This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Interpretable Zero-Shot Learning with Locally-Aligned Vision-Language Model

Shiming Chen[1], Bowen Duan[2], Salman Khan[1,3], Fahad Shahbaz Khan[1,4]
[1]Mohamed bin Zayed University of AI [2]Huazhong University of Science and Technology
[3]Australian National University [4]Linköping University
{shimingchen, bwduan9910}@gmail.com {salman.khan, fahad.khan}@mbzuai.ac.ae

Abstract

Large-scale vision-language models (VLMs), such as CLIP, have achieved remarkable success in zero-shot learning (ZSL) by leveraging large-scale visual-text pair datasets. However, these methods often lack interpretability, as they compute the similarity between an entire query image and the embedded category words, making it difficult to explain their predictions. One approach to address this issue is to develop interpretable models by integrating language, where classifiers are built using discrete attributes, similar to human perception. This introduces a new challenge: how to effectively align local visual features with corresponding attributes based on pre-trained VLMs. To tackle this, we propose LaZSL, a locally-aligned vision-language model for interpretable ZSL. LaZSL employs local visual-semantic alignment via optimal transport to perform interaction between visual regions and their associated attributes, facilitating effective alignment and providing interpretable similarity without the need for additional training. Extensive experiments demonstrate that our method offers several advantages, including enhanced interpretability, improved accuracy, and strong domain generalization. Codes available at: https://github.com/shiming-chen/LaZSL.

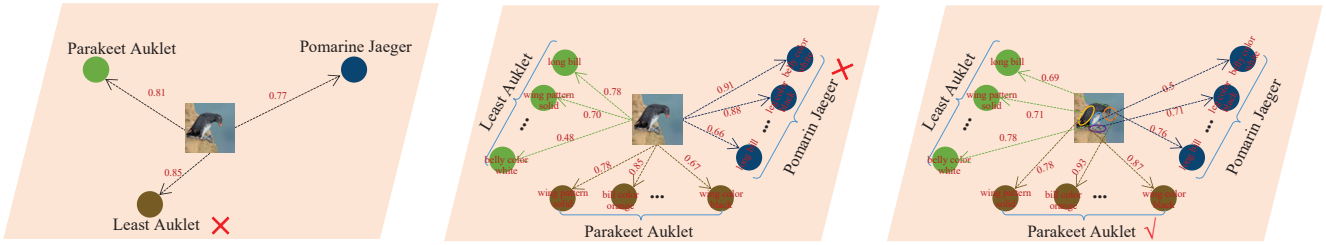
1. Introduction

Large-scale vision-language models (VLMs), such as CLIP [40], have achieved significant zero-shot learning (ZSL) by training with large amount visual-text pairs, sparking a new wave of research in ZSL [34, 51]. Specifically, CLIP performs a zero-shot classification procedure by computing the similarity between the whole query image and the embedded words for each category prompt (e.g., "*a photo of a class*"), then choosing the highest similarity for classification. Since its encoder includes rich information in the world and does not require additional annotation knowledge, it is also used in segmentation [45], detection [23], and retrieval [15] tasks.

Despite its achievements, CLIP's performance exhibits notable sensitivity to the text prompts used during the inference stage and the domain-specific datasets. Accordingly, many works focus on i) prompts learning [44, 55, 56] and ii) adapter learning [18, 24, 31]. Prompt learning discovers the domain knowledge from the downstream data for improving the generalization of CLIP. For example, Zhou introduced prompt learning using the downstream data for learning domain knowledge in the text prompt [55, 56]. Shu [44] and Feng [17] explored additional insights from the test sample itself to enhance the domain knowledge of the prompt. Adapter learning follows fine-tuning thoughts to learn lightweight parameters via additional visual-semantic interactions [24].

However, these methods compute the similarity between the whole query image and the embedded words for each category prompt following the standard CLIP, resulting in limited interpretability, as shown in Fig. 1(a). Because they cannot recognize classes based on the corresponding factors (e.g., attributes/semantics) of target classes. Meanwhile, they fail to capture fine-grained information of vision, their generalization is limited. Accordingly, a few works [12, 16, 30, 37, 42, 46] attempt to build interpretable models by integrating language, where classifiers are constructed with the discrete attributes. Specifically, they adopt large language models (LLMs) to generate multiple finer text descriptions with amounts of attribute for each category, and take these text as text prompts for computing similarity with image, as shown in Fig. 1(b). They can classify classes according to the key attributes corresponding to the target classes, and thus their classifiers have interpretability analogously to human perception, where humans recognize an object via the obvious factors (e.g., attributes). Unfortunately, these interpretable similarities are between the whole image and attributes, which cannot directly capture the relationships between fine-grained visual information and their corresponding attributes, inevitably resulting in wrong visual-semantic alignments and limiting the generalization of VLMs.

Naturally, interpretable ZSL raises a new challenge: how



（a）Similarity of CLIP　（b）Interpretable Similarity of DCLIP　（c）Interpretable Similarity of LaZSL

Figure 1. Comparison between the similarity of CLIP [40], DCLIP [30], and our LAZSL. (a) CLIP simply computes the similarity between the whole query image and the embedded words for each category, failing to explain their predictions. (b) DCLIP builds interpretable similarity based on alignments between whole query images and attributes, obtaining limited generalization. (c) Our LaZSL represents the optimally interpretable similarity between local visual regions and attributes.

to conduct effective alignment between local vision and attributes based on the pre-trained VLMs. Different from classical ZSL methods [6, 7, 29, 33] that can take attention mechanism in the network backbone to learn attribute localization for local alignment between visual regions and human annotated attributes, the network of VLMs are frozen and not easy to re-design for training as there are not sufficient data. Motivated by optimal transport (OT) theory [28, 36, 47], we can divide each image as a set of local patches over the visual space and view each attribute set of one class as a discrete distribution over the semantic space. With such formulation, the classification task becomes to measure the distance between the two distributions of visual and semantic spaces, as shown in Fig. 1(c). The plan of OT is then calculated between the local visual features and attribute features, enabling the local alignments. Accordingly, it is possible to predict the label with the detailed attributes and patch features, resulting in more effective alignment and classification accuracy.

