Efficient and Versatile Robust Fine-Tuning of Zero-shot Models —Supplementary Material—

Sungyeon Kim[1], Boseung Jeong[1], Donghyun Kim[2], and Suha Kwak[1]

[1] Pohang University of Science and Technology (POSTECH), Korea
[2] Korea University, Korea
[1]{sungyeon.kim,boseung01,suha.kwak}@postech.ac.kr,[2]d_kim@korea.ac.kr
http://cvlab.postech.ac.kr/research/R-Adapter

A Implementation Details

Training. The details of training configurations for full-/16-shot image classification, cross-modal retrieval, and open-vocabulary segmentation are presented in Table 1. Moreover, Table 2 presents the detailed training configuration for image classification in base-to-novel generalization (C.1).

Image Augmentation. Following CLIP [29] and the previous work [11, 33], training images are randomly cropped to match the default pixel resolution of the model (e.g., 224×224 or 336×336), without employing additional data augmentation techniques. For testing, images are simply resized to default image sizes.

Text Templates. For image classification tasks, regardless of the dataset, we utilize the 80 text templates related to ImageNet as proposed in CLIP [29]. In the full-shot learning setting, during training, we randomly sample one of the text templates to construct the text following FLYP [11]. For few-shot learning, we primarily use the single text template, "`a photo of a {class}`", following CoOp [39] and CoCoOp [38]. During the evaluation, we construct the classifier weights by employing an ensemble of prompts generated from the 80 text templates to construct the classifier weights, following CLIP, WiSE-FT [33], and FLYP.

Open-Vocabulary Segmentation. By following [19], the original OVSeg model consists of two components. One is a mask proposal network *i.e.*, MaskFormer [6], and the other is the CLIP image and text encoders [29]. Specifically, the mask proposal network with Swin-B [21] as a backbone pre-trained on the COCO-Stuff dataset [3] produces several segmentation masks given an image input. Meanwhile, the CLIP image encoder is trained on the image masks from COCO Captions [5] in two stages, starting with full fine-tuning followed by mask prompt tuning, while freezing the CLIP text encoder. OVSeg employs a dataset composed of mask proposals from MaskFormer and their predictions, which is used for training. In contrast, we adopt a training strategy for OVSeg that is significantly different from the original training strategy. OVSeg with our method involves a **single-stage training process** focused solely on our adapters in

Table 1: Training configurations of various tasks.

Configuration	Classification (full-shot)	Classification (16-shot)	Cross-Modal Retrieval	Open-Vocabulary Segmentation
Source dataset	ImageNet-1K [8]	ImageNet-1K [8]	COCO [20]	COCO Captions [5]
Image encoder	3 CLIP ViTs (B/32, B/16, L/14@336px)	CLIP ViT-B/16	2 CLIP ViTs (B/16, L/14)	CLIP ViT-L/14
Batch size	512	256	512	256
Total epochs	10	50	10	5
Optimizer	AdamW [23]			
Scheduler	Cosine-annealing schedule [22]			
Warm-up step	500			
Initial learning rate	5e−4			
Drop Probability p	0.2			
Momentum m	0.999			
Temperature τ	0.01			
Margin δ	0.05			
Label Smoothing Noise ϵ	0.05	0	0	0
Re-scaling coefficient α	0.5	0.5	0.8	0.4

Table 2: Training configurations of image classification in base-to-novel generalization setting.

Configuration	Classification (Base-to-Novel)
Image encoder	CLIP ViT-B/16
Batch size	32
Total epochs	100
Optimizer	AdamW [23]
Scheduler	Cosine [22]
Warm-up step	500
Initial learning rate	5e−4
Drop Probability p	0.2
Momentum m	0.9
Temperature τ	0.01
Margin δ	0.05
Label Smoothing Noise ϵ	0
Re-scaling coefficient α	0.5

