A Pre-training Details

We conduct prompt pre-training on the ImageNet-21K dataset (official winter 2021 released version[1]). We follow the processing methods in [23], which involves cleaning invalid classes, allocating 50 images per class for a validation split, and crop-resizing all the images to 224 resolution. We conduct all the experiments on $8\times$Nvidia V100 GPUs. For pre-training, the learnable vector is randomly initialized by drawing from a zero-mean Gaussian distribution with a standard deviation equal to 0.02. We use the SGD optimizer with an initial learning rate of 0.002, decayed by the cosine annealing rule. The batch size is 32, and the maximum epoch is 20.

For the mask proposal network and region proposal network pre-training, we strictly follow the settings of ZSSeg [29] and Detic [32], respectively. Specifically, we take MaskFormer [3] with ResNet-101 [10] as the mask proposal network. We use an AdamW optimizer with the initial learning rate of $1e$-4, weight decay of 1e-4, a backbone multiplier of 0.1, and a poly learning rate policy with a power of 0.9. Besides, we take CenterNet2 [33] detector with ImageNet-21k pre-trained ResNet-50 [23] as the region proposal network. We use an Adam optimizer with learning rate $2e$-4. Other tricks like Federated Loss, repeat factor sampling, and large scale jittering are incorporated to further improve the performance. As with Detic, we leverage both region-level and image-level supervision. We always first train a converged base-class-only model ($4\times$ schedule) and fine-tune it with additional image-labeled data for another $4\times$ schedule.

B Setting for Segmentation and Detection

Table 1 outlines the settings for semantic segmentation and object detection. We further introduce the settings in detail from three perspectives: backbone, data processing, and prompt.

Backbone. In general, we adopt a **two-stage framework** for these two tasks. At stage one, we use a pre-trained proposal network to generate a set of mask or region proposals. At stage two, we classify each proposal with the class features generated by our POMP prompt. For semantic segmentation, our POMP shares the same visual backbone as ZSSeg [29], which uses a pre-trained MaskFormer [3] with ResNet-101 [10] as default backbone to extract a set of binary masks. For object detection, our POMP shares the same visual backbone with Detic [32], which takes CenterNet2 [33] detector with ResNet-50 as its backbone, and leverages both region-level and image-level supervision.

Data Processing. We follow previous work [27, 29, 8, 32] to designate data belonging to two class sets as **source data** and **target data**, respectively. The proposal networks are pre-trained on the source data with the source class set, while conducting zero-shot evaluation on the target data with the target class set. There are two protocols for the source-target data split. The first is the **open-vocabulary protocol**, where the class set of one dataset is divided into two disjoint groups for the source and target data, respectively. The second protocol is the **cross-dataset protocol**, in which the source and target data are from two independent datasets with potentially overlapping class sets.

We introduce the details of class set splitting in the open-vocabulary protocol. COCO Stuff and Pascal VOC 2012 are the two semantic segmentation datasets using the open-vocabulary protocol. Following previous settings [27, 29], a total of 171 annotated classes in COCO Stuff are divided into 156 seen classes and 15 unseen classes. For Pascal VOC 2012, a total of 20 classes are divided into 15 seen classes and 5 unseen classes, and the provided augmented annotations are used. LVIS is the object detection dataset using the open-vocabulary protocol. The standard LVIS dataset contains object detection and instance segmentation labels for 1203 classes. The classes are divided into three groups: frequent, common, and rare, based on the number of training images. According to previous work [8], the data from the 866 frequent and common classes are considered the source data, while those from the remaining 337 rare classes are the target data in testing.

Prompt. ZSSeg provides two kinds of prompts: hand-crafted prompts and learning-based prompts. Hand-crafted prompts include *single prompt*, i.e., "`a sculpture of a [CLASSNAME]`", as well as *ImageNet prompts* [21] and *ViLD prompts* [8], which are used for prompt ensemble and consist of 80 and 14 hard prompts, respectively. The learning-based prompt is obtained by fine-tuning a randomly

[1] `https://image-net.org/`

Table 1: Settings for semantic segmentation and object detection.

