Open-set Fine-grained Retrieval via Prompting Vision-Language Evaluator

Shijie Wang[1], Jianlong Chang[2], Haojie Li[1,3*], Zhihui Wang[1], Wanli Ouyang[4], Qi Tian[2]
[1]International School of Information Science & Engineering, Dalian University of Technology, China
[2] Huawei Cloud & AI, China
[3]College of Computer and Engineering, Shandong University of Science and Technology, China
[4] Sense Time Computer Vision Research Group, The University of Sydney, Australia

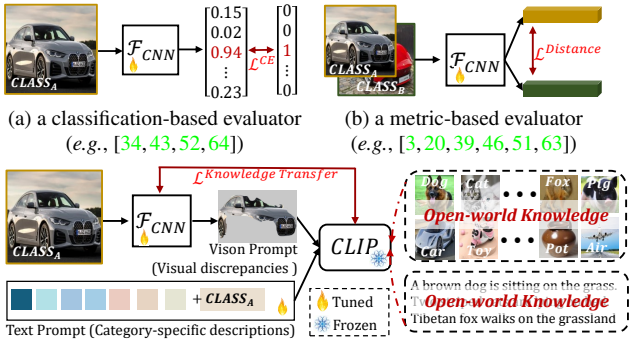
Abstract

Open-set fine-grained retrieval is an emerging challenge that requires an extra capability to retrieve unknown subcategories during evaluation. However, current works focus on close-set visual concepts, where all the subcategories are pre-defined, and make it hard to capture discriminative knowledge from unknown subcategories, consequently failing to handle unknown subcategories in open-world scenarios. In this work, we propose a novel Prompting vision-Language Evaluator (PLEor) framework based on the recently introduced contrastive language-image pretraining (CLIP) model, for open-set fine-grained retrieval. PLEor could leverage pre-trained CLIP model to infer the discrepancies encompassing both pre-defined and unknown subcategories, called category-specific discrepancies, and transfer them to the backbone network trained in the close-set scenarios. To make pre-trained CLIP model sensitive to category-specific discrepancies, we design a dual prompt scheme to learn a vision prompt specifying the category-specific discrepancies, and turn random vectors with category names in a text prompt into category-specific discrepancy descriptions. Moreover, a vision-language evaluator is proposed to semantically align the vision and text prompts based on CLIP model, and reinforce each other. In addition, we propose an open-set knowledge transfer to transfer the category-specific discrepancies into the backbone network using knowledge distillation mechanism. Quantitative and qualitative experiments show that our PLEor achieves promising performance on open-set fine-grained datasets.

1. Introduction

Open-set fine-grained retrieval (OSFR) attempts to build a well-generalized embedding space where the visual discrepancies among unknown subcategories are clearly reflected. It plays a vital role in numerous vision applications from fashion industry, *e.g.*, retrieval of diverse types of clothes [1, 31], to environmental conservation, *e.g.*, retrieving endangered species [7, 49, 50]. As shown in Fig. 1(a)(b), existing works follow a close-set learning setting, where all the subcategories are pre-defined, and evaluate embeddings identifying the visually similar objects of pre-defined subcategories. However, such evaluation focuses on closed-set visual concepts, limiting the model to a pre-defined list of subcategories, and is not generalizable when it comes to unknown subcategories unseen during training.

Fortunately, recent works [66, 67] using large-scale contrastive language-image pretraining (CLIP) model [37] have shown great potentials in alleviating this limitation. As shown in Fig. 1(c), CLIP model is pretrained from scratch on a dataset of 400 moillion image-text pairs, which are automatically collected from the publicly available sources on



(a) a classification-based evaluator (*e.g.*, [34, 43, 52, 64])
(b) a metric-based evaluator (*e.g.*, [3, 20, 39, 46, 51, 63])
(c) a vision-language evaluator with open-world knowledge

Figure 1. Comparison on existing evaluators in open-set fine-grained retrieval. Although our PLEor (c) is trained in a close-set scenarios, similar with previous works (a) (b), it could mine the category-specific discrepancies using pre-trained CLIP model aided by vision and text prompts, and transfer the discrepancies encompassing both pre-defined and unknown subcategories to our model. This enables our model to procure in-depth understanding for unknown subcategories owing to distilling the knowledge with open-world visual concepts from CLIP model, improving retrieval performance eventually in open-set scenarios.

*Corresponding author: hjli@dlut.edu.cn

the Internet. Based on this, CLIP model could associate much wider range of visual concepts in the images with their text descriptions, rather than a fixed set of pre-defined categories. Therefore, one question naturally arises: is it possible that we can effectively exploit the open-set visual concepts in CLIP model to solve OSFR task? It is already answered yes by recent studies exploring how to transfer the knowledge from CLIP model to other downstream tasks via prompt techniques [10, 16, 26, 37, 56, 66, 67]. However, their prompt strategies are tailored for capturing category-level semantic (*e.g.*, dog and cat) rather than more detailed visual discrepancies for distinguishing fine-grained objects (*e.g.*, different breeds of dogs). Therefore, how to effectively make pre-trained CLIP model sensitive to the visual discrepancies encompassing both pre-defined and unknown subcategories (termed as category-specific discrepancies), and transfer these discrepancies to the model trained in closed-set scenarios is worthy of investigation.

To this end, we design a novel prompting vision-language evaluator (PLEor) for OSFR, based on the power of recently introduced CLIP model. Technically, to make pre-trained CLIP model sensitive to category-specific discrepancy, we design a dual prompt scheme composed of vision prompt and text prompt for explicitly highlighting the category-specific discrepancies from the input perspective. Concretely, the vision prompt specifies the category-specific discrepancies via parsing semantic features inferred by the backbone network. And the text prompt turns random vectors with category names into category-specific discrepancy descriptions. Meanwhile, a vision-language evaluator is proposed to encourage pre-trained CLIP model to locate the category-specific descriptions in vision prompt and generate the category-specific visual semantics into text prompt. In this way, the OSFR task aided by the designed prompts is close to the solved task of pre-training CLIP model, thus making the CLIP model sensitive to category-specific discrepancy. Nevertheless, a non-negligible problem is that the corporation of CLIP model and backbone network is quite complex, leading to very time consuming and memory demanding during evaluation. Thereby, we propose an open-set knowledge transfer module to transfer the category-specific discrepancies from CLIP model to the backbone network using knowledge distillation mechanism.

