		Texture		Avg	
		AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓
0.0	0.0	95.70	18.78	93.83	28.25	91.30	34.85	87.36	49.26	92.05	32.78
1.5	0.0	96.00	20.13	95.87	21.53	92.92	30.49	90.96	41.63	93.94	28.45
1.5	0.25	97.13	14.58	96.26	20.12	93.64	26.94	91.17	37.96	94.55	24.90
1.5	0.5	97.13	15.25	96.74	18.26	94.01	25.74	91.13	41.26	94.75	25.13
1.5	0.75	97.22	14.57	96.58	18.68	93.46	27.70	91.37	39.31	94.66	25.06
1.5	1.0	96.56	17.22	96.49	17.25	93.64	25.44	89.74	40.40	94.11	25.08
1.5	1.25	96.39	17.85	96.37	18.48	93.63	26.08	91.24	37.34	94.06	24.94
1.5	1.5	97.07	14.26	96.52	17.66	93.88	24.54	90.91	37.84	94.59	23.57
1.5	1.75	97.02	15.58	96.31	18.75	93.55	26.92	90.77	39.57	94.42	25.21
1.5	2.0	96.96	15.51	96.30	19.27	93.51	27.38	90.96	38.65	94.43	25.20
1.5	2.25	96.80	16.95	95.90	21.32	93.07	28.51	90.56	40.87	94.08	26.91
1.5	2.5	97.18	15.21	96.25	20.03	93.19	28.48	91.13	39.22	94.44	25.73
0.0	0.5	98.02	9.37	95.57	21.38	93.66	26.52	89.32	44.15	94.14	25.35
0.25	0.5	97.75	10.14	96.20	19.02	93.73	26.61	90.45	39.91	94.53	23.92
0.5	0.5	96.98	16.51	96.78	17.39	94.16	25.00	90.41	42.55	94.59	25.36
0.75	0.5	97.14	14.80	96.00	21.17	93.34	28.43	91.00	39.06	94.37	25.87
1.0	0.5	96.68	16.24	96.59	18.19	93.78	25.73	90.71	40.30	94.44	25.12
1.25	0.5	97.23	14.70	96.81	17.99	93.99	25.79	91.01	40.89	94.76	24.84
1.5	0.5	97.13	15.25	96.74	18.26	94.01	25.74	91.13	41.26	94.75	25.13
1.75	0.5	96.99	15.57	96.40	19.06	93.63	27.15	91.14	38.88	94.54	25.17
2.0	0.5	96.57	18.36	96.41	19.03	93.47	28.14	91.57	38.43	94.51	25.99
2.25	0.5	96.91	17.08	96.56	19.37	93.91	26.59	91.78	37.57	94.79	25.15
2.5	0.5	96.56	17.62	96.35	20.03	93.53	27.81	91.21	38.33	94.41	25.95

Table 3. Results on ImageNet with different loss weight. OOD score: GL-MCM.

OOD Aug	iNaturalist		Places		SUN		Texture		Avg.	
	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓
None	95.90	20.97	94.69	28.09	92.02	34.71	89.27	46.38	92.97	32.54
Rep	95.60	20.23	95.46	22.36	92.58	32.17	90.08	44.50	93.43	29.82
Bg	97.07	14.59	96.77	17.19	93.85	25.95	90.90	41.35	94.65	24.77
Rep+Bg	97.13	15.25	96.74	18.26	94.01	25.74	91.13	41.26	94.75	25.13

Table 4. Ablation of OOD Augmentation. 'Rep' stands for the data generated by repeating local ID regions, 'Bg' stands using the decoupled OOD content with ID regions inpainted. OOD score: GL-MCM.

3. Additional Results

Detailed results on fewer shots and OOD scores. Tab. 1 and 2 present detailed results on various few-shot settings and different OOD scores. These results show that our method consistently achieves the best performance across all settings, regardless of the OOD scores used. Notably, our method maintains high performance even in the 1-shot settings, achieving AUR scores of 93.25%, 93.81% with the MCM [9] and GL-MCM [11], OOD scores, respectively. Results show the robustness and effectiveness of OSPCoOp for different scenarios.