Task	Proposal Network	Setting	Source Data and Class Set (for proposal network pre-training)	Target Data and Class Set (for zero-shot evaluation)
Semantic Segmentation	MaskFormer (Mask Proposal Network)	Open-vocab COCO Stuff	COCO Stuff (seen)	COCO Stuff (unseen)
		Open-vocab PASCAL VOC	PASCAL VOC (seen)	PASCAL VOC (unseen)
		Cross-dataset	COCO Stuff	ADE20K / PASCAL Context
Object Detection	CenterNet2 detector (Region Proposal Network)	Open-vocab LVIS	LVIS (frequent+common)	LVIS (rare)
		Cross-dataset	LVIS	COCO / Object365

Table 2: Datasets in our experiments.

Dataset	Classes	Train Size	Test Size	Metric
Datasets of Image Classification				
Caltech-101 [7]	102	3,060	6,086	mean per-class accuracy
Oxford-IIIT Pets [20]	37	3,680	3,669	mean per-class accuracy
Stanford Cars [15]	196	8,144	8,041	accuracy
Oxford Flowers-102 [19]	102	2,040	6,149	mean per-class accuracy
Food-101 [1]	101	75,750	25,250	accuracy
FGVC Aircraft [17]	100	6,667	3,333	mean per-class accuracy
SUN-397 [28]	397	15,880	19,850	accuracy
Describable Textures (DTD) [4]	47	3,760	1,880	accuracy
EuroSAT [11]	10	10,000	5,000	accuracy
UCF-101 [25]	101	7,639	3,783	accuracy
ImageNetV2 [22]	1,000	10,000	10,000	accuracy
ImageNet-S [26]	1,000	50,889	50,889	accuracy
ImageNet-A [13]	200	7,500	7,500	accuracy
ImageNet-R [12]	200	30,000	30,000	accuracy
Datasets of Semantic Segmentation				
COCO Stuff [2]	171	117K	5K	mIoU (seen/unseen), hIoU
PASCAL VOC [6]	20	11,185	1,449	mIoU (seen/unseen), hIoU
ADE20K [30]	150	20K	3K	mIoU, fwIoU, pACC
PASCAL Context [18]	59	10,103	9,637	mIoU, fwIoU, pACC
Datasets of Object Detection				
LVIS [9]	1,203	100,170	19,822	AP_r, AP_c, AP_f, AP
COCO [16]	80	118K	5K	AP, AP_{50}, AP_{75}, AP_s, AP_m, AP_l
Object365 [24]	365	600K	38K	AP, AP_{50}, AP_{75}, AP_s, AP_m, AP_l

initialized soft prompt on the source data. Accordingly, for a fair comparison, we conducted two sets of experiments based on whether to use the source data for prompt fine-tuning. (1) The results of ZSSeg with various hard-crafted prompts and the pre-trained POMP prompt without access to the source data can be found in Table 4 in Appendix E.3. (2) The results of ZSSeg with learning-based prompts initialized from random vectors and our pre-trained POMP prompt, both using source data for further fine-tuning, can be found in Table 4 and Table 5 in § 4.3.2. Detic has also extensively delved into intricate prompts, such as "`a photo of a [CLASS] in the scene`". Moreover, it has made endeavors to employ synonyms for each category. Nevertheless, its ultimate recommendation is to use a simple yet effective prompt, i.e., "`a [CLASSNAME]`", and all its released checkpoints are based on this prompt. We strictly adhere to Detic's best practice, the evaluation of Detic and POMP in § 4.3.3 are both conducted without any further prompt tuning on the source data.

C Datasets

The details of the downstream datasets for image classification, semantic segmentation, and object detection are shown in Table 2.

Image Classification. For cross-dataset image classification, we evaluate the performance of POMP on 10 downstream datasets, including Caltech-101 [7], Oxford-Pets [20], Stanford Cars [15], Oxford-Flowers102 [19], Food-101 [1], FGVC Aircraft [17], EuroSAT [11], SUN-397 [28], Describable

Table 3: Ablation on the sampling distribution in POMP based on CLIP (ViT/B-16) backbone.