Our contributions are summarized as follows:

- A prompting vision-language evaluator, *i.e.*, PLEor, is proposed. It can distill the knowledge with open-world visual concepts from CLIP model to alleviate the problems behind open-set scenarios. To our best knowledge, we are the first to regard CLIP model as an evaluator specifically for OSFR task.

- PLEor provides timely insights into the adaptation of pre-trained CLIP model adopting prompt learning, and crucially, demonstrates the effectiveness of a simple modification for inputs of CLIP model in OSFR.

- PLEor achieves new state-of-the-art results compared with classification-based and metric-based evaluators, which is significant gains of 8.0% average retrieval accuracy on three widely-used OSFR datasets.

2. Related Work

Open-set fine-grained retrieval. Existing open-set fine-grained retrieval works can be roughly divided into two groups. The first group, *classification-based schemes*, utilizes the supervision of category signals to learn discriminative embeddings [34, 48, 52, 64]. Although these works have made an inspiring achievement, their shortcoming lies in their narrow focus on individual samples, while overlooking inter-class and intra-class correlations between subcategories, ultimately leading to a decrease in retrieval performance. The second group of schemes, namely *metric-based schemes*, learn an embedding space that attracts similar examples and repels dissimilar [3, 18, 20, 39, 46, 51, 62, 63]. However, they are rooted in the close-set scenarios and thus make it hard to capture discriminative discrepancies from unknown subcategories, inevitably impairing the retrieval performance. To alleviate the problem behind open-set scenarios, we design a PLEor to transfer the visual discrepancies encompassing both pre-defined and unknown subcategories from pretrained CLIP model to our model trained in close-set scenarios.

Vision-language pretraining. Yasuhide et. al [33] has studied the connection between images and words using pair-wise text documents, and existing works [8, 53] proposed to jointly explore image-text alignment with the category names. Recently, CLIP [37], ALIGN [13] and FILIP [54] have further scaled up the training with large-scale Internet data. It is shown that powerful representation could be learned from image-text pairs via simple noise contrastive learning. A large amount of follow-up works have been proposed to utilize the pre-trained models for various downstream tasks, *e.g.*, few-shot transfer [9, 60, 67], point cloud understanding [38, 61] and video understanding [16, 47]. However, these works still follow the principles of using the large-scale vision-language models as backbone networks, leading to very time consuming and memory demanding. Differently, we are the first to tend to treat the pre-trained vision-language models as an evaluator and use it only during training.

Prompt learning. Prompting [29] in NLP reformulates the downstream tasks into a language modeling problem, enabling a pre-trained language model to be adapted more efficiently to new tasks. Hence, prompt techniques are now being used to address a wide range of NLP tasks, including language understanding and generation [15, 22, 23, 30].

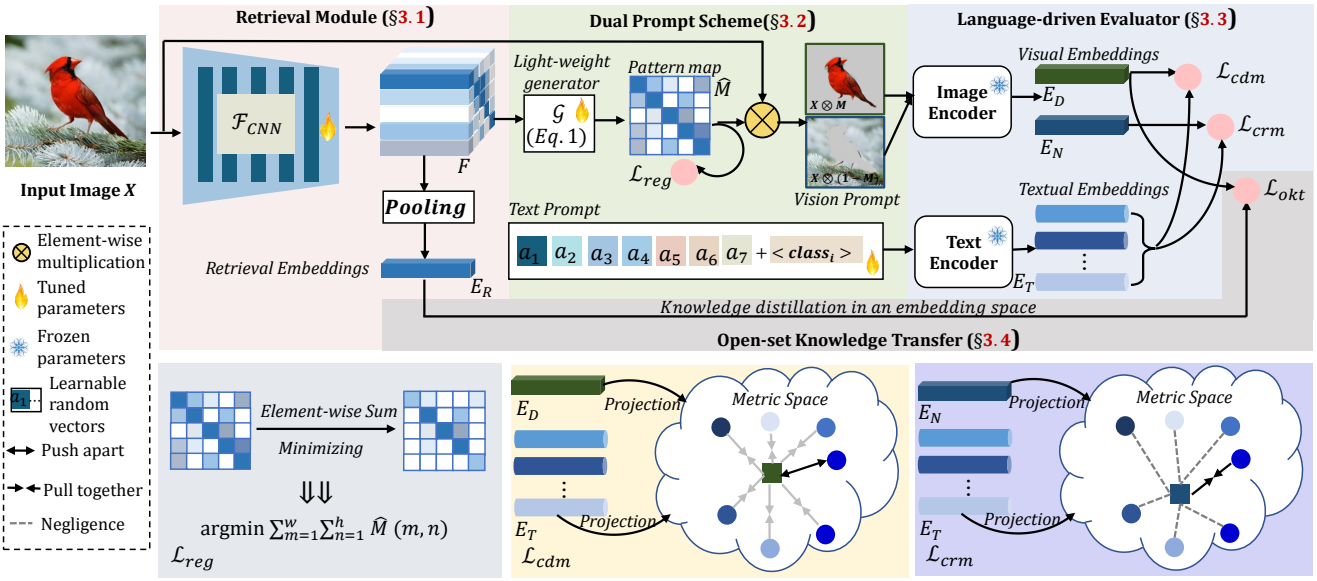


Figure 2. Detailed illustration of **prompting vision-language evaluator**. See §3 for more details.

Recently, prompt scheme has been integrated into multi-modal computer vision [10, 16, 37, 55, 66, 67]. However, current prompt techniques in multi-modal applications primarily extend the capabilities of language-based models, which cannot be directly applied to pre-trained vision models. In addition, recent works [14, 24, 28, 35] design a vision prompt introducing a few learnable parameters to steer the pre-trained models for general vision tasks. However, their prompt strategies are tailored for capturing category-level semantic rather than more detailed visual discrepancies for distinguishing fine-grained objects. Therefore, we design a dual prompt scheme module, *i.e.*, vision prompt and text prompt, to make pre-trained CLIP model sensitive to category-specific discrepancies.

Knowledge distillation. This paper is associated with knowledge distillation [2, 12, 58], which aims to transfer the knowledge from a well-trained teacher model to a student model. Most classification-based works pay attention to improving the student model by imitating the prediction output or distribution of teacher model. Moreover, existing researchers [5, 6, 19] also study knowledge distillation for image retrieval tasks via exploring distances between samples. This involves *e.g.*, learning to rank [6] and regression on quantities containing one or more pairs like distances [57] or angles [36]. Direct regression on embedding is not commonly used or demonstrated to be inferior [57], but we think it is actually much more effective than previously thought.