In this paper, we propose a locally-aligned vision-language model for interpretable ZSL, dubbed LaZSL. LaZSL first constructs the vision and semantic sets by randomly cropping the image and LLM, respectively. Then, LaZSL adopts local visual-semantic alignment via OT to formulate the interaction between the constructed visual sets and semantic sets, enabling effective alignments by optimizing the OT plan. Notably, we also incorporate the global visual information into the hybrid cost matrix to avoid knowledge forgetting in the pre-trained VLMs. Finally, we can predict their classes of input images by aligning hybrid similarity matrix with OT plan. The extensive experiments on nine widely-used data demonstrate our method offers several advantages, i.e., interpretability, improvements in accuracy, and good domain generalization.

Our main contributions are summarized in the following:

- We propose LaZSL to conduct effective alignment between local vision and attributes based on the pre-trained VLMs (e.g., CLIP) for interpretable ZSL. Different from most variants of CLIP that require additional model training, our LaZSL is training-free.
- We introduce local visual-semantic alignment using optimal transport to formulate interaction between the constructed visual regions and attributes, enabling effective alignment and obtaining interpretable similarity for ZSL prediction.
- We conduct extensive experiments on nine widely-used datasets to evaluate our methods, and results demonstrate that our LaZSL achieves competitive performances over baselines.

2. Related Works

Classical Zero-Shot Learning: Early ZSL methods utilize the human annotated attributes as side-information for knowledge transfer from seen classes to unseen ones [34, 51]. They target how to conduct effective visual-semantic interactions for knowledge transfer. Typically, there are three types of interactions: embedding-based methods, generative methods, and common space learning methods. Embedding-based methods map visual features into semantic space and search nearest-neighbor semantic prototypes for classification [1, 7, 9, 21]. Generative methods learn a semantic-conditioned generator to synthesize image/feature samples for unseen classes and transform the ZSL task as supervised classification [8, 10, 11, 52]. Common-space learning methods map the visual and semantic features into a common space and perform classification by nearest-neighbor search [5, 43, 49]. Although these methods have interpretability with the human-annotated attributes, they are time-consuming and labor-intensive to collect for various scene generations. Thus, they cannot work well on the large-scale dataset, e.g., ImageNet.

VLM-Based Zero-Shot Learning: VLMs take large-scale visual-text pairs for model training and have significant knowledge transfer capacity [22, 40]. For example, CLIP leads new trends in ZSL tasks [40]. Despite its advances,

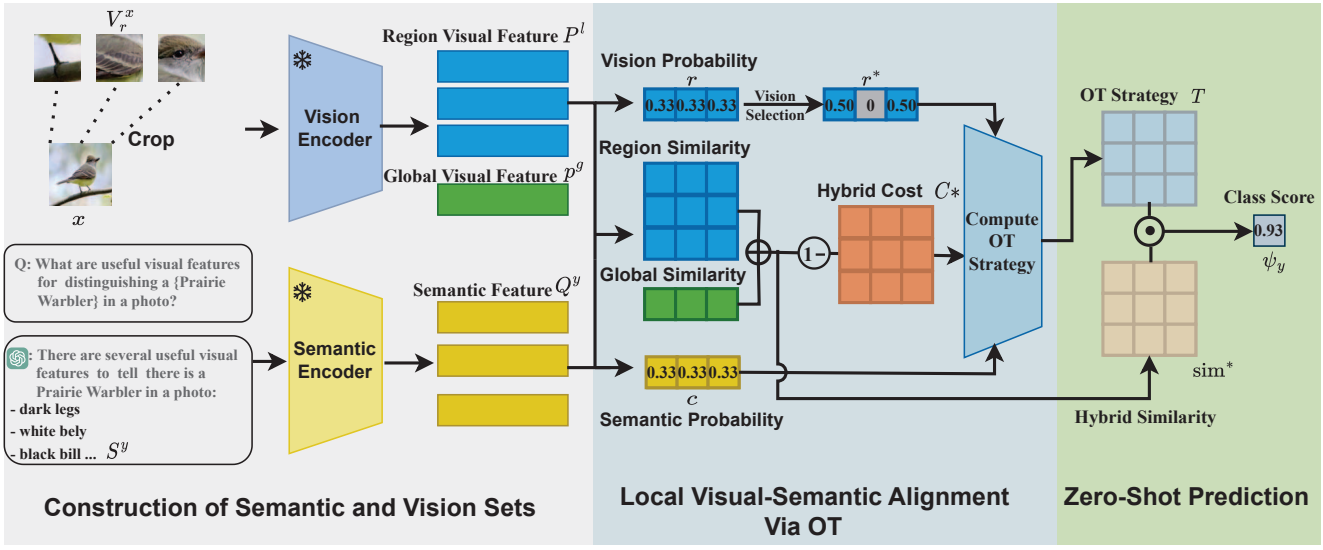


Figure 2. The pipeline of our LaZSL. LaZSL consists of three main module components, including the construction of semantic and vision sets, local visual-semantic alignment via OT, and zero-shot prediction.