the CLIP model. We utilize the ground truth masks and categories from COCO Caption for training. This approach was initially suggested to result in performance degradation in the original OVSeg paper (as mentioned in Table 2 of OVSeg paper [19]). In our implementation, our method overcomes the issues they identified and achieves even higher performance. Interestingly, we found that the performance degraded when mask prompt tuning was used in conjunction with our method. For testing, the final class predictions are computed by an ensemble of the prediction of MaskFormer model and the prediction of CLIP model following the same setting of OVSeg. Specifically, when the prediction

weight of CLIP is denoted as x and the prediction weight of MaskFormer as y, the ensemble is expressed as $y^{(1-\lambda)} * x^{\lambda}$. For the ensemble value λ, we used 0.8 in A-847, 0.75 in PC-459, 0.8 in A-150, 0.5 in PC-59, and 0.25 in PAS-20.

B Datasets Details

Image Classification. We use ImageNet (IN) [8] as the ID dataset for fine-tuning; we evaluate the robustness of the models on five standard OOD datasets that represent five different types of OOD scenarios: ImageNetV2 (IN-V2) [30] is a new test set for ImageNet with distribution shift. ImageNet-R (IN-R) [14] consists of various artistic renditions (*e.g.*, painting, cartoons) of 200 ImageNet classes ImageNet-Sketch (IN-Sketch) [32] contains sketch images of 1000 ImageNet classes. ObjectNet [1] is a test set that contains images with 313 object classes collected from new viewpoints on new backgrounds, where 113 classes overlap with ImageNet. ImageNet-A (IN-A) [15] consists of natural images that are misclassified by a pre-trained ResNet-50 [12] for 200 ImageNet classes.

Cross-Modal Retrieval. We utilize two standard benchmarks for image-text cross-modal retrieval, COCO [20] as ID and Flickr30K [36] as OOD. For these two datasets, each image is associated with the corresponding five captions. Specifically, COCO is exploited as the ID dataset, and Flickr30K is utilized as the OOD dataset which has distribution shifts in both image and text modalities. In COCO, there are 123,287 images, and we follow the data split of [16] with 113,287 images for training, and 5,000 images for testing. Flickr30K contains 29,000 images for training and 1,000 images for testing.

Open-Vocabulary Segmentation. By following [19], we train the models on COCO Captions [5] and evaluate them on ADE20K [37], Pascal Context [26], and Pascal VOC [9] with 20 categories (PAS-20). Specifically, we exploit ADE20K in two versions, one with 150 frequently used categories (A-150) and the other with diverse 847 categories (A-847). Moreover, we also utilize Pascal Context in two versions, one with 59 frequently used categories (PC-59) and the other with the whole 459 categories (PC-459). Following our baseline method [19], we train a CLIP model on the COCO Captions dataset [5] and test them on several benchmarks as OOD: ADE20K [37], Pascal Context [26], and Pascal VOC [9].

Image Classification in Base-to-Novel Generalization. In our study of base-to-novel generalization for image classification, we employed 11 image recognition datasets as used in CoOp [39], encompassing a wide range of recognition tasks. The benchmark includes: ImageNet [8] and Caltech101 [10] for generic object classification; OxfordPets [28], StanfordCars [18], Flowers102 [27], Food101 [2], and FGVCAircraft [25] for fine-grained classification; SUN397 [35] for scene recognition; UCF101 [31] for action recognition; DTD [7] for texture classification; and EuroSAT [13] for satellite imagery recognition.

C Additional Experiments

C.1 Generalization From Base to Novel Classes

We conduct experiments to further emphasize generalizability by utilizing 11 datasets to measure the generalization performance in a base-to-novel setting following CoCoOp [38]. On each of the 11 datasets, we divide the classes into two equal groups: base classes and novel classes. All models are trained using only the base classes, with 16 samples per class, while evaluation is conducted on both base and novel classes separately to test generalizability.

As a default setting for training on few-shot datasets, we construct a text description of the target class employing a single text-template. We experiment with two settings varying bottleneck dimensions of R-Adapter: 4-rank and full-rank, with fewer and more parameters, respectively. We further explore the model when employing a full-rank structure and sampling templates from a predefined set of multiple text templates. The results are shown in Table 3.