3. Methodology

The overall structure of PLEor is shown in Fig. 2. It is clear that our network is mainly organized by four modules: retrieval module, dual prompt scheme module, vision-language evaluator module and open-set knowledge transfer module. The dual prompt scheme module and vision-language evaluator module are designed to make pre-trained CLIP model sensitive to the discrepancies encompassing both pre-defined and unknown subcategories. In addition, the open-set knowledge transfer module is responsible for transferring these discrepancies to our model trained in close-set scenarios.

3.1. Retrieval Module

The retrieval module aims at extracting basic image representations and producing the final retrieval embeddings using the backbone network. Given an input image $\mathbf{X}$, let $\mathbf{F} \in \mathbb{R}^{W \times H \times C}$ be the C-dimensional with $H \times W$ feature planes encoded by a backbone network $\mathbf{F} = \mathcal{F}_{CNN}(\mathbf{X})$. Thus the most common way for retrieval is to embed the final feature $\mathbf{F}$ by using global average pooling operations (GAP), calculating mean values on the $H \times W$ feature plane and producing the final retrieval embeddings $\mathbf{E_R} \in \mathbb{R}^C$. It should be clarified that our PLEor does not introduce extra computation overhead during evaluation.

3.2. Dual Prompt Scheme

Subtle yet discriminative discrepancies are widely recognized to be significant for fine-grained understanding [34, 52, 64]. However, CLIP model is originally designed to model the visual concepts identifying various species (*e.g.*, cat, dog and person), instead of mining subtle discrepancies among subcategories within a certain species. To alleviate this, we devise a dual prompt scheme to solely modify the inputs of vision and text for pre-trained CLIP model.

This makes the fine-grained retrieval task aided by the dual prompt scheme similar to those solved of pre-training CLIP model. Specifically, the dual prompt scheme are composed of vision prompt and text prompt. The vision prompt specifies the category-specific discrepancies via parsing semantic features inferred by the backbone network. And the text prompt turns random vectors with category names into category-specific discrepancy descriptions.

Vision Prompt. To obtain the category-specific discrepancies, the vision prompt aims to project the semantic features into a new space where the location, scale and intensity of these discrepancies are specified. Concretely, we map the features into a pattern map $\hat{\mathbf{M}} \in \mathbb{R}^{W \times H}$. This map can be generated by a light-weight generator $\mathcal{G}$ as below:

$$\hat{\mathbf{M}} = \sigma(\mathcal{G}(\mathbf{F})), \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid activation function, and $\mathcal{G}$ is a convolution with kernel size 1. Then the pattern map is required to be resized to the size of the original input image.

With the amplified pattern map $\mathbf{M}$, we can split the original input image $\mathbf{X}$ into two vision prompts as follows:

$$\mathbf{V_D} = \mathbf{X} \otimes \mathbf{M}, \quad \mathbf{V_N} = \mathbf{X} \otimes (1 - \mathbf{M}), \tag{2}$$

where $\mathbf{V_D}$ and $\mathbf{V_N}$ are the category-specific vision prompt and category-irrelevant vision prompt, respectively. $\otimes$ denotes element-wise multiplication. Note that, since each element in the pattern map belongs to 0 to 1, it can weigh the importance of category-specific discrepancies instead of equally treating them.

For fine-grained understanding, the pattern map should solely cover the discrepancies of an object, so that we can better identify objects relying on the discrepancies involved in the pattern map. Therefore,we apply the regularization constraint to restrain the size of response in the pattern map, which ensures that the irrelevant responses are excluded:

$$\mathcal{L}_{reg} = \frac{1}{W \times H} \sum_{m=1}^{W} \sum_{n=1}^{H} \hat{\mathbf{M}}(m, n). \tag{3}$$

Text Prompt. When understanding a fine-grained object, human can instinctively seek help from discriminative visual clues. For example, the extra semantic information, such as appearance descriptions, will make it easier to distinguish fine-grained objects among subcategories. However, it is difficult to acquire such visual semantics in OSFR task due to only providing the category names, which are pre-defined and fixed. To handle this limitation, we design a text prompt to generate appropriate text descriptions automatically via keeping semantically coherent with the category-specific vision prompt.

Concretely, we construct the "virtual" prompt template via combining the category names and random vectors:

$$\mathcal{P}_{\mathbf{class}} = (a_1, a_2, \cdots, a_i, \cdots, a_k, < class >), \tag{4}$$

where $\mathcal{P}_{\mathbf{class}} \in \mathbb{R}^{N \times (k+1) \times D}$ is the text prompt of all subcategories, N is the number of subcategories, k is the number of prompt vectors and D is the vector dimension. $a_i \in \mathbb{R}^D$ denotes the i-th prompt vector, consisting of several learnable parameters. $< class > \in \mathbb{R}^D$ refers to the generated word embeddings for this category name. Note that these prompt vectors a are shared for all subcategories, thus they are only task-specific. Ultimately, these learnable prompt vectors cooperating with category names end up constructing virtual prompt templates, which can be understood by the text encoder of CLIP model to generate appropriate descriptions regarded as extra discriminative clues.

3.3. Vision-language Evaluator

Our goal is to make the pre-trained CLIP model sensitive to the category-specific discrepancies. Thus, the key challenge is to let the pre-trained CLIP model learn discriminative representation that can attend to vision prompt and text prompt. To achieve this, we design a vision-language evaluator to mutually align vision prompt and text prompt into semantic space via contrastive learning. In one word, the contrastive objective of vision-language evaluator encourages the pre-trained CLIP model to locate the category-specific descriptions in vision prompt and generate the category-specific semantics into text prompt.

By forwarding the vision prompt and text prompt to the image encoder Φ_I and text encoder Φ_T of pre-trained CLIP model, respectively, we can obtain the corresponding visual and textual embeddings:

$$\mathbf{E_D} = \Phi_I(V_D), \mathbf{E_N} = \Phi_I(V_N), \mathbf{E_T} = \Phi_T(\mathcal{P}_{\mathbf{class}}), \tag{5}$$

where $\mathbf{E_D} \in \mathbb{R}^C$ and $\mathbf{E_N} \in \mathbb{R}^C$ are category-specific and category-irrelevant visual embeddings, respectively. $\mathbf{E_T} \in \mathbb{R}^{N \times C}$ are category-specific textual embeddings.