CLIP heavily relies on prompt design and obtains limited performances in the domain-specific datasets [27, 55]. Many works attempt to improve CLIP by prompt learning [24, 53, 55], adapter learning [24, 31]. However, these methods compute the similarity between the whole query images and the word embedding of class names, which lacks interpretability. Recently, a few works [12, 16, 30, 37, 42, 46] build interpretable models by integrating language generated by LLMs, where classifiers can recognize target classes according to whether the object has corresponding key attributes analogously to human conception. However, they cannot well capture the relationships of local visual information with the corresponding attributes, resulting in wrong alignments between visual and semantic space and hampers the generalization capacity of VLMs. To tackle this challenge, we design a locally-aligned vision-language model to conduct effective alignment between local vision and attributes based on the pre-trained VLMs.

Optimal Transport in Vision: The Optimal Transport (OT) theory is initially introduced to solve the problem of how to reduce the cost when moving several items simultaneously [36]. Thanks to the property of distribution matching, OT has been widely applied in many computer vision tasks [2, 25, 26]. More related to our works, some works [4, 28, 47] take OT for prompt tuning. For example, Chen *et al.*[4] and Li *et al.*[28] utilize OT to learn single- or multi-mode prompts at the same time to optimize text prompts. Differently, we take OT to find better alignments between the local visual sets and attribute sets by local alignment and thus obtain accurate similarity for classification.

3. Locally-Aligned Vision-Language Model

In this section, we introduce our locally-aligned vision-language model for interpretable zero-shot learning (dubbed LaZSL), which performs effective alignment between local vision and attributes by optimizing the interactions of semantic and visual features using OT. LaZSL can accurately capture the local fine-grained visual information according to its corresponding attributes for class prediction. Thus, it achieves interpretability, accuracy improvements, and good domain generalization. As shown in Fig. 2, LaZSL consists of three main components: the construction of semantic and visual sets, the visual-semantic interaction at the attribute level based on OT, and the zero-shot prediction. We will introduce their details in the following subsections.

3.1. Construction of Semantic and Vision Sets

Although CLIP variants based on global features can handle ZSL tasks effectively, they often struggle with fine-grained classification or categories with semantically sparse names. Accordingly, enriching CLIP models with local attributes to capture fine-grained visual information is necessary. Furthermore, formulating an attribute classifier has the property of interpretability [30]. Targeting this goal, we first construct the semantic and vision sets.

Construction of Semantic Sets. To introduce semantic information at the attribute level, we follow [30] to construct a semantic set for each category with the help of LLMs. Briefly, given a class label y, the semantic sets S^y can be

generated:

$$S^y = h(prompt(y)) = \{s_i^y | i = 1, \dots, M\}, \quad (1)$$

where $h(\cdot)$ is an LLM (e.g., GPT-3) and $prompt(\cdot)$ is a template function that generates queries for LLMs by utilizing fixed prompts.

In fact, after obtaining the semantic set for each category, most interpretable ZSL methods simply compute the category score by averaging the similarity between global visual features and the attributes in the semantic set [30, 37, 42]. While this approach can partially alleviate the deficiencies caused by missing attribute semantics, it essentially performs global alignment with attribute information rather than achieving true attribute-level visual-semantic interaction. To address this, we further construct a visual set to capture local visual information, which is then used for local visual-semantic interaction.

Construction of Vision Sets. Specifically, we attempt to capture visual regions corresponding to the semantic set by performing random multi-scale cropping on the original images. Given an image $x \in \mathbb{R}^{H \times W \times 3}$, where H and W are its height and width respectively, we propose $P_r(\cdot, \cdot)$ to construct region vision sets:

$$V_r^x = \{v_i^x = P_r(x, \gamma_i min(W, H)) | i = 1, \dots, N\}, \quad (2)$$

where γ_i is a random number drawn from the uniform distribution $U(\alpha, \beta)$, where α and β are predefined hyperparameters that constrain the lower and upper bounds of the sampling range respectively. N is the number of cropped regions and typically set to $[60 - 90]$. The function $P_r(\cdot, \cdot)$ performs a random region cropping on an image, with γ_i specifying the scale of the cropped regions. It can be seen that $P_r(\cdot, \cdot)$ and γ_i ensure that the constructed region vision set contains visual features of various scales. This enables our visual set to better match the previously constructed attribute set.

3.2. Local Visual-Semantic Alignment Via OT

After constructing the visual and semantic sets, we use CLIP encoders to obtain their latent space representations $P_x = \{p_i\}_{i=1}^N$ and $Q_y = \{q_j\}_{j=1}^M$, respectively:

$$P^x = P^l \cup \{p^g\} = \sigma_v(V_r^x) \cup \{\sigma_v(x)\}, \quad (3)$$

$$Q^y = \sigma_s(S^y), \quad (4)$$

where, σ_v and σ_s are the vision encoder and semantic encoder from CLIP,respectively.

Unlike existing interpretable ZSL methods [30, 37, 42] that perform visual-semantic alignment using global features (1 to 1), our approach constructs local visual and semantic sets (N to M). Simply averaging multiple similarities is insufficient to fully estimate the similarity between the two sets. Therefore, we propose an OT matching method.

The optimal transport problem can be viewed as finding the minimum cost to transform one probability distribution into another. Since this cost can serve as a measure of the distance between two probability distributions, we employ OT theory to match the vision set and the semantic set. The process to compute OT is formulated as:

$$T = \mathrm{OT}(C, r, c), \quad (5)$$

where T is the OT plan, r and c are the discrete probability vectors belonging to the vision set and semantic set respectively, and can be initialized as uniform distributions. C is the cost matrix between the region vision set and the semantics set, and $C_{i,j} = 1 - \mathrm{sim}_{i,j} = 1 - \mathrm{cosine}(p_i^l, q_j^y)$

It can be seen that, by using the OT, we have found an effective way to evaluate the similarity between the vision set and the semantic set. However, this algorithm still has two limitations: i) the randomly obtained region visual features contain some noises, and ii) the region visual features may cause the knowledge forgetting of CLIP visual encoder and cannot well capture their corresponding categories. To address these issues, we propose a vision selection mechanism and a region-global hybrid cost approach.