Advantages in Performance. Our model with full-rank significantly outperformed the existing state of the art on most datasets by a large margin. Our method shows an average improvement of more than 1%p in base classes and over 0.7%p in novel classes compared to existing methods. Additionally, our model with 4-rank, which has a similar number of parameters as existing methods, performed better on new classes compared to our full-rank one, clearly achieving state-of-the-art performance. Overall, in terms of harmonic mean, our method achieves higher performance than existing methods, except for MaPLe [17]. Moreover, we found that using a set of multiple text-templates for sampling and training, instead of a single text-template, resulted in even greater performance gains. Consequently, this approach yields a 1.13%p improvement in the harmonic mean over the MaPLe, demonstrating the effectiveness of diversifying textual input during training.

Advantages in Efficiency. Our method easily adjusts to the required number of parameters by controlling the bottleneck dimension, without any added latency during inference. However, all existing baseline methods increase inference latency with added parameters since they involve adding input sequences. Especially, MaPLe, which achieved state-of-the-art performance, adds prompts to both text and visual encoders, significantly increasing its inference latency. Considering these factors, our method is highlighted for maintaining the same amount of computation as the original pre-trained model while achieving state-of-the-art performance.

C.2 Detailed Comparison to Parameter-Efficient Fine-Tuning

In this analysis, we conduct a detailed comparison among parameter-efficient fine-tuning (PEFT) methods, specifically focusing on LoRA, AdaptFormer, and RepAdapter. It's important to recall that R-Adapter utilizes a bottleneck module consisting of two matrices when the adapter rank is smaller than the hidden

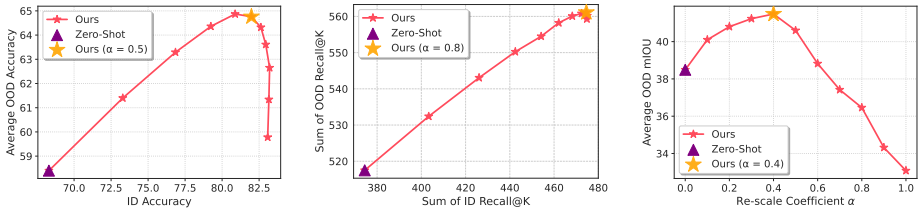
dimension of the backbone encoder. Conversely, R-Adapter with a full-rank employs a singular matrix due to the omission of non-linear layers, leveraging a multiplicative bottleneck structure. In our experiment, regardless of methods, all adapter modules are uniformly attached to both image and text encoders, ensuring fairness. However, the attachment locations and attachment manner differ among the approaches, leading to variations in the number of parameters even at the same rank.

We note that as the rank increases across all methods, there is a corresponding increase in the number of parameters, which significantly enhances performance in ID data. However, all existing methods show a decrease in OOD generalization performance as rank increases. In contrast, our method demonstrates robustness in OOD even at lower ranks and, unlike other methods, shows an improvement in OOD performance as the rank increases, creating a substantial gap in OOD performance between our method and existing approaches. Consequently, when using a similar number of parameters, our method not only outperforms existing PEFT methods in terms of performance but also ensures robustness irrespective of rank.

C.3 Additional Ablation Studies

Ablation Study on Re-scaling Coefficient. We investigate the impact of the re-scaling coefficient α in various tasks. The effect varies with each task and dataset, and as the distribution shift between in-distribution (ID) and out-of-distribution (OOD) data increases, performance improvement is noted when the re-scaling parameter value is smaller. In ImageNet classification, as analyzed in WiSE-FT [33], fixing the scaling parameter to 0.5 yields sufficiently high performance for both ID and OOD data, and tuning it can achieve even higher performance. In Cross-modal Retrieval, although the distribution gap between COCO and Flickr30K is not very large, a continuous increase is observed as the scaling parameter increases. However, performance improvement is still noted compared to when scaling is not applied. In open vocabulary segmentation, we observe that the mIOU performance generally improves as the coefficient moderately increases, but it tends to decrease again when the coefficient becomes too large.