Detailed results on different loss weight. Tab. 3 presents detailed results from ablation experiments on the loss weights, compared to those reported in the main paper. When both λ_{out}^{g} and λ_{out}^{r} are set to 0, our method is equivalent to CoOp. Using $\mathcal{L}_{out}^{r}$ and $\mathcal{L}_{out}^{g}$ individually yields improvements of 1.89% and 2.09% in AUR, respectively. This demonstrates that applying pseudo-OOD supervision can significantly enhance the OOD ablility, whether optimizing from the perspective of global features or regional features. Furthermore, when both $\mathcal{L}_{out}^{g}$ and $\mathcal{L}_{out}^{r}$ are utilized, the final OOD performance improves further, regardless of the values of λ_{out}^{g} and λ_{out}^{r}. This indicates that integrating constraints from both global and regional features can lead to better ID/OOD decision boundaries. Results also suggest that the final performance remains stable and is not significantly affected by parameter variations.

Detailed results on OOD augmentation. Tab. 4 and 10 presents detailed results of the OOD augmentation. The

ID Aug	iNaturalist		SUN		Places		Texture		Avg	
	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓
One-shot										
None	96.57	18.05	95.83	21.59	92.72	30.17	90.12	41.10	93.81	27.73
Inpaint	96.85	16.13	96.31	18.86	93.22	28.27	89.37	43.34	93.94	26.65
Texture	96.73	16.72	96.32	19.25	93.47	27.48	89.72	43.56	94.07	26.76
Mix	96.36	18.93	96.15	20.70	93.20	29.12	89.96	42.31	93.92	27.77
Two-shot										
None	95.58	23.81	95.83	21.10	92.52	30.76	90.10	43.93	93.51	29.90
Inpaint	96.64	16.87	96.36	18.50	93.24	27.81	89.65	42.96	93.98	26.53
Texture	96.41	18.62	96.40	19.02	93.45	27.75	89.66	44.18	93.98	27.39
Mix	96.31	19.20	96.33	19.24	93.20	28.18	89.80	44.22	93.91	27.71
Four-shot										
None	95.86	21.54	96.01	20.81	92.99	29.20	90.56	41.51	93.86	28.26
Inpaint	96.20	19.10	96.44	18.66	93.29	28.23	90.30	41.58	94.06	26.90
Texture	95.93	19.60	96.35	18.23	93.29	27.25	89.87	42.22	93.86	26.82
Mix	96.59	17.24	96.42	19.07	93.36	27.72	90.26	41.57	94.16	26.40
Eight-shot										
None	96.04	21.26	96.19	19.72	93.20	28.15	90.54	41.26	93.99	27.60
Inpaint	95.79	21.09	96.42	18.84	93.43	27.41	90.00	42.37	93.91	27.43
Texture	95.98	19.39	96.29	18.95	93.36	27.09	90.35	39.88	94.00	26.33
Mix	96.47	17.41	96.59	17.41	93.53	26.54	90.68	40.54	94.32	25.48

Table 5. Results on ImageNet-1K as ID datasets with different ID augmentation. OOD score: GL-MCM.

Methods	LSUN-C		LSUN-R		Places		Texture		iSUN		Avg	
	AUR↑	FPR ↓	AUR↑	FPR ↓	AUR↑	FPR ↓	AUR↑	FPR ↓	AUR↑	FPR ↓	AUR↑	FPR ↓
GLMCM [11]	90.50	49.86	81.54	79.69	57.28	98.18	75.34	86.83	79.81	82.04	73.90	82.42
CoOp [22]	90.45	44.99	84.41	61.31	48.93	99.52	76.38	84.64	81.91	74.41	76.42	72.97
LoCoOp [10]	91.13	42.17	84.76	68.38	58.28	98.82	77.02	80.5	83.47	72.53	78.93	72.48
CLIP-OS [13] †	87.50	-	85.69	-	60.04	-	70.88	-	85.82	-	78.24	-
OSPCoOp	92.18	40.25	84.98	69.82	69.13	96.64	76.53	80.32	85.34	70.37	81.63	71.48

Table 6. 1-shots results on CIFAR-100 with ViT-B/16 and GL-MCM OOD score. Results marked with † are taken from [13]. 'AUR' stands for AUROC and 'FPR' stands for FPR95.

results indicate that OOD augmentation not only improves performance on the scene OOD dataset but also leads to enhancements in other datasets.