Vision Selection. Specifically, the original region visual features are divided into a relevant region set and an irrelevant region set based on a threshold δ, which is obtained according to the average similarity of all region visions and the global image. Then, we remove irrelevant noises by modifying the initialization of r that is relevant to δ. This process can be expressed as:

$$\delta = \mathrm{AVG}(\sum_{i=1}^{N} \mathrm{cosine}(p_i^l, p^g)), \quad (6)$$

$$P_{pos}^l = \{p_i^l | \mathrm{cosine}(p_i^l, p^g) \geq \delta\}, \quad (7)$$

$$P_{neg}^l = \{p_i^l | \mathrm{cosine}(p_i^l, p^g) < \delta\}, \quad (8)$$

$$r_i^* = \begin{cases} \frac{1}{|P_{pos}^l|} & \text{if } p_i^l \in P_{pos}^l, \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

Thus, the visual probability r is updated as r^* according to Eq. 9. Then, the positive region vision set $p_i^l \in P_{pos}^l$ is used for visual-semantic interaction via OT.

Region-Global Hybrid Cost. In OT, the cost matrix is a crucial source of prior knowledge. In the visual-semantic interaction based on OT, we observe that our cost is solely composed of the similarity between the randomly cropped region vision set and the semantic set. This inevitably makes the OT strategy T, computed from the cost, overly sensitive to the noise introduced by the cropping process, resulting in knowledge forgetting of the vision encoder in CLIP. Therefore, we choose to incorporate additional global prior information into the cost matrix to address this issue:

$$C_i^* = 1 - (\theta \mathrm{sim}_i + (1 - \theta) p^{g\top} Q^y) \quad (10)$$

where C_i^* is the i-th row of the hybrid cost matrix C^*, $\theta \in (0, 1)$ is a hyper-parameter, which can be seen as the hybrid confidence between region feature and global feature.

Compute OT Strategy. Once we have obtained C^*, we can compute the optimal transport plan T. Specifically, we choose the Sinkhorn algorithm [13] to solve the optimal transport distance, which uses an entropic constraint for fast optimization. By iteratively updating the strategy matrix between the region vision set and semantic set, Eq. 5 is specifically formulated as:

$$T = \text{diag}(\mathcal{U})\mathcal{M}\text{diag}(\mathcal{V}), \tag{11}$$

$$u_i^{k+1} = \frac{r_i^*}{\sum_j \mathcal{M}_{i,j} v_j^k}, \tag{12}$$

$$v_j^{k+1} = \frac{c_j}{\sum_i \mathcal{M}_{i,j} u_i^k}, \tag{13}$$

$$\mathcal{M} = \exp(\frac{-C^*}{\lambda}), \tag{14}$$

where T is the OT plan, r^* and c are the discrete probability vectors belonging to the vision set and semantic set. C^* is the hybrid cost matrix between the region vision set and the semantics set, k is the number of iterations for T, $\text{diag}(\cdot)$ is a function used to construct a diagonal matrix, $\mathcal{U} = \{u_i\}_{i=1}^N$ and $\mathcal{V} = \{v_j\}_{j=1}^M$.

3.3. Zero-Shot Prediction

After the visual-semantic interaction, we get the OT plan T. Similarly, to align with the cost matrix, we calculate the category scores using a hybrid similarity approach:

$$\psi_y = < T, \text{sim}^* >_F, \tag{15}$$

$$\text{sim}_i^* = \theta \text{sim}_i + (1 - \theta) p^{g\top} Q^y, \tag{16}$$

where sim_i^* is the i-th row of the hybrid similarity matrix sim^*, $< \cdot, \cdot >_F$ is the Frobenius inner product between two matrix. It can be observed that for each category y, we can obtain its corresponding category score. Next, we use this score to perform zero-shot predictions:

$$y^* = \arg\max_y \psi_y, \tag{17}$$

where y^* represents the predicted category label.

4. Experiments

4.1. Experiment Setup

Datasets. To make a comprehensive evaluation of our work, we conduct extensive ZSL experiments on cross-dataset transfer learning and domain generalization. These experiments are evaluated on nine widely used image datasets, varying in scale and domains. For example, ImageNet [14] is used for recognizing daily objects; CUB [50] is used for fine-grained classification of birds, Oxford Pets [35] is used for recognizing common animals, Food101 [3] is specifically designed for food classification, Place365 [54] is applied to scene recognition. We also take the variant datasets of ImageNet with natural domain shifts to validate the domain generalization capacity of LaZSL, including, ImageNet-V2 [41], ImageNet-Sketch [48], ImageNet-A [20] and ImageNet-R [19], each dataset represents a unique distribution shift from ImageNet.

Compared Baselines. Since our method is VLM-based ZSL, we primarily take VLM-based ZSL methods for fair comparison on various network backbones, i.e., CLIP models with ViT-B/32, ViT-B/16, ViT-L/14. Compared baselines includes the standard VLM-based ZSL methods (i.e., CLIP [40], CoOp [56], CoCoOp [55], TPT [44], MaPLe [24], ProGrad [57]) and interpretable VLM-based ZSL methods (i.e., DCLIP [30], WaffleCLIP [42], CuPL [37], ArGue [46]).

Implementation Details. In this work, we mainly follow [30] to take GPT-3 to generate the attribute sets for each class. Specifically, we take the following prompts to generate attribute descriptions for each class:

Q: What are useful features for distinguishing a {class name} in a photo?

A: There are several useful visual features to tell there is a {class name} in a photo:

–

To ensure more consistent output from LLM, we additionally included two specific examples in our prompt. The text prompt template of text encoder for classification is "*A photo of a {class name}, which (is/has/etc) {attribute}*". During visual sets construction, we set the cropping scale to 0.6 and each image is typically cropped into $[60, 90]$ region patches. The hybrid coefficient θ is set to 0.8 for all datasets. All experiments are performed on a single NVIDIA H100 graphic card with 80GB memory.

