

ECO: Ensembling Context Optimization for Vision-Language Models

Lorenzo Agnolucci*
University of Florence

lorenzo.agnolucci@unifi.it

Alberto Baldrati*
University of Florence

alberto.baldrati@unifi.it

Francesco Todino
University of Florence

francesco.todino@stud.unifi.it

Federico Becattini
University of Siena

federico.becattini@unisi.it

Marco Bertini
University of Florence

marco.bertini@unifi.it

Alberto Del Bimbo
University of Florence

alberto.delbimbo@unifi.it

Abstract

Image recognition has recently witnessed a paradigm shift, where vision-language models are now used to perform few-shot classification based on textual prompts. Among these, the CLIP model has shown remarkable capabilities for zero-shot transfer by matching an image and a custom textual prompt in its latent space. This has paved the way for several works that focus on engineering or learning textual contexts for maximizing CLIP’s classification capabilities. In this paper, we follow this trend by learning an ensemble of prompts for image classification. We show that learning diverse and possibly shorter contexts improves considerably and consistently the results rather than relying on a single trainable prompt. In particular, we report better few-shot capabilities with no additional cost at inference time. We demonstrate the capabilities of our approach on 11 different benchmarks.

1. Introduction

Thanks to their large-scale pre-training, foundational vision-language models proved to be very effective at generalizing to downstream tasks. In particular, CLIP (Contrastive Language-Image Pre-training) [15] has achieved surprising performance in several different fields, such as image generation [7], image retrieval [2, 1] and image quality assessment [18]. Specifically, CLIP can be employed for zero-shot classification by predicting the output class based on the similarity between the image features and the textual features of words belonging to a given vocabulary.

*Equal contribution

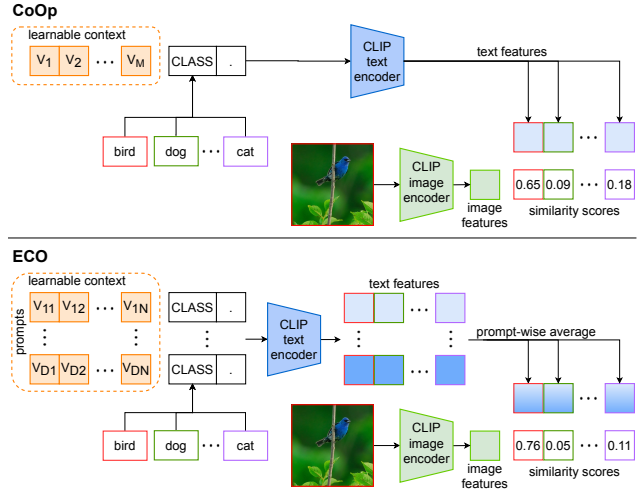


Figure 1: Overview of our approach. While CoOp uses a single prompt with M context tokens, ECO trains D prompts with N context tokens each, such that $M = D * N$. Given the same number of trainable parameters, ensembling multiple prompts with a reduced number of context tokens performs better than using a single prompt with a larger number of context tokens.

However, the textual input – referred to as *prompt* – greatly influences the performance in downstream tasks. For example, [21] reports a 5% increase in accuracy by adding an “a” before the class token in the prompt “a photo of [CLASS]” for few-shot classification with the Caltech101 [6] dataset. Given the significant difference in performance caused by slight changes in wording, crafting prompts by hand to find the best-performing one is a non-trivial task.

Therefore, *prompt ensembling* is often employed to improve the robustness and achieve better results [15]. Prompt ensembling consists of computing the textual features of several different prompts, such as “*a photo of a [CLASS]*”, “*an illustration of a [CLASS]*” etc., and then using the average of them for the downstream task.