Category-specific discrepancy matching. Given the category-specific visual embeddings $\mathbf{E_D}$ and the category-specific textual embeddings $\mathbf{E_T}$, we can calculate the category-specific discrepancy matching loss:

$$\mathcal{L}_{cdm} = -\sum_{i=1}^{N} y_i \cdot log \frac{exp(cos < \mathbf{E_D}, \mathbf{E_T^i} > /\tau)}{\sum_{i=1}^{N} exp(cos < \mathbf{E_D}, \mathbf{E_T^i} > /\tau)}, \tag{6}$$

where y is the class label, $\mathbf{E_T^i}$ indicates the corresponding textual embeddings of the y_i-th category, τ denotes the hyperparameter of temperature, and $cos <, >$ indicates the cosine similarity between visual and textual embeddings. The category-specific vision prompt and text prompt will gradually approach category-specific discrepancies under the supervision of $\mathcal{L}_{cdm}$. However, $\mathcal{L}_{cdm}$ alone can not encourage CLIP model to inject the complementary discrepancies into the category-specific vision and text prompts.

Category-irrelevant region matching. To improve the completeness of category-specific discrepancies, we devise

the category-irrelevant region matching constraint $\mathcal{L}_{crm}$. Formally, we denote the category-irrelevant vision-text pair as $(\mathbf{E_N}, \mathbf{E_T})$ which contains category-irrelevant visual embeddings $\mathbf{E_N}$ and category-specific textual embeddings $\mathbf{E_T}$. Therefore, $\mathcal{L}_{crm}$ can be formulated as:

$$\mathcal{L}_{crm} = -\sum_{i=1}^{N} y_i \cdot log(1 - \frac{exp(cos < \mathbf{E_N}, \mathbf{E_T^i} > /\tau)}{\sum_{i=1}^{N} exp(cos < \mathbf{E_N}, \mathbf{E_T^i} > /\tau)}). \tag{7}$$

Optimizing $\mathcal{L}_{crm}$ can make the missing discrepancies recovered in category-specific vision and text prompts and thus ensure that more complete discrepancies are perceived by pre-trained CLIP model.

3.4. Open-set Knowledge Transfer

After the friendly cooperation of dual prompt scheme and vision-language evaluator, the pre-trained CLIP model with the aid of the backbone network could provide a remarkable retrieval performance under open-set scenarios. However, the complex combination is very time consuming and memory demanding for retrieval evaluation. Network distillation is proven to be one of the solutions to handle this problem in the classification filed [12]. Inspired by this, we propose an open-set knowledge transfer module to extend the theory of knowledge distillation to retrieval tasks that aims to project an image into an embedding space. Concretely, this module aims to transfer the category-specific discrepancy knowledge containing unknown subcategories from pre-trained CLIP model to the backbone network trained in close-set scenarios.

Formally, the retrieval embeddings $\mathbf{E_R}$ and the category-specific visual embeddings $\mathbf{E_D}$ are used for distillation:

$$\mathcal{L}_{okt} = \| \mathbf{E_R} - \mathbf{E_D} \|, \tag{8}$$

where $\| \cdot \|$ refers to the Frobenius norm. The retrieval embeddings can only learn from pre-defined subcategories. In contrast, the category-specific visual embeddings could contain the discriminative knowledge of both pre-defined and unknown subcategories, as pre-trained CLIP model can generalize. After optimizing $\mathcal{L}_{okt}$, the retrieval embeddings are sufficiently discriminative and generalized, thus better retrieving the visually similar objects under open-set scenarios accordingly.

3.5. Overall Training Objective

The overall training loss for the proposed PLEor can be formulated as:

$$\mathcal{L} = \alpha\mathcal{L}_{cdm} + \beta\mathcal{L}_{crm} + \gamma\mathcal{L}_{reg} + \lambda\mathcal{L}_{okt}, \tag{9}$$

where α, β, γ, and λ are the hyper-parameters to weight the four loss items.

Table 1. Comparison of performance and efficiency on CUB-200-2011 using different combinations of constraints. The first row indicates that we use the tranditional classification-based classifier (*i.e.*, ResNet-50) as supervision, to replace the proposed PLEor for comparison. "Time" is the time of extracted retrieval embeddings.

$\mathcal{L}_{cdm}$	$\mathcal{L}_{crm}$	$\mathcal{L}_{reg}$	$\mathcal{L}_{okt}$	Recall@1	Time
				66.3%	21.1ms
✓				72.1%	42.3ms
✓	✓			74.4%	42.3ms
✓	✓	✓		75.1%	42.3ms
✓	✓	✓	✓	**74.8%**	**21.1ms**

Table 2. Evaluation results of retrieval performance on CUB-200-2011 dataset with/without the prompt learning. Hand-craft prompt denotes that we use the handcrafted prompt template ("a photo of a [·].") in text prompt.

Prompt	Recall@1
CLIP + Hand-craft prompt	71.5%
CLIP + Text Prompt	$73.3\%_{+1.8}$
CLIP + Vision&Hand-craft Prompt	$72.4\%_{+0.9}$
CLIP + Vision&Text Prompt	$\mathbf{74.8\%}_{+3.3}$

4. Experiments

4.1. Experimental Setup

Datasets. CUB-200-2011 [4] consists of 200 bird species. We use the first 100 subcategories (5,864 images) for training and the consists of (5,924 images) for testing. The Stanford Cars [21] includes 196 car models. Similarly, we use the first 98 classes, which contain 8,054 images, for training and the remaining classes, which contain 8,131 images, for testing. Finally, FGVC Aircraft [32] is split into first 50 classes, containing 5,000 images, for training and the remaining 50 classes with 5,000 images, for testing.

Evaluation protocols. To evaluate the retrieval performance, we use *Recall@K* with cosine distance, which calculates the average recall score over all query images in the test set and strictly follows the setting in previous work [44]. Specifically, for each query, our model returns the top K similar images. In the top K returning images, the score will be 1 if there exists at least one positive image, and 0 otherwise.

Implementation Details. In our experiments, we use the widely-used ResNet-50 [11] as our backbone network, with pre-trained parameters. For CLIP model, the image and text encoders are adopted from pre-trained CLIP Resnet-50 and ViT-B/16, respectively. More importantly, **their parameters in CLIP model keep frozen.** The only trainable parameters are backbone network and vision&text prompts. Before feeding the images into our model, we resize them

Table 3. Compared with competitive methods on CUB-200-2011, Stanford Cars 196 and FGVC Aircraft datasets. "Arch" represents the architecture of utilizing backbone network. "R50" and "BN" respectively denote Resnet50 [11] and Inception V3 with BatchNorm [45].