Method	ImageNet-21K	Cross-dataset (10 Avg.)	Cross-domain (4 Avg.)
POMP (uniform distribution)	25.3	67.0	60.8
POMP (frequency distribution)	24.9 **(-0.4)**	66.2 **(-0.8)**	60.1 **(-0.7)**
POMP (similarity distribution)	23.6 **(-1.7)**	64.2 **(-2.8)**	59.2 **(-1.6)**

Textures (DTD) [4], UCF-101 [25]. We also conduct zero-shot evaluation on 4 out-of-domain datasets derived from ImageNet [5], including ImageNetV2 [22], ImageNet-S [26], ImageNet-A [13], and ImageNet-R [12], to evaluate the domain generalization capability of our method.

Semantic Segmentation. We perform open-vocab semantic segmentation on COCO Stuff [2] and Pascal VOC 2012 [6]. Following previous notation and settings [27, 29], we split the class set into seen and unseen classes, where data for seen classes is considered the source data and data for unseen classes is considered the target data. The major measures for evaluation include mIoU and the harmonic mean IoU (hIoU) among both seen and unseen classes [29]. The hIoU is defined as:

$$\text{hIoU} = \frac{2 \times \text{mIoU}_{\text{seen}} \times \text{mIoU}_{\text{unseen}}}{\text{mIoU}_{\text{seen}} + \text{mIoU}_{\text{unseen}}}$$

We also conduct cross-dataset evaluation, which takes the standard COCO Stuff dataset as the source dataset for pre-training a mask proposal network, and then conducts zero-shot inference on ADE20K [30] and PASCAL Context [18].

Object Detection. We evaluate the performance of POMP on the object detection dataset LVIS [9] under the open-vocabulary setting proposed by [8]. The data from the 866 frequent and common classes are considered the source data, while those from the remaining 337 rare classes are the target data in testing. We take AP_r, i.e., AP on rare classes, as the major measure. AP_f and AP_c, i.e., AP on frequent and common classes, are also reported. In the cross-dataset setting, the region proposal network is pre-trained on the source dataset of standard LVIS, and then directly conducts inference on two target datasets, including COCO [16] and Object365 [24]. We use AP, AP_{50}, AP_{75}, AP_s, AP_m, and AP_l the evaluation metrics.

D Qualitative Results for Semantic Segmentation and Object Detection

In this section, we provide more qualitative results of our POMP for semantic segmentation and object detection. Figure 1 shows another three cases on open-vocabulary COCO-Stuff segmentation. POMP demonstrates a stronger ability than ZSSeg in the recognition of background classes. In case (1), POMP correctly identified the *dirt* and *plant-other* in the scene, instead of marking all these areas as *grass*. In case (2) and (3), POMP recognizes the classes of *clouds* and *tree*, respectively, while ZSSeg misclassifies them as *sky-other* and *bush*. However, POMP misses some objects of *sheep* located at the edge in case (2) and neglects the object of *branch* in case (3), indicating it still has insufficient recognition of small objects. For object detection, Figure 2 illustrates qualitative results on LVIS images. Base and novel categories are shown in purple and green, respectively. POMP identifies regions from the novel class without using the corresponding 1.2K detection annotations, demonstrating its generalization in the wild.

E More Ablation Study

E.1 Ablation on Proposal Distribution

As introduced in § 3.2, we also investigate other types of proposal distribution for local contrast and negative class sampling. The first is the frequency distribution $Q^{(f)}$, which samples the negative class i based on the number of training samples belonging to this class. Note that the original ImageNet-21K is class-imbalanced, i.e., the number of training samples belonging to common classes is larger than those belonging to rare classes, which can roughly reflect the long-tail distribution of