Recently, several works have proposed to employ prompt learning to substitute hand-crafted prompts with learned context word vectors. CoOp [21] was the first work to propose to use prompt learning for vision-language models, improving over hand-crafted prompts. CoCoOp [20] trains a neural network to generate an input-conditional token for each image. MaPLe [10] proposes to learn a multi-modal prompt instead of a textual-only one. However, all existing methods only learn a single prompt, thus not exploiting the potential of prompt ensembling.

For this reason, we present ECO (Ensembling Context Optimization), a method for merging prompt learning and prompt ensembling. The main idea of our approach is conceptually quite straightforward: learning multiple prompts with a reduced number of context tokens instead of a single one with a larger number of context tokens, and then combining them with prompt ensembling. Figure 1 shows an overview of the proposed method and a comparison with CoOp [21]. Note that ECO is orthogonal to the prompt learning technique being used, as it focuses on how to take full advantage of the information of the learned prompts rather than how to obtain it. Despite its apparent simplicity, our approach performs significantly better than the competing methods on 11 testing datasets. Moreover, it proves to be a more data-efficient and effective few-shot learner, since the largest gains in performance are observed for as few as 1 and 2 shots. Finally, ECO does not add any computational overhead at inference time since the textual features used for the classification can be precomputed.

We summarize the contributions of this work as follows:

- We propose ECO, an approach for prompt learning that employs prompt ensembling to combine multiple prompts with reduced learned context tokens;
- ECO can be combined with any prompt learning strategy, making it a simple and versatile tool for improving accuracy with no overhead at inference time;
- We obtain significant improvements over the competing methods on 11 testing datasets, showing the effectiveness of our method.

2. Method

2.1. Preliminaries

The vision-language model CLIP [15] is designed to align visual and textual data within a common embedding

space. It consists of two encoders: a visual encoder denoted as f_θ and a text encoder represented as g_ϕ . These encoders extract feature representations $f_\theta(I) \in \mathbb{R}^d$ and $g_\phi(E_w(Y)) \in \mathbb{R}^d$ from an input image I and its corresponding text caption Y , respectively. Here, d indicates the dimension of the CLIP embedding space, while E_w represents the word-embedding layer, which maps each tokenized word in Y to the token embedding space \mathcal{W} . The primary objective of training the CLIP model is to ensure a high similarity between the feature representations of corresponding images and text, i.e. $f_\theta(I) \approx g_\phi(E_w(Y))$.

In the zero-shot classification setup using CLIP, we start with an image I and a set of text prompts $\{Y_i\}_{i=1}^K$, where K represents the number of classes. Each text prompt Y_i is of the form “*a photo of a [CLASS_i]*”, with CLASS_i denoting a specific class name, such as “*bird*”, “*dog*”, “*cat*”, etc. We then extract feature representations from the image and the text prompts using the CLIP encoders. The image feature representation is denoted as $\psi_I = f_\theta(I)$, while the text feature representation for each prompt is represented as $\psi_T^i = g_\phi(E_w(Y_i))$. Finally, we can compute the prediction probability for each class as follows:

$$p(y = i|I) = \frac{\exp(\cos(\psi_T^i, \psi_I)/\tau)}{\sum_{j=1}^K \exp(\cos(\psi_T^j, \psi_I)/\tau)}, \quad (1)$$

Here, τ is a temperature parameter that is learned during the training of the CLIP model, and $\cos(\cdot, \cdot)$ represents the cosine similarity between the image and text features.

2.2. ECO

Our approach, named ECO, aims to enhance the adaptability of frozen pre-trained CLIP models to downstream tasks by overcoming the inefficiency of hand-crafted prompts. Previous methods, such as CoOp [21], CoCoOp [20], and MaPLe [10], learn a single set of context tokens. On the contrary, drawing inspiration from prompt ensembling techniques that have proven to boost performance over using a single prompt [15], we learn multiple sets of context tokens. In other words, while standard prompt learning techniques learn only a single prompt, we learn multiple prompts that we combine together to improve performance.