4.2. Comparison with State of the Arts

Results on Cross Datasets. We first compared our LaZSL with the interpretable ZSL methods (i.e., DCLIP [30], WaffleCLIP [42], CuPL [37]) on five image classification datasets. Results are shown in Table 1. We find that LaZSL achieves the best average performances across five datasets when using various network backbones, i.e., 66.8%, 69.7%, and 74.0% average accuracy on ViT-B/32, ViT-B/16, and ViT-L/14, respectively. Compared to DCLIP [30] that is the first interpretable ZSL method based on VLM, LaZSL consistently obtains performance gains on all datasets with various network backbones. Especially, the performance gains are larger on the challenging fine-grained datasets (e.g., CUB, Place365). Because aligning key fine-grained visual details is important for the alignment of vision and

Table 1. Comparison of ZSL performance (accuracy in %) across five image classification benchmarks (i.e., ImageNet, CUB, Oxford Pets, Food101, and Place365) using three different CLIP models (ViT-B/32, ViT-B/16, ViT-L/14). △ DCLIP denotes the improvements of our LaZSL over DCLIP. For each group, the best results are marked with **Bold**.

Method	ImageNet	CUB	Oxford Pets	Food101	Place365	Average
	ViT-B / 32 with pre-trained weights from CLIP					
CLIP [40]	62.1	51.2	85.0	82.6	38.5	63.3
DCLIP [30]	63.0	52.7	84.5	84.1	39.9	64.8
WaffleCLIP [42]	63.3	52.0	85.5	84.0	39.5	64.9
CuPL [37]	64.4	49.8	87.0	84.2	39.1	64.9
ProAPO [39]	64.1	53.6	**88.7**	84.2	**42.7**	66.7
LaZSL (Ours)	**65.3**	**56.5**	84.7	**85.9**	41.5	**66.8**
△ DCLIP	+2.2	+3.8	+0.2	+1.8	+1.6	+2.0
	ViT-B / 16 with pre-trained weights from CLIP					
CLIP [40]	66.7	56.0	88.1	88.4	39.3	67.7
DCLIP [30]	67.9	57.1	86.9	88.5	40.3	68.1
WaffleCLIP [42]	68.1	56.9	86.5	89.1	40.8	68.3
CuPL [37]	**69.6**	56.4	**91.1**	89.0	39.8	69.2
LaZSL (Ours)	69.2	**60.3**	87.4	**89.7**	**42.0**	**69.7**
△ DCLIP	+1.3	+3.1	+0.52	+1.2	+1.7	+1.6
	ViT-L / 14 with pre-trained weights from CLIP					
CLIP [40]	73.5	62.1	93.2	92.6	39.6	72.2
DCLIP [30]	74.9	63.5	92.4	93.0	40.3	72.8
WaffleCLIP [42]	75.3	62.3	91.6	93.3	40.9	72.7
CuPL [37]	**76.6**	62.2	**94.3**	93.4	40.8	73.5
LaZSL (Ours)	75.7	**66.1**	92.7	**93.5**	**41.8**	**74.0**
△ DCLIP	+0.8	+2.6	+0.3	+0.5	+1.5	+1.2

attributes. These results reveal that LaZSL is effective in conducting alignment between local vision and attributes based on the pre-trained VLM for interpretable ZSL.

Results on Cross Domains. We also evaluate the domain generalization capacity of LaZSL on the variants of ImageNet with natural domain shifts, as shown in Table 2. Compared to the standard VLM-based methods (e.g., CoOp [56], CoCoOp [55], TPT [44], MaPLe [24], ProGrad [57]) that require additional training for prompt or models, LaZSL advances the interpretable ZSL methods to obtain state-of-the-art average performance (i.e., 60.9%) over all datasets. This reveals that the interpretable classifier with attributes has great potential for ZSL by effective local alignment without additional training. When comparing with other interpretable ZSL methods, LaZSL achieves performance gains of 1.7% and 0.9% over DCLIP [30] and CuPL [37], respectively. LaZSL even outperforms ArGue [46] which is interpretable ZSL method tuning prompts with additional training. These should be thanks to the local visual-semantic alignment via OT, enabling to learn interpretable classifiers that can accurately capture the local visual information corresponding to their attributes.

4.3. Ablation Study

To provide further insight into LaZSL, we conduct ablation studies to evaluate the effects of various model components, including locally visual-semantic alignment via OT (denoted as "OT"), vision selection (denoted as "VS"), and hybrid the local and global visual information (denoted as "Hybrid"). Results on three datasets are presented in Table 3. The baseline is the typical interpretable ZSL method DCLIP [30]. When baseline using OT to conduct local visual-semantic alignment, its performances consistently improved by 0.6%, 1.2%, and 1.3% on ImageNet, CUB,

Table 2. ZSL performances (accuracy in %) on ImageNet with natural distribution shifts of the state-of-the-art methods. ViT-B/16 is used as a network backbone for fair comparison. Methods take the source data for training marked with ✔, otherwise, marked with ✗. † denotes interpretable ZSL with attributes generated by LLMs. For each group, the best results are marked with **Bold**.