Method	Arch	CUB-200-2011				Stanford Cars 196				FGVC Aircraft			
		1	2	4	8	1	2	4	8	1	2	4	8
SCDA $_{\text{TIP}_{17}}$ [52]	R50	57.3	70.2	81.0	88.4	48.3	60.2	71.8	81.8	56.5	67.7	77.6	85.7
CRL $_{\text{IJCAI}_{18}}$ [64]	R50	62.5	74.2	82.9	89.7	57.8	69.1	78.6	86.6	61.1	71.6	80.9	88.2
CEP $_{\text{ECCV}_{20}}$ [3]	R50	69.2	79.2	86.9	91.6	89.3	93.9	96.6	98.1	-	-	-	-
HDCL $_{\text{IJON}_{21}}$ [59]	R50	69.5	79.6	86.8	92.4	84.4	90.1	94.1	96.5	71.1	81.0	88.3	93.3
DGCRL $_{\text{AAAI}_{19}}$ [65]	R50	67.9	79.1	86.2	91.8	75.9	83.9	89.7	94.0	70.1	79.6	88.0	93.0
DCML $_{\text{CVPR}_{21}}$ [62]	R50	68.4	77.9	86.1	91.7	85.2	91.8	96.0	98.0	-	-	-	-
DAS $_{\text{ECCV}_{22}}$ [27]	R50	69.2	79.3	87.1	92.6	87.8	93.2	96.0	97.9	-	-	-	-
IBC $_{\text{ICML}_{21}}$ [41]	R50	70.3	80.3	87.6	92.7	88.1	93.3	96.2	98.2	-	-	-	-
NIA $_{\text{CVPR}_{22}}$ [40]	R50	70.5	80.6	-	-	89.1	93.4	-	-	-	-	-	-
Proxy $_{\text{CVPR}_{21}}$ [17]	BN	71.1	80.4	87.4	92.5	88.3	93.1	95.7	97.5	-	-	-	-
HIST $_{\text{CVPR}_{22}}$ [25]	R50	71.4	81.1	88.1	-	89.6	93.9	96.4	-	-	-	-	-
ETLR $_{\text{CVPR}_{21}}$ [18]	BN	72.1	81.3	87.6	-	89.6	94.0	96.5	-	-	-	-	-
PNCA++ $_{\text{ECCV}_{20}}$ [46]	R50	72.2	82.0	89.2	93.5	90.1	94.5	97.0	98.4	-	-	-	-
Our PLEor	R50	**74.8**	**84.5**	**91.3**	**94.9**	**94.4**	**96.9**	**98.3**	**98.9**	**86.3**	**91.7**	**95.1**	**96.7**

to 256×256 and then crop them into 224×224. During training, we utilize Stochastic Gradient Descent optimizer with a weight decay of 0.0001, momentum of 0.9, and batch size of 16. We adopt the widely-used data augmentation techniques, such as, random cropping, left-right flipping, and color jittering. We train our model end-to-end on a single NVIDIA 2080Ti GPU to accelerate the training process. The initial learning rate is set to 10^{-5}, with exponential decay of 0.9 after every 5 epochs. The total number of training epochs is set to 200. The momentum coefficient in semantic regularization is set to 0.2. The maximum number of textual tokens (D) in text prompt is 77 (following the official CLIP design), and the temperature hyper-parameter τ in Eq. (6)(7) is set to 0.01.

4.2. Ablation Studies

Efficacy of various constraints. To evaluate the efficacy of proposed prompting vision-language evaluator, we first employ ResNet-50 [11] with a fully connected layer as a classification-based evaluator, e.g., 66.3% on CUB-200-2011 dataset. It can be found in Tab. 1 that adopting the category-specific discrepancies matching constraint ($\mathcal{L}_{cdm}$) for mining the category-specific discrepancies, the performance boosts by 5.8%. While exploiting more complete discrepancies by introducing the category-irrelevant region matching constraint ($\mathcal{L}_{crm}$) can notably improve the retrieval performance. Besides, the regularization constraint ($\mathcal{L}_{reg}$) acting on the pattern map also enhances the focusing discrepancies, providing a stable improvement on result performance. More importantly, based on this high-performance baseline, we further add the open-set knowledge transfer constraint ($\mathcal{L}_{okt}$) to solve the problem of time consuming and memory demanding during evaluation and obtain a satisfactory performance compared to baseline.

Importance of the prompt scheme. Tab.2 presents the results for various prompt schemes. As the baseline, we directly use the handcrafted prompt templates ("a photo of a [·]") provided by the official CLIP in text prompt to guide the backbone network to discover category-specific discrepancies from input images. Although CLIP model may neglect vital discrepancies due to using the handcrafted prompt templates, the performance is much higher than baseline in Tab. 1 owing to its open-set visual concepts. Adding learnable text prompt templates instead of handcrafted prompt templates brings immediate benefits, with gains of 1.8%. This result reflects the learnable vectors in text prompt can learn a set of words describing the differences between subcategories, thus obtaining a performance boost. When we design the vision prompt scheme that emphasizes the visual discrepancies in images and combine it with handcrafted prompt templates, the retrieval performance only reaches 72.4%. Moreover, combining vision prompt with text prompt can discover more discrepancies among subcategories, thus improving the retrieval accuracy by 3.3%. Overall, all results suggest that, both learnable vision prompt and text prompt can reinforce each other and further capture more precise discrepancies to identify visually similar objects under open-world settings.

4.3. Comparisons with the State-of-the-Arts

We first compare the quality of the proposed prompting vision-language evaluator with previous open-set fine-

Table 4. Evaluation results of the text prompt with different number of category-specific vectors on CUB-200-2011 dataset.

Number (k)	4	8	16	32
Recall@1	72.4%	73.9%	**74.8%**	74.6%

Table 5. Comparison of the retrieval performance on CUB-200-2011 dataset using pretraining and fine-tuning CLIP models, respectively. It should be clarified that the parameters do not contain the ones of backbone network.

Optimization	Parameter	Recall@1	Recall@2
Fine-tuning	114.13M	65.4%	77.8%
Pre-training	**0.03M**	**74.8%**$_{+8.1}$	**84.5%**$_{+7.7}$

grained retrieval methods. Table 3 reports the performance of extensive competitive works on three datasets, *i.e.*, CUB-200-2011, Stanford Cars-196, and FGVC Aircraft datasets. The methods are separated into three groups from top to bottom of Tab. 3, which are (1) classification-based evaluators, (2) metric-based evaluators, and (3) the vision-language evaluator, termed as PLEor.