We denote the multiple sets of context tokens (*i.e.* the learnable prompts) as $\{v_{i1}, \dots, v_{iN}\}_{i=1}^D$, where each context vector v_{ij} belongs to the CLIP token embedding space \mathcal{W} . Here, N represents the number of context tokens per prompt, while D is the total number of prompts. For the k -th class of a dataset, the inputs to the text encoder are defined as $\{v_{i1}, \dots, v_{iN}, c_k\}_{i=1}^D$, where $c_k = E_w([\text{CLASS}_k])$. Similarly to CoOp, we share the same set of context vectors among all classes. We then extract the textual features using the textual encoder, averaging across prompts

$\psi_T^k = \frac{1}{D} \sum_{i=1}^D g_\phi(\{v_{i1}, \dots, v_{iN}, c_k\})$. Consequently, we can compute the probability $p(y = k|I)$ using Eq. (1).

The key innovation lies in our use of multiple prompts. We learn distinct sets of context vectors from data instead of relying on hand-crafted prompts like "a photo of a [CLASS]". Intuitively, each prompt contributes to a diverse feature extraction process, and we effectively blend the prompt-specific features by performing an element-wise average. This prompt-wise average conceptually emulates prompt ensembling, known to enhance CLIP’s zero-shot classification performance [15]. However, unlike standard prompt ensembling with hand-created prompts, our method learns context vectors directly from the data. To summarize, ECO seamlessly combines the concepts of prompt learning and prompt ensembling, a novel combination not previously explored in vision-language tasks.

During training, we employ cross-entropy as the loss function, allowing the gradients to flow through the text encoder to update the weights of the context vectors. Importantly, the CLIP base model remains frozen throughout the entire training process. To ensure a fair comparison with CoOp, we keep the number of trainable parameters constant. If CoOp uses M context vectors, we set N and D such that $M = N * D$. Note that our method coincides with CoOp when $D = 1$ and $N = M$. Although in our experiments we extend the CoOp method, what we propose is a general framework that can be extended to all prompt learning techniques that learn a single set of context tokens. In addition, ECO does not add any computational overhead at inference time. Despite learning multiple contexts, after training these are fixed and their encodings are averaged into a single latent vector ψ_T^k . Since ψ_T^k does not depend on the input, it can be stored and used as a single prompt, requiring no additional computation compared to non-ensembling models like CoOp.

3. Experimental Results

Since ECO does not depend on a specific prompt learning technique, we choose to compare our approach to the most basic one, *i.e.* CoOp [21]. In future work, we will extend the proposed method to other prompt learning works, such as CoCoOp [20] and MaPLe [10].

3.1. Evaluation Protocol

We follow the few-shot evaluation protocol of [15, 21], using 1, 2, 4, 8, and 16 shots for training and evaluating the performance of each model in the full test sets. We report the average results over three seeds.

Similarly to [21], we evaluate our approach on 11 image classification datasets: ImageNet [5], Caltech101 [6], OxfordPets [14], StanfordCars [11], Flowers102 [13], Food101 [3], FGVCAircraft [12], SUN397 [19], DTD [4], EuroSAT [9] and UCF101 [16].

Method	Shots				
	1	2	4	8	16
Zero-Shot CLIP [‡] [15]	58.77	58.77	58.77	58.77	58.77
Linear Probe CLIP [15]	36.67	47.61	57.19	64.98	71.10
CoOp [21]	59.59	62.32	66.77	69.89	73.42
ECO ($D=16, N=1$)	62.42	63.97	66.10	69.72	72.82
ECO ($D=8, N=2$)	63.18	65.16	67.90	70.72	73.45
ECO ($D=2, N=8$)	61.76	64.51	67.26	70.95	73.71
CoOp [†] ($D=1, N=16$)	59.43	62.36	66.49	69.74	73.18
ECO ($D=4, N=4$)	62.90	65.24	68.26	71.33	74.03
	+3.47	+2.88	+1.77	+1.59	+0.85

Table 1: Detailed comparison of the results on the average of the 11 datasets. Best scores are highlighted in bold. [‡] uses always zero shots. [†] indicates results obtained with our implementation. Note that CoOp[†] coincides with ECO ($D = 1, N = 16$). Absolute gains over CoOp[†] [21] are indicated in blue.