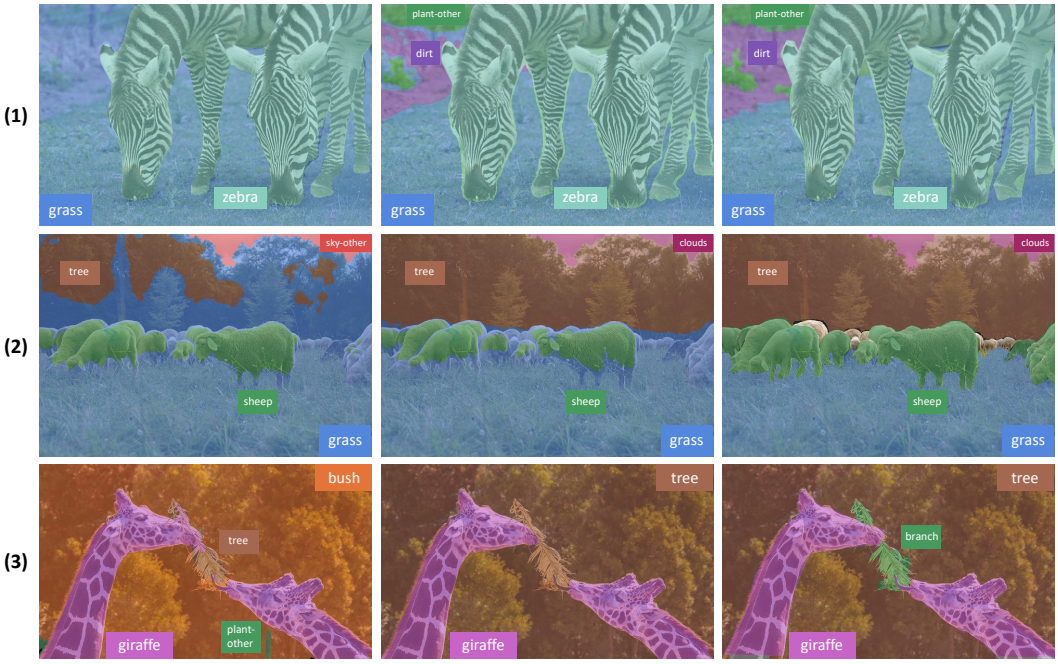

Figure 1: More qualitative results on open-vocabulary COCO-Stuff segmentation.

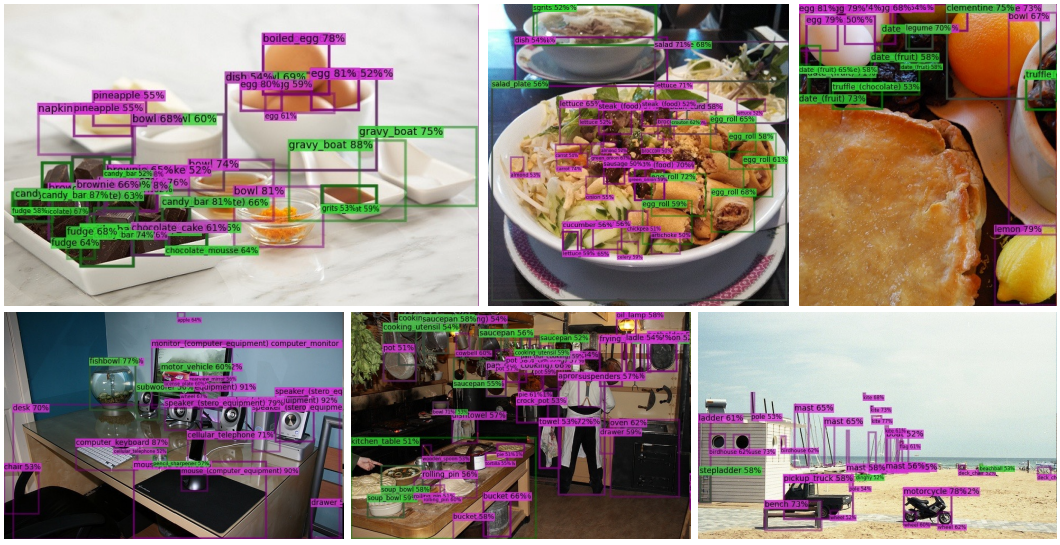

Figure 2: Qualitative results on LVIS images. Base and novel categories are shown in **purple** and **green** colors respectively. We use a score threshold of 0.5 and show the most confident class for each box.

