Improving Zero-Shot Generalization for CLIP with Variational Adapter

Ziqian Lu[1], Fengli Shen[2], Mushui Liu[1], Yunlong Yu[1]*, and Xi Li[1]

[1] Zhejiang University, Hangzhou 310007, China
[2] Yangtze Delta Region Institute (Huzhou) University of Electronic Science and Technology of China, Huzhou 313001, China
{ziqianlu,lms,yuyunlong,xilizju}@zju.edu.cn
fenglishen@csj.uestc.edu.cn

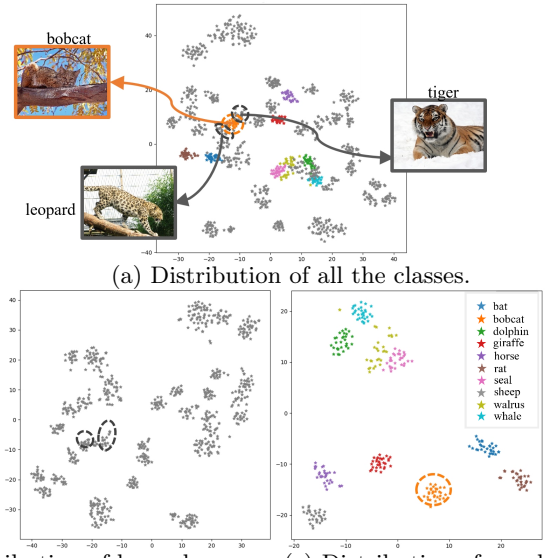
Abstract. The excellent generalization capability of pre-trained Vision-Language Models (VLMs) makes fine-tuning VLMs for downstream zero-shot tasks a popular choice. Despite achieving promising performance in the professionality of base classes, most existing fine-tuned methods suffer from feature confusion of novel classes, resulting in unsatisfactory transferability. To address this problem, we propose a divide-and-conquer approach called Prompt-based Variational Adapter (PVA) that explicitly reduces the prediction bias by separating base and novel samples. Specifically, we design two variational adapters with learnable textual tokens to align latent representations for each modality in a shared latent space. Once trained, we can separate novel samples from entangled space using the similarity metric of latent features, *i.e.,* converting confusion task into two independent ones (One for base classes and the other for novel classes). Moreover, to improve the transferability for novel classes, we further refine the output features of the learned adapters with the global features via a residual connection. We conduct extensive experiments on Generalized Zero-Shot Learning and Cross-Dataset Transfer Learning to demonstrate the superiority of our approach and establish a new state-of-the-art on four popular benchmarks.

Keywords: Visual-Language Models · Zero-Shot Generalization · Variational Adapter

1 Introduction

Recently, pre-trained Vision-Language Models (VLMs) such as CLIP [31] and ALIGN [18] have demonstrated remarkable applicability across various downstream tasks such as Zero-Shot Learning (ZSL) [1, 38]. To further boost the performance on downstream tasks, recent studies have proposed several fine-tuning methods based on the VLMs structure. In general, existing fine-tuning methods can be grouped into two lines: (1) Prompt Learning (PL) [19, 44, 45] is a paradigm that leverages a minimal set of learnable parameters, known as

* Corresponding author



(a) Distribution of all the classes.

(b) Distribution of base classes. (c) Distribution of novel classes.

Fig. 1: t-SNE visualization of AWA2 dataset with CoOP [45]. The color and grey pentagrams represent novel and base samples, respectively. (a) There is obvious overlap and confusion in the distribution of bobcat, tiger, and leopard. (b-c) The confusion issue could be efficiently addressed if the two class sets were meticulously segregated.

prompts, to efficiently adapt pre-trained models to the downstream tasks in a data-efficient manner while maintaining competitive performance. (2) Adapter-style Tuning (AT) [13, 43] involves introducing adapter layers or modules into the pre-trained models, allowing only these adapter layers to be updated during fine-tuning while the rest of the models' parameters remain frozen.

However, both PL and AT paradigms only fine-tune VLMs on base classes, resulting in extreme prediction bias towards base classes. This bias is exacerbated in challenging open-world vision tasks (*e.g.,* Generalized Zero-Shot Learning, GZSL [5, 22, 39]) due to the feature confusion. As shown in Fig. 1(a), the novel samples (color pentagrams) are located around the base samples (grey pentagrams) owing to the lack of prior knowledge of the former. These fine-tuning methods trained on unbalanced base data may tend to classify all samples into base classes and struggle to generalize to novel classes, reducing their transferability.