Detailed results on ID augmentation We replace the background with inpainted pictures or textured pictures of higher quality which are non-repetitive. Besides, We take many measures to ensure that the augmented samples added to training can improve the model's performance. Since the augmented images feature foregrounds and backgrounds that are completely unrelated and unnatural, we set the maximum proportion of augmented samples that can be used (e.g., 0.2). For the inpainted images, we set constraints to ensure that the background is not replaced with images from the same category. For images where the ID-relevant area occupies a small percentage of the picture, classification inevitably relies on ID-irrelevant background information. Therefore, we use a threshold (e.g., 0.5) for the ratio of the image covered by the mask to select the images used for augmentation. To ensure category balance, we implement a reintroduction strategy for categories with insufficiently augmented samples. For overly difficult augmented samples, we filter them based on their cosine similarity to the original samples in the visual embedding space. Those with a similarity lower than the threshold (e.g., 0.7) are excluded from training. Tab. 5 presents detailed ablation experiment results of the ID augmentation compared to those reported in the main paper. In most few-shot settings, the proposed augmentations can bring about performance improvements.

Detailed results on different region delineation methods Tab. 11 presents detailed results of the region delineation methods compared to those reported in the main paper.

Results on additional ID datasets. We evaluate our ap-

ID / OOD	IN-100 / IN-10		IN-10 / IN-20		IN-10 / IN-100		IN-20 / IN-10		Avg.	
	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓
LoCoOp [10] †	81.97	61.40	92.75	28.20	93.00	30.08	92.34	34.40	90.02	38.52
SCT [20] †	82.60	57.80	94.33	25.10	93.90	26.64	94.95	25.00	91.45	33.64
OSPCoOp	84.30	53.80	98.38	5.40	99.13	3.08	98.24	5.23	95.01	16.88

Table 7. Hard OOD results on ImageNet subsets with ViT-B/16 and GL-MCM OOD score. Results marked with † are taken from [20].

Methods	RESISC45		UC Merced Land Use		SUN		Avg.	
	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓
GLMCM [11]	77.92	71.30	76.13	63.33	83.17	59.03	79.07	64.55
CoOp [22]	90.87	40.41	97.37	17.58	91.77	33.36	93.34	30.45
LoCoOp [10]	91.19	42.22	96.69	19.70	93.89	26.07	93.92	29.33
OSPCoOp	92.59	38.82	97.46	13.94	96.48	19.52	95.51	24.09

Table 8. Results on RESISC45 with ViT-B/32 and GL-MCM OOD score.

proach on additional ID datasets, including the widely used CIFAR-100 [8] and a remote sensing dataset, NWPU-RESISC45 [1]. RESISC45 features landscape images as ID samples, where the distinction between foreground and background is often less clearly defined compared to other datasets. For CIFAR-100, we utilize LSUN [19], Places [21], Texture [2], and iSUN [17] as OOD datasets. For RESISC45, we designate 22 classes as ID samples and 23 classes as OOD samples. Additionally, we incorporate SUN and 10 non-overlapping classes from the UC Merced Land Use dataset [18] as OOD samples. Results for CIFAR-100 and RESISC45 are shown in Table 6 and Table 8. The results above demonstrate the strong generalization capability of OSPCoOp across diverse ID datasets.

Results on Hard-OOD Detection. Following MCM [9], we evaluate OSPCoOp in hard OOD scenarios. As shown in Table 7, we present hard-OOD detection results on subsets of ImageNet (IN). Additionally, the results in Table 8 for RESISC45 and UCM Land Use also fall under the category of hard OOD detection. In hard-OOD scenarios, the contribution of pseudo-OOD supervision becomes less significant; however, the mask-guided region regularization continues to enhance the model's ability to focus on id regions. These results demonstrate that OSPCoOp retains strong discriminative power for hard OOD detection.