Ablation Study on Label Smoothing Coefficient. We conducted an ablation study on the label smoothing coefficient ϵ, which is not included in the main text of the paper due to space limitations. The results of experiments on ImageNet using ViT-B/32 are presented in Table 6. We observe that increasing the label smoothing parameter up to 0.05 leads to performance improvements in both In-Distribution (ID) and Out-of-Distribution (OOD) settings. However, we also notice that label smoothing does not always benefit all tasks. While there is a clear performance improvement in the full-shot setting of ImageNet classification, in cases with fewer samples like the few-shot setting, or in settings other than classification, even a weak label smoothing noise can deteriorate performance. Our proposed loss, MPM-NCE, can consider multiple positive samples



(a) ImageNet Classification (ViT-B/16) **(b)** Cross-modal Retrieval (ViT-B/16) **(c)** Open Vocabulary Segmentation (ViT-L/14)

Fig. 1: Performance of our method varying re-scaling coefficient α in Eq. 9. The accuracy of each Cross-modal Retrieval is the sum of the performances in recall@K for Image retrieval (R@1, R@5, R@10) and the performances in recall@K for text retrieval (R@1, R@5, R@10). The accuracy of open vocabulary segmentation is the average of mIOU of 5 standard datasets.

and also easily apply traditional regularization techniques like label smoothing, and thus get benefit from them.

D Training Time Comparison

We compare and discuss the training latency of our method with the existing state-of-the-art method, Mask-Fill. The training latency for Mask-Fill is **8.44ms per image**, whereas, for our method, it is only **1.82ms per image**, tested on 64 batches with 3090 GPU. The training latency for Mask-Fill is computed using its official implementation[1]. The reasons for the increased latency during training time and discussion comparing with our method are as follows:

Mask-Fill enhances robustness by using masked images as counterfactual samples, which helps improve the robustness of the fine-tuning model. It generates masked images and then distills the information for the masked parts from a pre-trained model. This process involves extra computation time for creating masks and generating new images by combining different images. Moreover, for distillation, two images need to be forwarded by the training model, and one of them is forwarded by a pre-trained model during each iteration. Consequently, this training method results in longer time consumption compared to conventional fine-tuning methods. In contrast, our method avoids such complex processing and learns fewer parameters, enabling faster training speeds. This experiment demonstrates that our method not only surpasses the existing state-of-the-art method in performance but is also superior in terms of training time.

[1] https://github.com/Coxy7/robust-finetuning

Table 3: Comparison with fine-tuned methods from CLIP in base-to-novel generalization. All methods are trained from the base classes (16 shots). HM denotes Harmonic mean [34] which emphasizes the generalization trade-off. Superscripts denote the rank of adapter modules. "MT" represents that text description of the target class is constructed by sampling from a set of multiple predefined templates as used in FLYP [11].

	Base	Novel	HM
CLIP [29]	69.34	74.22	71.70
CoOp [39]	82.63	67.99	74.60
CoCoOp [38]	80.47	71.69	75.73
KgCoOp [39]	80.73	73.60	77.00
MaPLe [17]	82.28	75.14	78.55
Ours^{4}	80.06	**76.27**	78.11
Ours^{8}	81.74	76.45	79.01
Ours^{16}	82.34	76.25	79.18
Ours^{32}	83.00	76.16	79.43
Ours$^{\text{Full}}$	83.21	75.82	79.34
Ours$^{\text{Full}}$ (MT)	**83.64**	76.08	**79.68**

(a) Average over 11 datasets

	Base	Novel	HM
CLIP [29]	72.43	68.14	70.22
CoOp [39]	76.46	66.31	71.02
CoCoOp [38]	75.98	70.43	73.10
KgCoOp [39]	75.73	69.96	72.78
MaPLe [17]	76.66	70.54	73.47
Ours^{4}	76.38	71.38	73.87
Ours^{8}	76.39	71.81	74.03
Ours^{16}	76.38	71.58	73.90
Ours^{32}	76.76	71.64	74.11
Ours$^{\text{Full}}$	77.57	71.58	74.46
Ours$^{\text{Full}}$ (MT)	**77.74**	**71.70**	**74.60**