Method	Training	ImageNet-V2	ImageNet-R	ImageNet-S	ImageNet-A	Average
CLIP [40]	✗	60.8	74.0	47.8	46.1	57.2
CoOp [56]	✔	64.2	75.2	47.9	49.7	59.3
CoCoOp [55]	✔	64.1	76.2	48.8	50.6	59.9
TPT [44]	✔	64.3	73.9	46.4	53.6	59.5
MaPLe [24]	✔	64.1	77.0	**49.2**	50.9	60.3
ProGrad [57]	✔	**64.7**	74.5	47.9	49.3	59.1
DCLIP† [30]	✗	61.6	75.0	47.1	49.2	58.2
CuPL† [38]	✗	63.3	**77.1**	48.8	50.8	60.0
SaLS [32]	✔	64.2	74.4	46.3	48.5	58.4
ArGue† [46]	✔	64.6	76.6	48.9	50.9	60.3
LaZSL† (Ours)	✗	63.3	75.6	48.2	**56.2**	**60.9**

Table 3. Ablation studies for different components of our LaZSL on three datasets. "OT" denotes local visual-semantic alignment via OT, "VS" denotes vision selection, "Hybrid" denotes hybrid the local and global visual information. We conduct these experiments using the CLIP model of ViT-B/16.

Method	ImageNet	CUB	Place365
Baseline	67.9	57.8	40.3
Baseline+OT	68.5	59.0	41.6
Baseline+OT+VS	69.0	60.0	41.8
Baseline+OT+Hybrid	69.0	59.3	41.9
LaZSL (full)	**69.2**	**60.3**	**42.0**

and Place365, respectively. This demonstrates it's necessary to perform local alignment in interpretable ZSL and our method can effectively advance it. Furthermore, the visual selection and hybrid local and global visual information can further enable LaZSL to conduct accurate local visual regions with their corresponding attributes. Thus, LaZSL can learn an accurate and interpretable classifier for ZSL based on the pre-trained VLM.

4.4. Qualitative Analysis

Interpretable Class Prediction of LaZSL. As shown in Fig. 3, we show the classifier of our LaZSL on various samples. For example, LaZSL recognizes the first image as "Scissor Tailed Flycatcher", as it formulates the classifier with attributes "*black wings with white wingbars*", "*black head with a white throat*", "*a long, forked tail*", which are important factors of "Scissor Tailed Flycatcher". Different from standard CLIP which only computes a similarity between the whole image and prompts of class names, LaZSL can accurately predict classes according to the corresponding attributes. That is, attributes are the interpretable factors for our method.

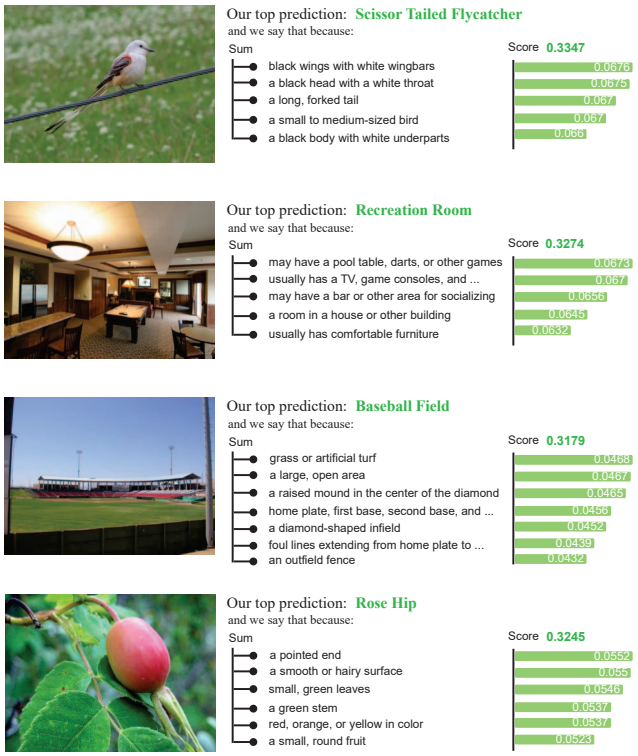


Figure 3. Visualization of interpretable classification of LaZSL.

Classification Comparison between LaZSL and DCLIP. To intuitively show the effectiveness of local alignment in interpretable ZSL, we show the top-5 classification results

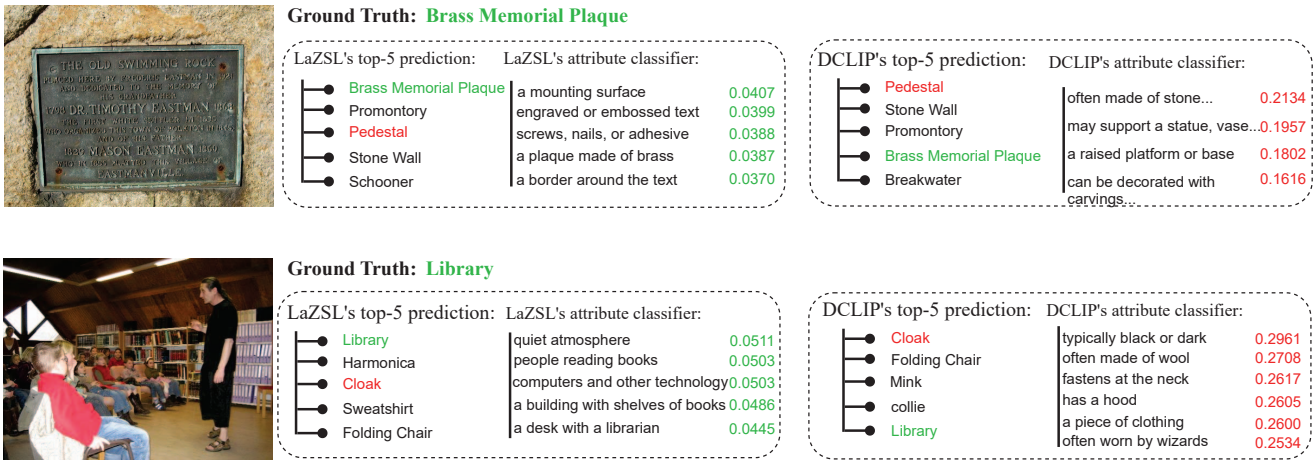
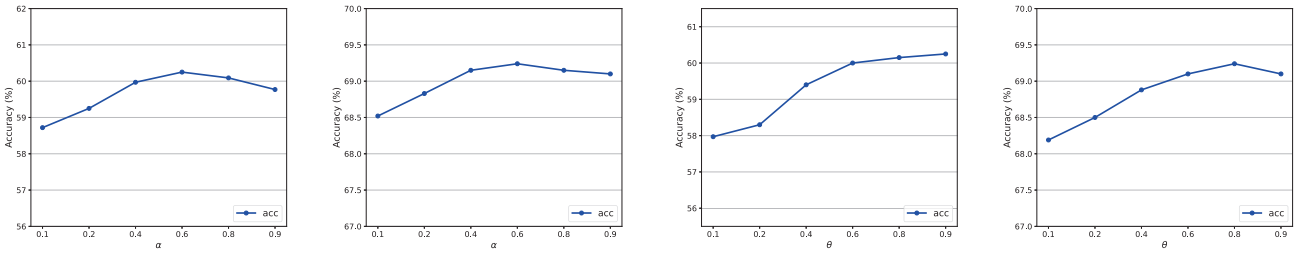