Baseline performance with equivalent augmented data. Tab. 9 compares baselines and OSPCoOp on 1-shot scenario with the same data augmentation, where *OOD* and *ID* represent using corresponding augmentations. With pseudo-OOD supervision, all baselines significantly improve performance. Local-optimized methods (LoCoOp and OSPCoOp) further enhance performance with ID data augmentation. As ID-augmented images have little ID information in their backgrounds, these methods can focus more on the ID regions and better leverage these samples.

Computational cost. As shown in Tab. 12, we present a comprehensive comparison of GPU memory usage and running time between OSPCoOp and other approaches on ImageNet benchmark. For the running time evaluation, we compute each iteration's training time and evaluation time. While OSPCoOp involves additional computational resource consumption during the masking and inpainting stages, these requirements are substantially lower compared to those in the training stage. Compared to LoCoOp, OSPCoOp requires approximately 50% additional time during the training phase, but it brings significant performance improvements. During the inference phase, OSPCoOp does not incur additional computational overhead.

Ablation study on training modules Tab. 13 shows the results of inserting parameters at different locations in the CLIP architecture. The experimental results indicate that inserting parameters in the vision encoder in 16-shot settings tends to cause overfitting, resulting in a decrease in OOD detection performance.

4. Visualization

To better show the enhancement of attention to ID-relevant regions by our method, we visualize the responses of both ID and OOD samples separately. Fig. 3 shows the responses of our method, LoCoOp, and the pre-trained CLIP on the ImageNet-1K validation set. Our method focuses on more ID regions while reducing attention on background regions, making it more robust to background noise. Fig. 4, 5 present the responses of our method, LoCoOp, and pre-trained CLIP to the background regions. We visualize the top 25 regions with the highest responses, where darker colors indicate higher responses. It can be seen that our method has a lower response to the background regions.

ID Aug	iNaturalist		SUN		Places		Texture		Avg	
	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓
None										
CoOp [22]	93.23	30.44	91.93	34.28	88.99	42.39	87.42	48.12	90.39	38.81
LoCoOp [10]	96.33	17.47	95.12	23.04	91.70	32.67	86.69	51.64	92.46	31.21
OOD Aug										
CoOp [22]	96.06	19.21	94.53	26.18	92.15	32.98	88.08	47.06	92.70	31.35
LoCoOp [10]	95.53	22.69	95.56	22.98	92.56	30.74	88.80	47.93	93.11	31.08
OSPCoOp	96.57	18.05	95.83	21.59	92.72	30.17	90.12	41.10	93.81	27.73
OOD+ID Aug										
CoOp [22]	96.48	17.21	94.46	29.07	91.93	34.90	87.42	47.32	92.57	32.13
LoCoOp[10]	96.44	17.18	95.71	21.73	92.79	29.93	89.10	44.46	93.51	28.32
OSPCoOp	96.36	18.93	96.15	20.70	93.20	29.12	89.96	42.31	93.92	27.77

Table 9. 1-shot results on ImageNet-1K as ID datasets with different ID augmentation. OOD score: GL-MCM.

Filter Thre	iNaturalist		Places		SUN		Texture		Avg.	
	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓
10	97.03	14.78	96.06	20.16	93.09	28.75	89.99	43.55	94.04	26.81
8	97.19	14.98	96.71	18.17	93.78	26.74	90.83	39.54	94.63	24.86
7	96.87	15.80	96.05	21.34	93.33	28.28	91.22	38.42	94.37	25.96
6	96.91	16.25	95.98	22.49	93.29	28.72	91.48	39.79	94.42	26.81
5	97.21	14.08	96.49	19.36	93.65	26.88	91.37	40.53	94.68	25.21
4	97.09	15.11	96.25	20.16	93.50	27.66	91.05	42.62	94.47	26.39
3	96.87	15.69	96.66	17.62	93.73	26.25	90.67	40.37	94.51	24.98
2	97.26	14.11	96.15	19.62	93.35	28.04	90.59	40.78	94.34	25.64

Table 10. Results with different filter threshold. OOD score: GL-MCM.