To this end, we propose a divide-and-conquer approach called Prompt-based Variational Adapter (PVA) that improves both the professionality and transferability of the VLMs model. Inspired by previous PL work, we learn task-specific textual tokens instead of subjective hand-crafted attributes. Compared to the static attributes, these learnable prompts can not only capture more discriminative information but also provide better diversity for subsequent adaptation learning. Then, two variational adapters are constructed to generate latent repre-

sentations in a shared latent space, in which we adopt class-wise and distribution-wise vision-language constraints to ensure alignment and interaction properties for each branch. With the help of variational structures, adapters carefully align multi-modal features in redundance-free latent space, providing accurate visual-semantic interactions. After that, we can search for a threshold to distinguish the base and novel domains by measuring the similarity between the latent representations of the two domains. Based on the similarity, GZSL can thus be converted into a base class classification and a zero-shot task of novel classes, as shown in Fig. 1(b-c). Once base and novel samples are distinguished, the prediction bias problem will naturally be addressed.

In addition, even if the test spaces are accurately distinguished, how to classify novel samples remains a challenging task. Thus, we further refine the output feature of the learned adapters with the global features (extracted from the original CLIP encoder) via a residual connection to improve the transferability. Benefiting from residual connection, variational adapters can fully exploit the transferable general knowledge stored in the original CLIP and freshly learned knowledge originated from training samples. In fact, different from existing methods that train an extra binary classifier to distinguish domains, we efficiently model the entire process (*i.e.*, data separation and classification) in a unified Prompt-Adapter hybrid framework to simultaneously preserve the professionality and transferability of the VLMs model.

In summary, our main contributions are:

- We introduce a novel approach, the Prompt-based Variational Adapter (PVA), specifically designed to address the challenges of GZSL. To the best of our knowledge, our method marks the inaugural exploration of the synergy between PL and AT in the context of fine-tuning CLIP models, offering a novel and effective solution to this complex task.
- We propose to utilize the latent representations of the adapter, which are tightly governed by the vision-language distribution, to distinguish novel classes from base classes. By segmenting the entire test space into two distinct and independent subspaces, the prediction bias that typically favors base classes is explicitly mitigated, thereby enhancing the accuracy and fairness of the classification process.
- Our experimental results demonstrate that the proposed PVA achieves state-of-the-art performance across a wide range of downstream tasks, including GZSL and Cross-Dataset transfer learning. Notably, PVA exhibits significant improvements in recognizing novel classes, showcasing its effectiveness in addressing the inherent challenges of these tasks.

2 Related Work

2.1 Generalized Zero-Shot Learning

How to recognize objects in an open-world visual environment has attracted increasing interest in recent years. GZSL [5,22,23,39] is one of the relevant research

fields focusing on open-world visual tasks. Specifically, GZSL aims to recognize both base and novel objects, relying solely on labeled samples from the base classes. To achieve this goal, existing methods can be grouped into two folds: (1) Embedding-based methods [4, 23, 42], which focus on learning a visual-semantic mapping in visual, semantic or latent space for cross-modal interactions. However, a major drawback of Embedding-based methods is the prediction bias toward base classes for the lack of novel labeled data. Therefore, some researchers tend to explore feature-generation methods to reduce prediction bias. (2) Generative methods [7, 39, 41] aim to generate the visual features or prototype of the novel classes by various generative networks, such as generative adversarial networks (GANs) [14] and variational autoencoders (VAEs) [32]. Despite the promising results, these methods still require massive annotated data and semantic descriptions (*e.g.*, sentences, hand-crafted attributes). In addition, since the generated features come from noise inputs, they usually suffer from extreme domain shift problems.

Different from these existing methods, we explicitly aim to tackle the prediction bias problem by converting the GZSL into a ZSL classification task and a supervised task. One of the related methods is calibration, which has been explored in several recent works [2, 3, 8, 33]. ESZSL [3] calibrates the prediction by directly reducing the probability of base classes. COSMO [2] proposes a soft combination method without any training samples. OOD [8] and DUS [33] attempt to learn the boundary from base samples to separate novel classes from entangled space. Different from the aforementioned methods, we adopt adapter-style tuning equipped with learnable prompts based on VLMs to improve the transferability, not simply learn a domain detector.