(b) ImageNet

	Base	Novel	HM
CLIP [29]	96.84	94.00	94.50
CoOp [39]	98.11	93.52	95.76
CoCoOp [38]	97.96	93.81	95.84
KgCoOp [39]	97.72	94.39	96.03
MaPLe [17]	97.74	94.36	96.02
Ours^{4}	97.74	95.85	96.79
Ours^{8}	98.44	96.02	97.22
Ours^{16}	98.21	96.19	97.19
Ours^{32}	98.67	95.77	97.20
Ours$^{\text{Full}}$	**98.83**	95.67	97.23
Ours$^{\text{Full}}$ (MT)	98.21	**96.36**	**97.28**

(c) Caltech101

	Base	Novel	HM
CLIP [29]	91.17	97.26	94.12
CoOp [39]	94.24	96.66	95.43
CoCoOp [38]	95.20	97.69	96.43
KgCoOp [39]	94.65	**97.76**	96.18
MaPLe [17]	95.43	**97.76**	96.58
Ours^{4}	93.73	97.71	95.68
Ours^{8}	95.75	96.92	96.33
Ours^{16}	95.96	98.04	96.99
Ours^{32}	95.91	97.54	96.72
Ours$^{\text{Full}}$	**95.80**	97.37	96.58
Ours$^{\text{Full}}$ (MT)	96.65	97.48	**97.07**

(d) OxfordPets

	Base	Novel	HM
CLIP [29]	63.37	74.89	68.65
CoOp [39]	76.20	60.40	72.49
CoCoOp [38]	70.49	73.59	72.01
KgCoOp [39]	71.76	75.04	73.36
MaPLe [17]	72.94	74.00	73.47
Ours^{4}	79.11	74.85	76.92
Ours^{8}	78.74	75.58	77.12
Ours^{16}	77.87	75.05	76.44
Ours^{32}	78.91	75.88	77.36
Ours$^{\text{Full}}$	81.24	**75.98**	**78.52**
Ours$^{\text{Full}}$ (MT)	**81.88**	74.15	77.82

(e) StanfordCars

	Base	Novel	HM
CLIP [29]	72.08	77.80	74.83
CoOp [39]	97.63	69.55	81.23
CoCoOp [38]	94.87	71.75	81.71
KgCoOp [39]	95.00	**74.73**	**83.65**
MaPLe [17]	**95.92**	72.46	82.56
Ours^{4}	87.34	74.03	80.14
Ours^{8}	91.42	72.63	80.95
Ours^{16}	91.33	73.79	81.63
Ours^{32}	92.86	74.34	82.57
Ours$^{\text{Full}}$	90.09	73.25	81.16
Ours$^{\text{Full}}$ (MT)	95.07	73.56	82.94

(f) Flowers102

	Base	Novel	HM
CLIP [29]	90.10	91.22	90.66
CoOp [39]	89.44	87.50	88.46
CoCoOp [38]	90.70	91.29	90.99
KgCoOp [39]	90.50	91.70	91.90
MaPLe [17]	**90.71**	**92.05**	**91.38**
Ours^{4}	90.28	90.79	90.54
Ours^{8}	90.50	91.31	90.9
Ours^{16}	90.58	91.33	90.95
Ours^{32}	90.55	91.42	90.98
Ours$^{\text{Full}}$	90.29	90.05	90.17
Ours$^{\text{Full}}$ (MT)	90.46	91.33	90.89

(g) Food101

	Base	Novel	HM
CLIP [29]	27.19	36.29	31.09
CoOp [39]	39.24	30.49	34.30
CoCoOp [38]	33.41	23.71	27.74
KgCoOp [39]	36.21	35.55	34.83
MaPLe [17]	37.44	35.61	36.50
Ours^{4}	36.01	**37.07**	36.54
Ours^{8}	35.71	37.55	36.61
Ours^{16}	39.20	36.71	37.91
Ours^{32}	39.56	35.33	37.33
Ours$^{\text{Full}}$	**41.48**	36.17	**38.64**
Ours$^{\text{Full}}$ (MT)	40.04	35.73	37.77