Methods	iNaturalist		Places		SUN		Texture		Avg.	
	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓
Rank	97.11	15.04	96.56	19.41	93.56	28.23	90.52	42.98	94.44	26.41
Mask+Rank	97.17	15.57	96.26	19.98	93.70	28.24	90.97	39.80	94.52	25.90
Mask	97.13	15.25	96.74	18.26	94.01	25.74	91.13	41.26	94.75	25.13

Table 11. Comparison of different region delineation methods. 'RANK' refers to method of LoCoOp , 'Mask' denotes ours, and 'RANK+Mask' indicates using CLIP's modal feature similarity for those images without mask.

Methods	GPU memory usage			Running time			Performances	
	Masking	Inpainting	Training	Training(s)	Eval(s)	Training iterations	AUROC↑	FPR95 ↓
LoCoOp [22]	-	-	20.87G	0.487	0.491	500	93.52	28.66
SCT [20]	-	-	20.83G	0.489	0.491	500	93.37	26.47
OSPCoOp	7.52G	7.28G	20.89G	0.489	0.486	781	94.75	25.13

Table 12. Computational cost on ImageNet benchmark in the 16-shot settings with ViT-B/16. For training, we set the batch size to 32, while for evaluation, we set the batch size to 512.

Methods	iNaturalist		Places		SUN		Texture		Avg.	
	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓	AUR ↑	FPR ↓
Adapter [4]	95.49	26.47	91.83	38.41	92.86	33.08	92.12	32.85	93.08	32.70
TPT [22]+Adapter [4]	96.82	16.39	94.10	29.00	93.36	28.22	92.92	30.48	94.30	26.02
VPT[6]	95.54	25.13	93.36	33.44	91.36	36.56	87.06	49.01	91.83	36.03
MaPLe[7]	95.58	20.96	95.90	19.67	93.04	28.34	90.47	39.86	93.75	27.21
TPT [22]	97.13	15.25	96.74	18.26	94.01	25.74	91.13	41.26	94.75	25.13

Table 13. Ablation study on training parameters. 'TPT' refers to text prompt tuning, 'Adapter' refers to the insertion of an adapter after the Vision encoder, refining both the image features and region image features.