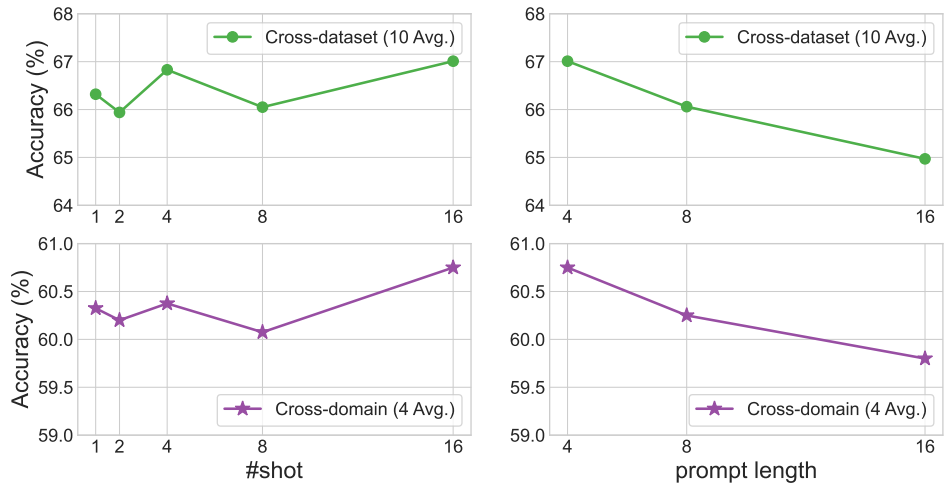

Figure 3: Ablation study on #shot and prompt length. When varying #shot, the prompt length is 4, and when varying the prompt length, #shot is 16.

object categories in nature. The frequency distribution will allow for more sampling of common classes while suppressing the exposure of rare classes in prompt tuning. Let M_i be the number of training samples belonging to the negative class i, the frequency distribution is defined as:

$$Q_i^{(f)} = \frac{M_i}{\sum_{j=1}^{N} M_j}. \tag{1}$$

The second is the similarity distribution $Q^{(s)}$, which aims to sample more hard negative classes. Hard negative classes are those that have a higher similarity between their features and the features of the input images, and are more likely to be confused with the positive class. Accordingly, in the similarity distribution, the likelihood of a negative class being sampled increases as the similarity between its feature and the image feature increases. To achieve this, we pre-encode features of all classes represented by a hand-crafted prompt (i.e., "`a photo of a [CLASSNAME]`"). The feature of class i is denoted as $\mathbf{w}_i$. The likelihood of sampling a negative class is determined by the similarity between the class feature $\mathbf{w}_i$ and the image feature $\mathbf{x}$:

$$Q_i^{(s)}(\mathbf{x}) = \frac{\exp(\mathbf{x}^\top \mathbf{w}_i/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{x}^\top \mathbf{w}_j/\tau)}. \tag{2}$$

Table 3 illustrates the performance of different proposal distributions. Compared to the uniform distribution, using the frequency distribution for sampling leads to degraded performance, particularly in cross-dataset and cross-domain settings, due to reduced sampling of rare categories. This highlights the importance of a large number of long-tail categories in the ImageNet-21K dataset for the generalization of the soft prompt. Additionally, the performance of the similarity distribution is also not as strong as that of the uniform distribution. The reason for this may be that as the soft prompt evolves, the features of hard negative classes change. However, the negative features used in (2) are obtained from the hard prompt, creating a fixed proposal distribution that is unable to adapt to these changes, potentially causing the soft prompt to converge to a local optimum. In contrast, POMP with the simple uniform distribution considers both common and rare classes, as well as easy and difficult classes, leading to the best performance for both the soft prompt and class features.

E.2 Ablation on #shot and Prompt Length

We further conduct ablation on the number of pre-training instances per class (#shot) and the prompt length to analyze their influence on the generalization ability of POMP. The left panel in Figure 3 illustrates the results of #shot. The green curve represents the average accuracy of 10 datasets under the cross-dataset evaluation, while the purple curve represents the avergaed accuracy of 4 datasets

under the cross-domain evaluation. Overall, the performance of POMP improves as #shot increases. We find that POMP can achieve decent cross-dataset and cross-domain accuracy even with #shot=1. This is due to the huge number of classes in ImageNet-21K. Even if there are only one instance per class, the overall amount of data (21K instances for 21K classes) is enough for training a soft prompt with only 0.012 M learnable parameters.

The right panel in the figure shows the results of the prompt length. The soft prompt of length 16 achieves 65.0% accuracy across datasets, which is lower than the soft prompt of length 4 with 67.0% cross-dataset accuracy. It indicates that the prompt with too large lengths impairs its generalization, which consistent with the findings from previous work [31, 14].