As shown in Tab. 3, it is obvious that the retrieval performance obtained by our evaluator is better than other methods no matter whether the classification-based or metric-based evaluators are adopted. Concretely, existing works based on classification evaluators, *i.e.*, CEP [3] and HDCL [59], tend to project the final retrieval embeddings into a category space. Despite the encouraging achievement, the shortcoming of these works is that they only focus on individual samples while neglecting the correlations among subcategories, thus limiting the retrieval performance. To address this problem, the effectiveness of these models based on metric evaluator, *i.e.*, ETLR [18] and PNCA++ [46], can be largely attributed to their precise identification of negative/positive pairs through the manipulation of distances, which indirectly enhances the discriminative power of features. However, these existing works, *e.g.*, CEP [3], HIST [25] and PNCA++ [46], follow a close-set learning setting, where all the categories are pre-defined, to learn the discriminative and generalizable embeddings for identifying the visually similar objects of unknown subcategories. It is thus very challenging for a feature extractor trained in closed-set scenarios with classification or metric supervisions to capture discriminative discrepancies from unknown subcategories, consequently impairing the retrieval performance. Compared to these works, although our PLEor also follows the close-set learning setting, it can transfer the discrepancies encompassing unknown subcategories from CLIP model to our model, thus achieving the state-of-the-art performance.

Table 6. Results of the text prompt with random and manual initialization on CUB-200-2011 dataset, respectively.

Initialization	Recall@1
Random [a_1, a_2, a_3, a_4]	72.4%
Manual ["a", "photo", "of", "a"]	72.4%

Table 7. Results comparing to various vision prompts based on Recall@K on CUB-200-2011.

Method	R@1	R@2	R@4	R@8
CAM [42]	69.8%	79.7%	84.2%	91.6%
Bounding box	73.9%	82.6%	90.5%	94.2%
Our PLEor	**74.8%**	**84.5%**	**91.3%**	**94.9%**

4.4. Further Analysis

Investigation on the length of text prompt. How many category-specific vectors in the text prompt should be used? And is it better to have more category-specific words? Here we study the impact of this hyperparameter of Eq. (4) on retrieval performance. Specifically, we repeat experiments on CUB-200-2011 dataset by varying the number of category-specific vectors from 4 to 32 with a stride of 4. The retrieval performance can be found in Tab. 4, which indicates that the performance drops when the number of vectors in the text prompt increases to 32. The possible reason of the performance drop is that after using more vectors, the vision-language evaluator may force our network to focus on more visual clues and even the useless.

Comparison of fine-tuning and pretraining schemes. Tab. 5 shows that fine-tuning pre-trained CLIP model can actually reduce retrieval accuracy compared to freezing it. This phenomenon is reasonable since fine-tuning CLIP model on the close-set scenarios could impair the ability of visual modelling in open-set scenarios due to discarding the remarkable zero-shot generalisation ability. Besides, we apply pretrained CLIP model as an evaluator, which only introduces only about 0.03 M of learnable parameters. Thereby, its parameters are less than a classification-based evaluator containing a fully connected layer.

Initialization of text prompt. In our comparison between random initialization and manual initialization, we used the embeddings of "a photo of a" to initialize the category-specific words in the latter. To ensure a fair comparison, we set the length of the learnable vectors in the text prompt to 4 when using random initialization. As shown in Tab 6, using a meaningful initialization did not result in a vital difference in performance. Although further tuning of the initialization with meaningful words may be helpful, we suggest using the random initialization method.

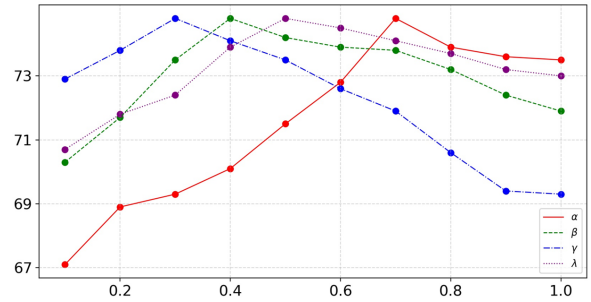


Figure 3. Analyses of hyper-parameters α, β, γ and λ in Eq. (9). The results denote Recall@1 on CUB-200-2011.

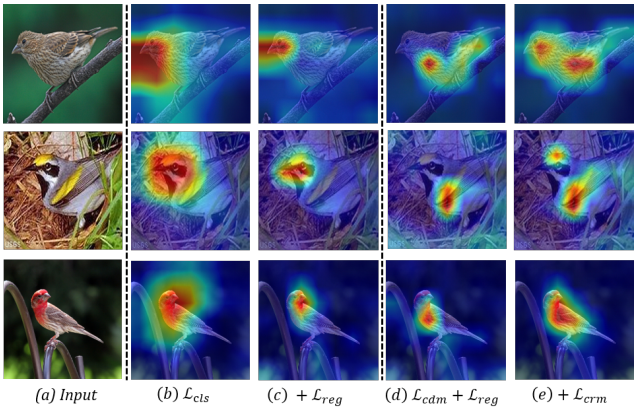


Figure 4. Visualization of vision prompt based on classification-based evaluator (b)(c) and our vision-language evaluator (d)(e), respectively. $+\mathcal{L}$ means that we successively add this constraint, *i.e.*, $+\mathcal{L}_{reg}$= $\mathcal{L}_{cls} + \mathcal{L}_{reg}$, $+\mathcal{L}_{crm}$= $\mathcal{L}_{cdm} + \mathcal{L}_{reg} + \mathcal{L}_{crm}$.

Analyses of vision prompt. By switching the processing method of input images, we can gain further insight into the effectiveness of the vision prompt scheme. As shown in Tab. 7, switching from our vision prompt strategy to the fixed prompt strategy, such as directly highlighting objects, leads to a significant drop in performance. Specifically, we apply the class activation map or the bounding boxes provided by the dataset to localize the objects from the original images. These cropped objects are regarded as fixed prompts. However, the fixed prompt strategy only provides the location of object or parts instead of the category-specific discrepancies, thus making it hard for CLIP models to capture category-specific discrepancies. Therefore, our PLEor makes the open-set retrieval task aided by the learnable prompt strategy similar to the original pre-training task, resulting in a steady improvement in performance.