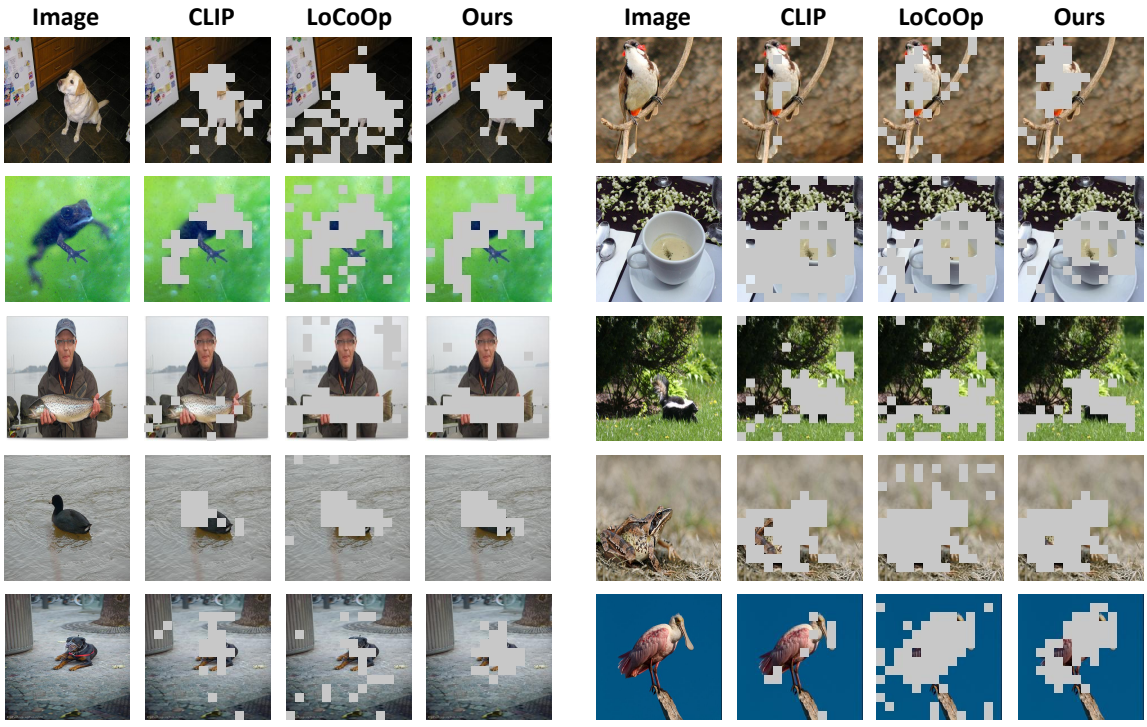

Figure 3. Visualization of responses on ID sample. The gray rectangular box highlights regions where the visual feature responses to ground truth rank in the top 200 across all labels.

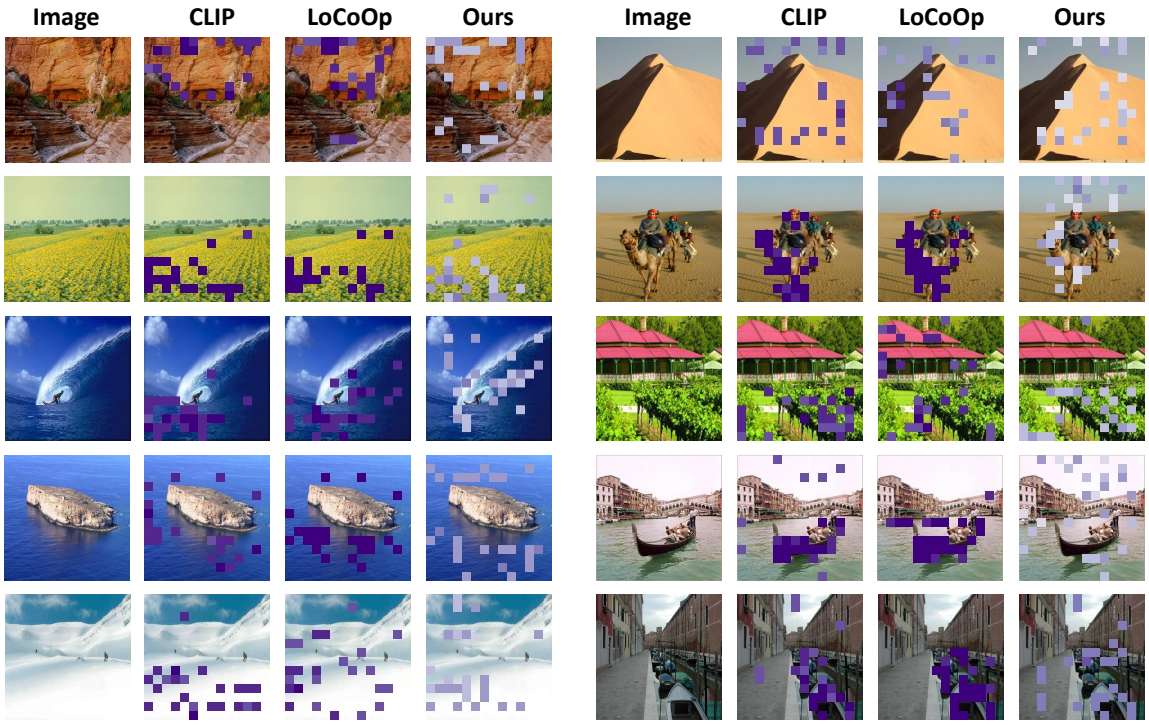

Figure 4. Visualization of responses on SUN, darker colors indicate higher logits for the region.