(h) FGVCAircraft

	Base	Novel	HM
CLIP [29]	69.36	75.35	72.23
CoOp [39]	80.85	68.34	74.07
CoCoOp [38]	79.74	76.86	78.27
KgCoOp [39]	80.29	76.53	78.36
MaPLe [17]	80.82	**78.70**	79.75
Ours^{4}	80.42	78.43	79.42
Ours^{8}	81.41	78.63	79.99
Ours^{16}	81.76	77.98	79.82
Ours^{32}	82.08	78.24	80.12
Ours$^{\text{Full}}$	81.38	78.06	79.68
Ours$^{\text{Full}}$ (MT)	**82.70**	78.36	**80.48**

(i) SUN397

	Base	Novel	HM
CLIP [29]	53.24	59.90	56.37
CoOp [39]	80.17	47.54	59.68
CoCoOp [38]	77.01	56.00	64.85
KgCoOp [39]	77.55	54.99	64.35
MaPLe [17]	80.36	59.18	68.16
Ours^{4}	73.15	**66.43**	69.62
Ours^{8}	77.43	66.91	71.79
Ours^{16}	79.05	63.89	70.67
Ours^{32}	79.86	64.37	71.28
Ours$^{\text{Full}}$	**83.45**	64.13	72.53
Ours$^{\text{Full}}$ (MT)	83.33	64.62	**72.79**

(j) DTD

	Base	Novel	HM
CLIP [29]	56.48	64.05	60.03
CoOp [39]	91.54	54.44	68.27
CoCoOp [38]	87.49	60.04	71.21
KgCoOp [39]	85.64	64.34	73.48
MaPLe [17]	**94.07**	73.23	**82.35**
Ours^{4}	84.88	**75.85**	80.11
Ours^{8}	90.33	76.03	82.56
Ours^{16}	90.74	75.54	82.44
Ours^{32}	92.74	74.79	82.81
Ours$^{\text{Full}}$	90.14	74.26	81.43
Ours$^{\text{Full}}$ (MT)	88.74	74.92	81.25

(k) EuroSAT

	Base	Novel	HM
CLIP [29]	70.53	77.50	73.85
CoOp [39]	85.14	64.47	73.37
CoCoOp [38]	82.33	73.45	77.64
KgCoOp [39]	82.89	76.67	79.65
MaPLe [17]	83.00	**78.66**	80.77
Ours^{4}	81.64	76.58	79.03
Ours^{8}	83.04	77.61	80.23
Ours^{16}	84.64	78.69	81.56
Ours^{32}	85.06	78.47	81.63
Ours$^{\text{Full}}$	85.06	77.45	81.07
Ours$^{\text{Full}}$ (MT)	**85.21**	78.64	**81.79**

(l) UCF101

Table 4: Harmonic mean accuracy on base and novel classes. All methods are fine-tuned with 16 shots per base class.

Methods	Param	Avg	IN	Cal	Pets	Cars	Flo	Food	Air	SUN	DTD	Euro	UCF
MaPLE	3.55 M	78.6	73.5	96.0	96.6	73.5	82.6	91.4	36.5	79.8	68.2	82.4	80.8
Ours^{4}	0.25 M	78.1	73.9	96.8	95.7	76.9	80.1	90.5	36.5	79.4	69.6	80.1	79.0
Ours^{8}	0.49 M	79.0	74.0	**97.2**	96.3	77.1	81.0	90.9	36.6	80.0	**71.8**	82.6	80.2
Ours^{16}	0.98 M	79.2	73.9	**97.2**	**97.0**	76.4	81.6	81.0	**37.9**	79.8	70.7	82.4	81.5
Ours^{32}	1.97 M	**79.4**	**74.1**	**97.2**	96.7	**77.4**	**82.6**	**91.0**	37.3	**80.1**	71.3	**82.8**	**81.6**

Table 5: Top-1 accuracy of parameter-efficient fine-tuning methods on ImageNet (ID) and OOD datasets with ViT-B/32. Superscripts denote the rank of adapter or LoRA.