E.3 Ablation on Prompt Types for Semantic Segmentation

Table 4: Cross-dataset evaluation for semantic segmentation. All methods share the same visual backbone with ZSSeg, but use different prompts.

Method	Source Dataset: Standard COCO Stuff				Target Dataset: ADE20K				Target Dataset: PASCAL Context			
	mIoU	fwIoU	mACC	pACC	mIoU	fwIoU	mACC	pACC	mIoU	fwIoU	mACC	pACC
ZSSeg (single prompt)	40.5	47.8	53.5	61.7	17.8	44.0	31.0	52.9	51.8	64.6	69.9	74.3
ZSSeg (ImageNet prompts)	40.9	48.4	54.7	62.3	17.7	46.5	31.8	57.1	52.0	64.7	70.3	75.4
ZSSeg (ViLD prompts)	40.9	48.6	54.2	62.3	20.2	49.1	33.4	60.7	51.8	63.8	69.6	73.8
ZSSeg (POMP prompt, ours)	**41.2**	**49.0**	**54.7**	**62.6**	**20.6**	**49.3**	**35.0**	**61.7**	**52.4**	**65.3**	**70.6**	**76.4**

We perform an ablation study on prompt types for cross-dataset semantic segmentation to further demonstrate the superior generalization ability of our prompt on downstream tasks. Specifically, we take ZSSeg as the backbone and evaluate the performance of four types of prompts, as described in Appendix B. As shown in Table 4, ZSSeg with our POMP prompt achieves the highest performance on the three datasets. It is noteworthy that, despite using 80 hard prompts for *ImageNet prompts* and 14 for *ViLD prompts* for prompt ensemble, their performance was consistently worse than our POMP with just one soft prompt, highlighting the effectiveness of our method.

F Limitations

To facilitate future research, we analyze the limitations in our work and propose potential solutions. **(1)** We present the local contrast and use the loss within a subsampled class set as an empirical estimation for the expected contrastive loss within the full class set. However, the theoretical risk of such an estimation is urged to be investigated. **(2)** ImageNet-21k comprises a vast number of classes that are organized based on a semantic structure. By leveraging the hyponym and hypernym relations provided by WordNet synsets, we can derive the parent class and a list of child classes for each class. We believe that utilizing the semantic information holds the potential to further enhance performance. **(3)** Despite the excellent performance exhibited by our pre-trained prompt, its interpretability poses a significant challenge because the context vectors are optimized in a continuous space. We leave it as future work.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *ECCV*, 2014.

[2] Holger Caesar, Jasper R. R. Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2016.

[3] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Neural Information Processing Systems*, 2021.

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[6] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010.

[7] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178, 2004.

[8] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Zero-shot detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[9] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. Lvis: A dataset for large vocabulary instance segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5351–5359, 2019.

[10] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[11] Patrick Helber, Benjamin Bischke, Andreas R. Dengel, and Damian Borth. 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*, 12:2217–2226, 2019.

[12] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Lixuan Zhu, Samyak Parajuli, Mike Guo, Dawn Xiaodong Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8320–8329, 2020.

[13] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Xiaodong Song. Natural adversarial examples. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15257–15266, 2019.

[14] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. *ArXiv*, abs/2210.03117, 2022.

[15] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013.

[16] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

[17] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *ArXiv*, abs/1306.5151, 2013.

[18] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Loddon Yuille. The role of context for object detection and semantic segmentation in the wild. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.

[19] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008.

[20] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 3498–3505. IEEE Computer Society, 2012.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[22] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 2019.

[23] T. Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *ArXiv*, abs/2104.10972, 2021.

[24] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438, 2019.

[25] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *ArXiv*, abs/1212.0402, 2012.

[26] Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary Chase Lipton. Learning robust global representations by penalizing local predictive power. In *Neural Information Processing Systems*, 2019.

[27] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero- and few-label semantic segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8248–8257, 2019.

[28] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010.

[29] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *ArXiv*, abs/2112.14757, 2021.

[30] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017.

[31] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16795–16804, 2022.

[32] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krahenbuhl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, 2022.

[33] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection. *ArXiv*, abs/2103.07461, 2021.