We consider the version of CoOp with the class token positioned at the end, ResNet 50 [8] as the backbone and with the number of context tokens $M = 16$. We inherit the training details from [21]. For a fair comparison, in the experiments, we vary the number of prompts D and the number of context tokens N for each of them so that the number of trainable parameters stays the same (*i.e.* $M = D * N$). Note that, for $D = 1$ and $N = 16$, ECO coincides with CoOp.

3.2. Quantitative Results

Figure 2 shows the results of ECO for all the testing datasets. For completeness, we also report the performance of zero-shot CLIP, which is based on hand-crafted prompts. ECO obtains significant improvements over the baselines on all the datasets. The version of ECO with $N = 4$ and $D = 4$ achieves the best performance on average and proves to be the best tradeoff between the number of prompts and context tokens. In addition, ECO is less sensitive to noisy labels than CoOp [21], as it achieves better performance than zero-shot CLIP also on the Food101 dataset, which is known to have noisy annotations [21].

In Table 1 we provide a comparison between the different versions of ECO and the baselines by reporting the average accuracy on the 11 benchmarks, varying the number of shots. For a fair comparison, for CoOp we report the results we obtained with the version of our model with $D = 1$ and $N = 16$, since they coincide. We denote this in Table 1 as CoOp[†]. However, we observed a difference of only 0.32% on average with the values of the original paper [21], which can be attributed to different seeds and hardware. For completeness, we also provide the results of the linear probe model of CLIP, which is considered a strong few-shot learning baseline [17]. Our approach consistently outperforms all the competing methods. In particular, we observe abso-