Figure 4. Classification comparison between LaZSL and DCLIP [30].



(a) Varying α on CUB (b) Varying α on ImageNet (c) Varying θ on CUB (d) Varying θ on ImageNet

Figure 5. Hyper-parameters analysis. We show the ZSL performance variations on CUB and ImageNet by adjusting the value of cropping scale α in (a, b), the value of hybrid coefficient θ in (c, d).

of LaZSL and DCLIP [30]. Part results are shown in Fig. 4. We can find that LaZSL accurately predicts classes for images, while DCLIP fails. Especially for the challenging fine-grained images, there is little difference between various fine-grained classes, e.g., "Brass Memorial Plaque" and "Pedestal". Because LaZSL effectively aligns the fine-grained visual regions with their corresponding attributes via OT, which formulates a strong classifier for class prediction.

4.5. Hyper-Parameter Analysis

Cropping Scale α. We study the cropping scale α to determine its effectiveness on our LaZSL. As can be seen from Fig. 5(a)(b), LaZSL is not sensitive to α, and it achieves best performances on all datasets when α is set to 0.6. Additionally, when α is set to too large (e.g., $\alpha \geq 0.6$), the performance of LaZSL decreases as it cannot conduct accurate local vision regions with attributes. Accordingly, we experimentally set the $\alpha = 0.6$ for all datasets.

Hybrid Coefficient θ. We study the hybrid coefficient θ to determine its effectiveness on our LaZSL. As illustrated in Fig. 5(c)(d), the performance of LaZSL improves as local visual information is progressively integrated into visual-semantic alignment and class prediction. This indicates that local alignment is crucial for interpretable ZSL. However, if θ is set too high (e.g., $\theta \geq 0.9$), performance may degrade. This is because excessive local information can lead to knowledge forgetting in the visual encoder of CLIP. Based on our experiments, we set $\theta = 0.8$ to fuse key local visual features with a few global features in LaZSL.

5. Conclusion

In this paper, we address the visual-semantic alignment challenge in interpretable VLM-based ZSL. To this end, we propose LaZSL, which leverages local visual-semantic alignment through Optimal Transport to effectively align fine-grained visual information with semantic attributes. This enables LaZSL to learn an accurate and interpretable classifier for ZSL, achieving strong performance and domain generalization. Notably, in contrast to most existing VLM-based ZSL methods that require additional training for model fine-tuning or prompt adjustment, LaZSL enhances the ZSL performance of CLIP without any extra training. Both qualitative and quantitative results demonstrate the effectiveness of LaZSL.

References

[1] Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(7):1425–1438, 2016. 2

[2] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017. 3

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *ECCV*, pages 446–461, 2014. 5

[4] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. PLOT: prompt learning with optimal transport for vision-language models. In *ICLR*, 2023. 3

[5] Shiming Chen, Guo-Sen Xie, Yang Yang Liu, Qinmu Peng, Baigui Sun, Hao Li, Xinge You, and Ling Shao. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. In *NeurIPS*, 2021. 2

[6] Shiming Chen, Ziming Hong, Guo-Sen Xie, Wenhan Yang, Qinmu Peng, Kai Wang, Jian Zhao, and Xinge You. MSDN: mutually semantic distillation network for zero-shot learning. In *CVPR*, pages 7602–7611, 2022. 2

[7] Shiming Chen, Ziming Hong, Wenjin Hou, Guo-Sen Xie, Yibing Song, Jian Zhao, Xinge You, Shuicheng Yan, and Ling Shao. Transzero++: Cross attribute-guided transformer for zero-shot learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):12844–12861, 2023. 2

[8] Shiming Chen, Wen Qing Hou, Ziming Hong, Xiaohan Ding, Yibing Song, Xinge You, Tongliang Liu, and Kun Zhang. Evolving semantic prototype improves generative zero-shot learning. In *ICML*, 2023. 2

[9] Shiming Chen, Wenjin Hou, Salman H. Khan, and Fahad Shahbaz Khan. Progressive semantic-guided vision transformer for zero-shot learning. In *CVPR*, pages 23964–23974, 2024. 2

[10] Shiming Chen, Dingjie Fu, Salman H. Khan, and Fahad Shahbaz Khan. Genzsl: Generative zero-shot learning via inductive variational autoencoder. 2025. 2

[11] Shiming Chen, Ziming Hong, Xinge You, and Ling Shao. Semantics-conditioned generative zero-shot learning via feature refinement. *International Journal of Computer Vision*, 2025. 2

[12] Mia Chiquier, Utkarsh Mall, and Carl Vondrick. Evolving interpretable visual classifiers with large language models. In *ECCV*, 2024. 1, 3