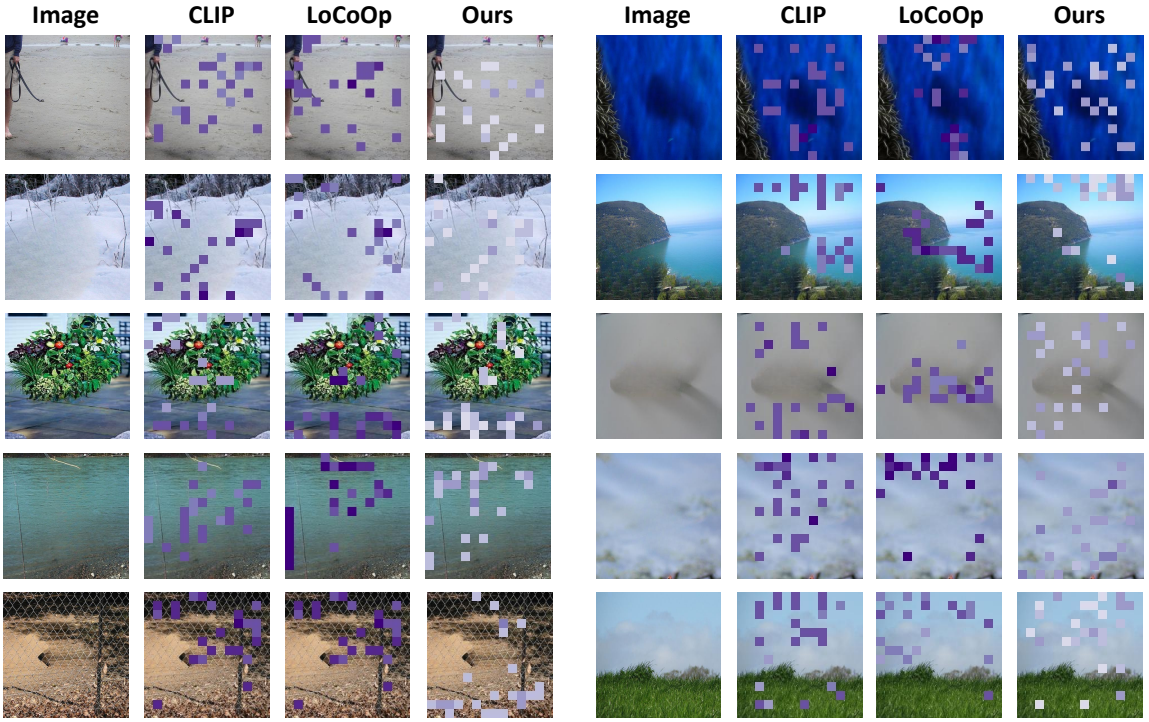

Figure 5. Visualization of responses on ImageNet-Bg, darker colors indicate higher logits for the region.

References

[1] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 6

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 1, 6

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitionn*, pages 248–255. Ieee, 2009. 1

[4] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 8

[5] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021. 1

[6] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 8

[7] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 8

[8] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[9] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in Neural Information Processing Systems*, 35:35087–35102, 2022. 4, 6

[10] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023. 3, 5, 6, 7

[11] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Gl-mcm: Global and local maximum concept matching for zero-shot out-of-distribution detection. *IJCV*, 2025. 4, 5, 6

[12] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 1

[13] Hao Sun, Rundong He, Zhongyi Han, Zhicong Lin, Yongshun Gong, and Yilong Yin. Clip-driven outliers synthesis for few-shot ood detection, 2024. 5

[14] 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. 1

[15] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018. 1

[16] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 1

[17] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015. 6

[18] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279, 2010. 6

[19] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 6

[20] Geng Yu, Jianing Zhu, Jiangchao Yao, and Bo Han. Self-calibrated tuning of vision-language models for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pages 56322–56348. Curran Associates, Inc., 2024. 6, 7

[21] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 1, 6

[22] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 3, 5, 6, 7, 8