Hyper-parameter analyses. The sensitivity analyses of the hyper-parameters in Eq. (9) are conducted, and the evaluation results are presented in Fig. 3. It is observed that the performance of our PLEor is a little sensitive with the variation of α, β, γ and δ. In our experiments, the default values of α, β, γ and δ are set to 0.7, 0.4, 0.3 and 0.5, respectively.

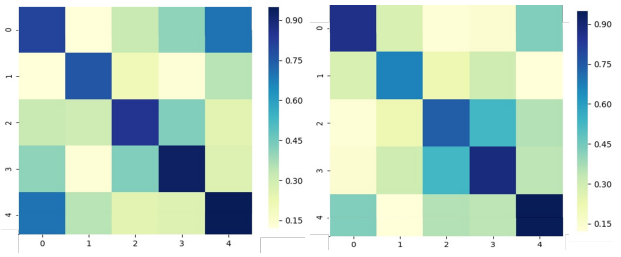


Figure 5. The nearest description for text prompt learned by PLEor, with their similarity shown in grids.

What makes a network retrieve objects visually? With this question in our mind, we exhibit the visualization results of vision prompt V_D. Since the values in V_D are continuous, we visualize them in a class activation map manner [42] to better display them. The referred results of classification-based and vision-language evaluators are shown in Fig. 4. It is shown that different constraints of our PLEor have different impact on category-specific discrepancies. Concretely, $\mathcal{L}_{reg}$ could efficiently constrain the size of activated regions, the combination of $\mathcal{L}_{cdm}$ and $\mathcal{L}_{crm}$ significantly activates more complete discrepancies.

The interpretation of the learned text prompt can be challenging due to the optimization of category-specific vectors in a continuous space. Therefore, we use an indirect method to interpret it by comparing the similarities between our text prompt and the actual descriptions. In Fig. 5, we calculate these similarities for the first 5 subcategories from both the train and test sets. We observe that the nearest descriptions to the learned text prompt are mostly the corresponding real descriptions. Overall, our text prompt learns category-specific descriptions, which guide the CLIP model to identify category-specific discrepancies.

5. Conclusion

In this paper, we propose to exploit the pre-trained CLIP model as an evaluator in place of traditional classification-based and metric-based evaluators for open-set fine-grained retrieval. The designed prompting vision-language evaluator, termed PLEor, makes the pretrained CLIP model refer the category-specific discrepancies with the appropriate prompt technologies, and transfers these discrepancies encompassing pre-defined and unknown subcategories to our model trained in close-set scenarios. This is the last tip, but the most important: This model could be end-to-end trained and gain competitive performance in three widely-used open-set fine-grained retrieval datasets.

Acknowledgements: This work is supported in part by the National Natural Science Foundation of China under Grant NO. 61976038 and NO.61932020.

References

[1] Kenan E. Ak, Ashraf A. Kassim, Joo-Hwee Lim, and Jo Yew Tham. Learning attribute representations with localization for flexible fashion search. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7708–7717. Computer Vision Foundation / IEEE Computer Society, 2018. 1

[2] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2654–2662, 2014. 3

[3] Malik Boudiaf, Jérôme Rony, Imtiaz Masud Ziko, Eric Granger, Marco Pedersoli, Pablo Piantanida, and Ismail Ben Ayed. A unifying mutual information view of metric learning: Cross-entropy vs. pairwise losses. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, volume 12351 of *Lecture Notes in Computer Science*, pages 548–564. Springer, 2020. 1, 2, 6, 7

[4] Steve Branson, Grant Van Horn, Serge J. Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. *CoRR*, abs/1406.2952, 2014. 5

[5] Mateusz Budnik and Yannis Avrithis. Asymmetric metric learning for knowledge transfer. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8228–8238. Computer Vision Foundation / IEEE, 2021. 3

[6] Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 2852–2859. AAAI Press, 2018. 3

[7] Mohamed Elhoseiny, Yizhe Zhu, Han Zhang, and Ahmed M. Elgammal. Link the head to the "beak": Zero shot learning from noisy text description at part precision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6288–6297. IEEE Computer Society, 2017. 1

[8] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomás Mikolov. Devise: A deep visual-semantic embedding model. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2121–2129, 2013. 2

[9] 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. *CoRR*, abs/2110.04544, 2021. 2

[10] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *CoRR*, abs/2202.06687, 2022. 2, 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016. 5, 6

[12] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 3, 5

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *ICML*, volume 139, pages 4904–4916, 2021. 2

[14] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *CoRR*, abs/2203.12119, 2022. 3

[15] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438, 2020. 2

[16] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. *CoRR*, abs/2112.04478, 2021. 2, 3

[17] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *CVPR*, pages 3235–3244. Computer Vision Foundation / IEEE, 2020. 6

[18] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Embedding transfer with label relaxation for improved metric learning. In *CVPR*, pages 3967–3976. Computer Vision Foundation / IEEE, 2021. 2, 6, 7

[19] Sungyeon Kim, Minkyo Seo, Ivan Laptev, Minsu Cho, and Suha Kwak. Deep metric learning beyond binary supervision. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2288–2297. Computer Vision Foundation / IEEE, 2019. 3

[20] ByungSoo Ko, Geonmo Gu, Han-Gyu Kim, and ByungSoo Ko. Learning with memory-based virtual classes for deep metric learning. In *ICCV*, pages 11772–11781. IEEE, 2021. 1, 2

[21] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pages 554–561, 2013. 5

[22] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *EMNLP*, pages 3045–3059. Association for Computational Linguistics, 2021. 2

[23] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong,

Fei Xia, Wenjie Li, and Roberto Navigli, editors, *ACL*, pages 4582–4597. Association for Computational Linguistics, 2021. 2

[24] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *CoRR*, abs/2210.08823, 2022. 3

[25] Jongin Lim, Sangdoo Yun, Seulki Park, and Jin Young Choi. Hypergraph-induced semantic tuplet loss for deep metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 212–222. IEEE, 2022. 6, 7

[26] Junfan Lin, Jianlong Chang, Lingbo Liu, Guanbin Li, Liang Lin, Qi Tian, and Chang-wen Chen. Being comes from not-being: Open-vocabulary text-to-motion generation with wordless training. *arXiv preprint arXiv:2210.15929*, 2023. 2