[13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, pages 2292–2300, 2013. 5

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 5

[15] Yongchao Du, Min Wang, Wen gang Zhou, Shuping Hui, and Houqiang Li. Image2sentence based asymmetrical zero-shot composed image retrieval. In *ICLR*, 2024. 1

[16] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. In *NeurIPS*, 2023. 1, 3

[17] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *ICCV*, pages 2704–2714, 2023. 1

[18] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Jiao Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 2023. 1

[19] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8320–8329, 2021. 5

[20] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021. 5

[21] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *CVPR*, pages 4482–4492, 2020. 2

[22] 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 *ICML*, pages 4904–4916, 2021. 2

[23] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Multi-modal classifiers for open-vocabulary object detection. In *ICML*, 2023. 1

[24] Muhammad Uzair Khattak, Hanoona Abdul Rasheed, Muhammad Maaz, Salman H. Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023. 1, 3, 5, 6, 7

[25] Nicholas I. Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *CVPR*, pages 10051–10060, 2019. 3

[26] Charlotte Laclau, Ievgen Redko, Basarab Matei, Younès Bennani, and Vincent Brault. Co-clustering through optimal transport. In *ICML*, pages 1955–1964, 2017. 3

[27] Jinhao Li, Haopeng Li, Sarah Monazam Erfani, Lei Feng, James Bailey, and Feng Liu. Visual-text cross alignment: Refining the similarity score in vision-language models. In *ICML*, 2024. 3

[28] Miaoge Li, Dongsheng Wang, Xinyang Liu, Zequn Zeng, Ruiying Lu, Bo Chen, and Mingyuan Zhou. Patchct: Aligning patch set and label set with conditional transport for multi-label image classification. In *ICCV*, pages 15302–15312, 2023. 2, 3

[29] Man Liu, Feng Li, Chunjie Zhang, Yunchao Wei, Huihui Bai, and Yao Zhao. Progressive semantic-visual mutual adaption for generalized zero-shot learning. In *CVPR*, pages 15337–15346, 2023. 2

[30] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *ICLR*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[31] Balamurali Murugesan, Julio Silva-Rodríguez, Ismail Ben Ayed, and José Dolz. Robust calibration of large vision-language adapters. In *ECCV*, 2024. 1, 3

[32] Balamurali Murugesan, Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. Robust calibration of large vision-language adapters. In *ECCV*, pages 147–165, 2024. 7

[33] Muhammad Ferjad Naeem, Yongqin Xian, Luc Van Gool, and Federico Tombari. I2dformer+: Learning image to document summary attention for zero-shot image classification. *Internal Journal of Computer Vision*, 132(9):3806–3822, 2024. 2

[34] Mark Palatucci, D. Pomerleau, Geoffrey E. Hinton, and Tom Michael Mitchell. Zero-shot learning with semantic output codes. In *NeurIPS*, pages 1410–1418, 2009. 1, 2

[35] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505, 2012. 5

[36] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Found. Trends Mach. Learn.*, 11(5-6):355–607, 2019. 2, 3

[37] Sarah Pratt, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, pages 15645–15655, 2022. 1, 3, 4, 5, 6

[38] Sarah M. Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, pages 15645–15655, 2023. 7

[39] Xiangyan Qu, Gaopeng Gou, Jiamin Zhuang, Jing Yu, Kun Song, Qihao Wang, Yili Li, and Gang Xiong. Proapo: Progressively automatic prompt optimization for visual classification. In *CVPR*, 2025. 6

[40] 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*, pages 8748–8763, 2021. 1, 2, 5, 6, 7

[41] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400, 2019. 5

[42] Karsten Roth, Jae-Myung Kim, A. Sophia Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. In *ICCV*, pages 15700–15711, 2023. 1, 3, 4, 5, 6

[43] Edgar Schönfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero- and few-shot learning via aligned variational autoencoders. In *CVPR*, pages 8247–8255, 2019. 2

[44] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022. 1, 5, 6, 7

[45] Feilong Tang, Zhongxing Xu, Zhaojun Qu, Wei Feng, Xingjian Jiang, and Zongyuan Ge. Hunting attributes: Context prototype-aware learning for weakly supervised semantic segmentation. In *CVPR*, pages 3324–3334, 2024. 1

[46] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Argue: Attribute-guided prompt tuning for vision-language models. In *CVPR*, pages 28578–28587, 2024. 1, 3, 5, 6, 7

[47] Dongsheng Wang, Miaoge Li, Xinyang Liu, Mingsheng Xu, Bo Chen, and Hanwang Zhang. Tuning multi-mode token-level prompt alignment across modalities. In *NeurIPS*, 2023. 2, 3

[48] Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, pages 10506–10518, 2019. 5

[49] Qian Wang and Ke Chen. Zero-shot visual recognition via bidirectional latent embedding. *Internal Journal of Compution Vision*, 124(3):356–383, 2017. 2

[50] P. Welinder, S. Branson, T. Mita, C. Wah, Florian Schroff, Serge J. Belongie, and P. Perona. Caltech-ucsd birds 200. *Technical Report CNS-TR-2010-001, Caltech,*, 2010. 5

[51] Yongqin Xian, B. Schiele, and Zeynep Akata. Zero-shot learning — the good, the bad and the ugly. *CVPR*, pages 3077–3086, 2017. 1, 2

[52] Yongqin Xian, T. Lorenz, B. Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551, 2018. 2

[53] Sheng Zhang, Salman H. Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *CVPR*, pages 3479–3488, 2023. 3

[54] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, 2018. 5

[55] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16795–16804, 2022. 1, 3, 5, 6, 7

[56] 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, 5, 6, 7

[57] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *ICCV*, pages 15613–15623, 2023. 5, 6, 7