2.2 Vision-Language Models for Zero-Shot Generalization

Pre-trained vision-language models (VLMs) explore the relationship between two different modalities by contrastive learning with massive noisy data. With the help of reasonable structures, these models, such as CLIP [31] and ALIGN [18] can provide excellent representations for both visual and text branches. Based on the remarkable transferability of VLMs, the researchers adopt improved methods to address downstream zero-shot generalization problems. [19, 44, 45] apply an efficient fine-tuning paradigm by learning representation for the specific task. [13, 43] learn extra MLPs to adapt the downstream knowledge. However, these methods only consider knowledge transfer, ignoring the feature confusion that occurs in hybrid open-world spaces.

3 Method

3.1 Overall idea

Our method introduces a prompt-based variational adapter to address the prediction bias problem without sacrificing task-specific discrimination and transferability. Fig. 2 shows the overall architecture of the proposed PVA. Different

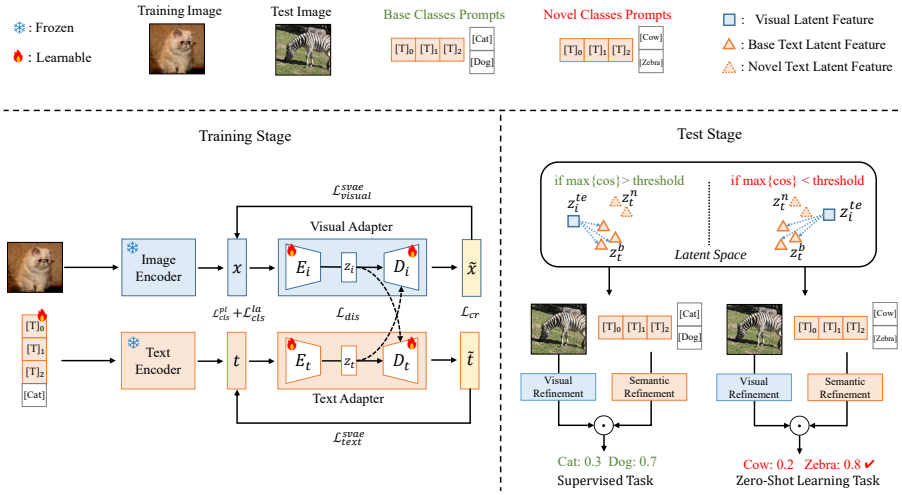


Fig. 2: Overview of the proposed Prompt-based Variational Adapter (PVA). **Left:**The training framework of PVA with detailed implementations. E_i, D_i, E_t and D_t refer to visual encoder, visual decoder, text encoder, and text decoder, respectively. z_i and z_t represent latent features for image and text aspects, where the adapter inputs are x and t, respectively. $\tilde{x}$ and $\tilde{t}$ denote the reconstructed features. **Right:** Test samples are grouped into base domain or novel domain by the cosine similarity between z_i and z_t (including both base and novel classes), and then classified by two experts.

from existing methods, we explicitly reduce the prediction bias by converting the GZSL problem into a ZSL and a supervised classification task. To separate novel (or base) samples from entangled space, we learn two variational adapters to align visual and semantic features, where the text features are equipped with learnable prompts. Then, we further refine the output features of adapters to balance task-specific and general knowledge.

3.2 Revisiting CLIP

CLIP [31] consists of two core encoders: a text encoder $\mathcal{T}$ and a visual encoder $\mathcal{I}$, which are jointly trained by massive noisy image-text pairs with contrastive loss. Each encoder is implemented by either repetitive Transformer [34] or ResNet [16] blocks. Once trained, we can build a zero-shot classifier with arbitrary object names and the corresponding visual features. For example, a N-way classification can be built by the similarity metric between visual features and text prompts for these classes, such as 'A photo of a [CLASS]'. Formally, the output probability of the model is:

$$p(y|x) = \frac{\exp(\cos(\mathcal{I}(x), \mathcal{T}(t_y))/\sigma)}{\sum_{n=1}^{N} \exp(\cos(\mathcal{I}(x), \mathcal{T}(t_n))/\sigma)}, \tag{1}$$

where σ is the temperature parameter and y represents the target class.