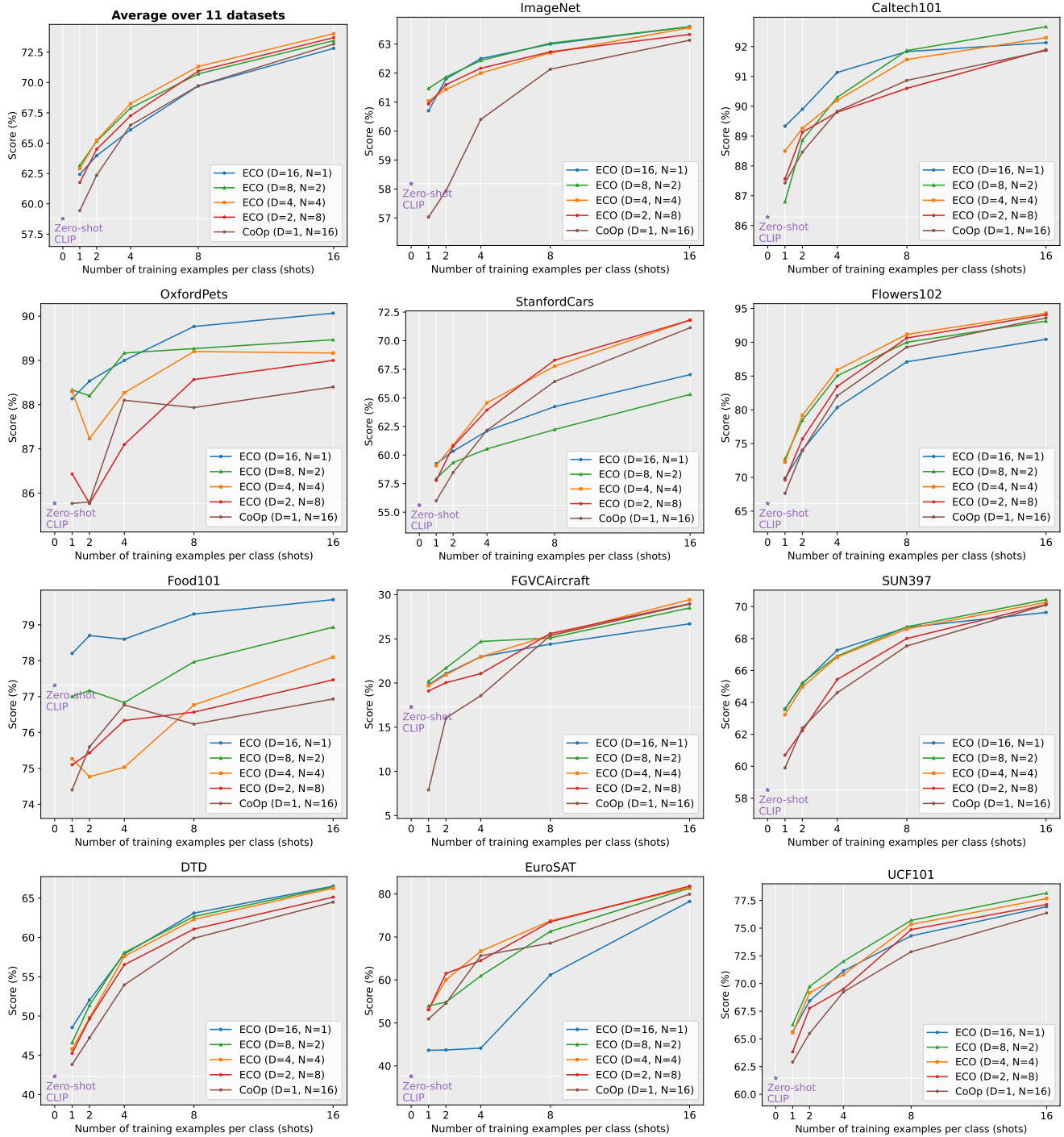


Figure 2: Quantitative results on the 11 test datasets varying the number of shots, prompts D and context tokens N for each of them. Note that CoOp [21] coincides with ECO when $D=1$ and $N=16$.

lute improvements up to 3.47 over CoOp. Moreover, the results show that ECO is a better few-shot learner and is more data-efficient than CoOp since the largest gains in performance are obtained for as few as 1 and 2 shots.

Overall, the experimental results demonstrate that, when utilizing an equivalent number of trainable parameters, employing an ensemble of multiple prompts with a reduced number of context tokens performs better than using a sin-

gle prompt with a larger number of context tokens.

4. Conclusion

In this paper, we have proposed a novel prompt learning strategy that consists in optimizing an ensemble of multiple contexts. Although simple, the method is effective, yielding consistent improvements over 11 different benchmarks, and versatile, being it applicable on top of potentially any existing prompt learning technique with no additional overhead at inference time. Interestingly, we found that balancing context length and number of prompts is beneficial for effectively exploit CLIP for few-shot image classification. This is particularly true for a reduced number of shots, such as 1 or 2, for which we report the bigger gains.

References

- [1] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. *arXiv preprint arXiv:2303.15247*, 2023.
- [2] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21466–21474, 2022.
- [3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014.
- [4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [6] 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. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.
- [7] Federico A Galatolo, Mario GCA Cimino, and Gigliola Vaglini. Generating images from caption and vice versa via clip-guided generative latent space search. *arXiv preprint arXiv:2102.01645*, 2021.
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [9] Patrick Helber, Benjamin Bischke, Andreas 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(7):2217–2226, 2019.
- [10] 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.
- [11] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.
- [12] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.
- [13] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.
- [14] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.
- [15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [16] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- [17] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 266–282. Springer, 2020.
- [18] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2555–2563, 2023.
- [19] 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.
- [20] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.
- [21] 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.