	Trainable	**ID**	**Out-Of-Distribution (OOD)**					
Methods	**Params (M)**	IN	OOD avg.	IN-V2	IN-R	IN-Sketch	ObjectNet	IN-A
AdaptFormer^{16} [4]	0.5	**74.7**	48.9	64.3	63.8	41.7	45.5	29.3
RepAdapter^{16} [24]	1.0	74.3	49.7	64.4	65.1	42.4	46.0	30.4
Ours^{16}	1.0	74.5	**52.5**	**65.1**	**69.5**	**45.8**	**47.9**	**34.0**
AdaptFormer^{128} [4]	3.9	75.6	48.3	64.5	61.7	41.0	45.0	29.3
RepAdapter^{128} [24]	7.8	76.3	48.9	65.2	62.7	41.9	45.7	29.2
Ours^{128}	7.8	**76.7**	**53.7**	**66.9**	**70.2**	**47.1**	**48.7**	**35.5**
LoRA$^{\text{Full}}$	163.6	**78.0**	48.2	66.2	60.0	42.3	45.0	27.4
AdaptFormer$^{\text{Full}}$ [4]	20.5	77.2	48.5	66.4	60.6	42.2	45.2	28.0
RepAdapter$^{\text{Full}}$ [24]	41.0	76.9	47.7	65.5	60.1	41.3	44.2	27.6
Ours$^{\text{Full}}$	20.5	77.7	**54.3**	**67.7**	**70.8**	**47.8**	**49.7**	**35.6**

Table 6: Ablation study on label smoothing coefficient ϵ in Eq. 10.

Label Smoothing Noise ϵ	ID	OOD
0	77.3	53.9
0.01	77.5	54.1
0.03	77.5	54.2
0.05	**77.7**	**54.3**

References

1. Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., Katz, B.: Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. Advances in neural information processing systems **32** (2019) 3
2. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: Proc. European Conference on Computer Vision (ECCV) (2014) 3
3. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 1
4. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. In: Proc. Neural Information Processing Systems (NeurIPS) (2022) 8
5. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015) 1, 2, 3
6. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems (2021) 1
7. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014) 3
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009) 2, 3
9. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision (IJCV) (2010) 3
10. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: 2004 conference on computer vision and pattern recognition workshop (2004) 3
11. Goyal, S., Kumar, A., Garg, S., Kolter, Z., Raghunathan, A.: Finetune like you pretrain: Improved finetuning of zero-shot vision models. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 1, 7
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016) 3
13. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2019) 3
14. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proc. IEEE International Conference on Computer Vision (ICCV) (2021) 3
15. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 3

16. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 3
17. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 4, 7
18. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 554–561 (2013) 3
19. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 1, 2, 3
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Proc. European Conference on Computer Vision (ECCV) (2014) 2, 3
21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proc. IEEE International Conference on Computer Vision (ICCV) (2021) 1
22. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016) 2
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proc. International Conference on Learning Representations (ICLR) (2019) 2
24. Luo, G., Huang, M., Zhou, Y., Sun, X., Jiang, G., Wang, Z., Ji, R.: Towards efficient visual adaption via structural re-parameterization. arXiv preprint arXiv:2302.08106 (2023) 8
25. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013) 3
26. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014) 3
27. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian conference on computer vision, graphics & image processing (2008) 3
28. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012) 3
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proc. International Conference on Machine Learning (ICML) (2021) 1, 7
30. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: Proc. International Conference on Machine Learning (ICML). PMLR (2019) 3
31. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012) 3
32. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. Advances in Neural Information Processing Systems **32** (2019) 3

33. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al.: Robust fine-tuning of zero-shot models. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 1, 5
34. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning-the good, the bad and the ugly. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 7
35. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE computer society conference on computer vision and pattern recognition (2010) 3
36. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics (2014) 3
37. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision (IJCV) (2019) 3
38. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 1, 4, 7
39. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision (IJCV) (2022) 1, 3, 7