[27] Lizhao Liu, Shangxin Huang, Zhuangwei Zhuang, Ran Yang, Mingkui Tan, and Yaowei Wang. DAS: densely-anchored sampling for deep metric learning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVI*, volume 13686 of *Lecture Notes in Computer Science*, pages 399–417. Springer, 2022. 6

[28] Lingbo Liu, Bruce X. B. Yu, Jianlong Chang, Qi Tian, and Chang Wen Chen. Prompt-matched semantic segmentation. *CoRR*, abs/2208.10159, 2022. 3

[29] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586, 2021. 2

[30] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602, 2021. 2

[31] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1096–1104. IEEE Computer Society, 2016. 1

[32] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. 5

[33] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *MISRM'99 First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999. 2

[34] Olga Moskvyak, Frédéric Maire, Feras Dayoub, and Mahsa Baktashmotlagh. Keypoint-aligned embeddings for image retrieval and re-identification. In *WACV*, pages 676–685. IEEE, 2021. 1, 2, 3

[35] Xing Nie, Bolin Ni, Jianlong Chang, Gaomeng Meng, Chunlei Huo, Zhaoxiang Zhang, Shiming Xiang, Qi Tian, and Chunhong Pan. Pro-tuning: Unified prompt tuning for vision tasks. *CoRR*, abs/2207.14381, 2022. 3

[36] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3967–3976. Computer Vision Foundation / IEEE, 2019. 3

[37] 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 Marina Meila and Tong Zhang, editors, *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 1, 2, 3

[38] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18061–18070. IEEE, 2022. 2

[39] Karsten Roth, Timo Milbich, Björn Ommer, Joseph Paul Cohen, and Marzyeh Ghassemi. Simultaneous similarity-based self-distillation for deep metric learning. In Marina Meila and Tong Zhang, editors, *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 9095–9106. PMLR, 2021. 1, 2

[40] Karsten Roth, Oriol Vinyals, and Zeynep Akata. Non-isotropy regularization for proxy-based deep metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7410–7420. IEEE, 2022. 6

[41] Jenny Seidenschwarz. Learning intra-batch connections for deep metric learning. In Marina Meila and Tong Zhang, editors, *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 9410–9421. PMLR, 2021. 6

[42] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.*, 128(2):336–359, 2020. 7, 8

[43] Chen Shen, Chang Zhou, Zhongming Jin, Wenqing Chu, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. Learning feature embedding with strong neural activations for fine-grained retrieval. In Wanmin Wu, Jianchao Yang, Qi Tian, and Roger Zimmermann, editors, *ACM MM*, pages 424–432. ACM, 2017. 1

[44] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, pages 4004–4012. IEEE Computer Society, 2016. 5

[45] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826. IEEE Computer Society, 2016. 6

[46] Eu Wern Teh, Terrance DeVries, Graham W. Taylor, and Graham. Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors,

ECCV, volume 12369 of *Lecture Notes in Computer Science*, pages 448–464. Springer, 2020. 1, 2, 6, 7

[47] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *CoRR*, abs/2109.08472, 2021. 2

[48] Shijie Wang, Jianlong Chang, Zhihui Wang, Haojie Li, Wanli Ouyang, and Qi Tian. Fine-grained retrieval prompt tuning. *CoRR*, abs/2207.14465, 2022. 2

[49] Shijie Wang, Haojie Li, Zhihui Wang, and Wanli Ouyang. Dynamic position-aware network for fine-grained image recognition. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 2791–2799. AAAI Press, 2021. 1

[50] Shijie Wang, Zhihui Wang, Haojie Li, and Wanli Ouyang. Category-specific semantic coherency learning for fine-grained image recognition. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann, editors, *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 174–183. ACM, 2020. 1

[51] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *CVPR*, pages 5022–5030. Computer Vision Foundation / IEEE, 2019. 1, 2

[52] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Trans. Image Process.*, 26(6):2868–2881, 2017. 1, 2, 3, 6

[53] Jason Weston, Samy Bengio, and Nicolas Usunier. WSABIE: scaling up to large vocabulary image annotation. In Toby Walsh, editor, *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 2764–2770. IJCAI/AAAI, 2011. 2

[54] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: fine-grained interactive language-image pre-training. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 2

[55] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. CPT: colorful prompt tuning for pre-trained vision-language models. *CoRR*, abs/2109.11797, 2021. 3

[56] Bruce X. B. Yu, Jianlong Chang, Lingbo Liu, Qi Tian, and Chang Wen Chen. Towards a unified view on visual parameter-efficient transfer learning. *CoRR*, abs/2210.00788, 2022. 2

[57] Lu Yu, Vacit Oguz Yazici, Xialei Liu, Joost van de Weijer, Yongmei Cheng, and Arnau Ramisa. Learning metrics from teachers: Compact networks for image embedding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2907–2916. Computer Vision Foundation / IEEE, 2019. 3

[58] Li Yuan, Francis E. H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3902–3910. Computer Vision Foundation / IEEE, 2020. 3

[59] Xianxian Zeng, Shun Liu, Xiaodong Wang, Yun Zhang, Kairui Chen, and Dong Li. Hard decorrelated centralized loss for fine-grained image retrieval. *Neurocomputing*, 453:26–37, 2021. 6, 7

[60] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *CoRR*, abs/2111.03930, 2021. 2

[61] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 8542–8552. IEEE, 2022. 2

[62] Wenzhao Zheng, Chengkun Wang, Jiwen Lu, and Jie Zhou. Deep compositional metric learning. In *CVPR*, pages 9320–9329. Computer Vision Foundation / IEEE, 2021. 2, 6

[63] Wenzhao Zheng, Borui Zhang, Jiwen Lu, and Jie Zhou. Deep relational metric learning. In *ICCV*, pages 12045–12054. IEEE, 2021. 1, 2

[64] Xiawu Zheng, Rongrong Ji, Xiaoshuai Sun, Yongjian Wu, Feiyue Huang, and Yanhua Yang. Centralized ranking loss with weakly supervised localization for fine-grained object retrieval. In Jérôme Lang, editor, *IJCAI*, pages 1226–1233. ijcai.org, 2018. 1, 2, 3, 6

[65] Xiawu Zheng, Rongrong Ji, Xiaoshuai Sun, Baochang Zhang, Yongjian Wu, and Feiyue Huang. Towards optimal fine grained retrieval via decorrelated centralized loss with normalize-scale layer. In *AAAI*, pages 9291–9298. AAAI Press, 2019. 6

[66] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16795–16804. IEEE, 2022. 1, 2, 3

[67] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022. 1, 2, 3