3.3 Prompt-based Variational Adapter

Problem Setting. GZSL attempts to recognize novel categories by the prior knowledge transferred from the base domain (*i.e.*, seen domain) to the novel domain (*i.e.*, unseen domain). Here, base domain represented by $\mathcal{D}^s = \{(x, y, t_y)|x \in \mathcal{X}^s, y \in \mathcal{Y}^s, t_y \in \mathcal{A}^s\}$, where y is the class label, and x and t_y refer to features extracted by the visual and text encoder of CLIP, respectively. Similarly, novel domain is given by $\mathcal{D}^u = \{(x^u, u, t_u)|x^u \in \mathcal{X}^u, u \in \mathcal{Y}^u, t_u \in \mathcal{A}^u\}$, where u refers the label of novel classes. The overall goal of GZSL is to learn $f_{GZSL} : \mathcal{X} \rightarrow \mathcal{Y}^s \cup \mathcal{Y}^u$, where $\mathcal{Y}^s \cap \mathcal{Y}^u = \emptyset$.

Learn the Prompt-based Variational Adapter. Inspired by the effectiveness of prompt learning, we also learn textual tokens instead of hand-crafted attributes. Formally, the prompts are given in the following form:

$$t = [\mathrm{T}]_1[\mathrm{T}]_2 \dots [\mathrm{T}]_M[\mathrm{CLASS}], \tag{2}$$

where each $[T]_M$ is a vector with the same dimension (*e.g.*, 512 in CoOP [45]) and M is the length of the learned prompts. Subsequently, prompts t and corresponding images are fed into the frozen text and image encoders for adaption learning.

By forwarding the original training samples and learnable prompts from base classes, we can then build two variational adapters. Inspired by the excellent representations of Spherical Variational Autoencoder (SVAE) [10,32] on low dimensional data, we employ two SVAE as the adapters to align different modalities, while preserving the transferability of classes knowledge. To achieve this goal, we model visual and semantic distributions by vMF and treat each distribution obtained from the visual and semantic branch as prior knowledge for another to align these modalities in spherical space.

Specifically, as shown in the left side of Fig 2, each adapter (*e.g.*, SVAE) consists of an encoder and a decoder with inherent and cross-modal constraints. For the visual branch, we can formulate the variational optimization by:

$$\mathcal{L}_{visual}^{svae} = \mathbb{E}_{q_{\delta_1}(z_i|x)}\left[\log p_{\omega_1}\left(x \mid z_i\right)\right] - \mathrm{KL}\left(q_{\delta_1}\left(z_i \mid x\right) \| p_{\omega_1}\left(z_i\right)\right), \tag{3}$$

where the first term refers to the reconstruction loss for visual features x over latent features z_i, and the second part denotes the distance metric between different distributions. Here $q_{\delta_1}(z_i|x)$ represents the features encoded by the visual encoder E_i, while $p_{\omega_1}\left(x \mid z_i\right)$ comes from D_i.

Similarly, the text adapter can be optimized by another SVAE loss:

$$\mathcal{L}_{text}^{svae} = \mathbb{E}_{q_{\delta_2}(z_t|t)}\left[\log p_{\omega_2}\left(t \mid z_t\right)\right] - \mathrm{KL}\left(q_{\delta_2}\left(z_t \mid t\right) \| p_{\omega_2}\left(z_t\right)\right), \tag{4}$$

where $q_{\delta_2}(z_t|t)$ and $p_{\omega_2}\left(t \mid z_t\right)$ denote the distribution represented by text encoder and decoder, respectively. By formulating the adapter into an SVAE style, the latent features can only depict each distribution independently. Therefore, to align the latent features of different modalities, we propose to minimize the

distance between them. Formally, we adopt optimal transport (Earth Mover's Distance) [35] to match visual latent distribution to text distribution by:

$$\mathcal{L}_{dis} = \inf_{\theta \in \Pi\left(P_{z_i}, P_{z_t}\right)} \mathbb{E}_{(z_i, z_t)\sim\theta}[\|z_i - z_t\|], \tag{5}$$

where $\Pi\left(P_{z_i}, P_{z_t}\right)$ denotes the joint distribution of two latent features z_i and z_t. With the help of $\mathcal{L}_{dis}$, the corresponding categories of visual and semantic features will be aligned in the latent space.

To further ensure visual-semantic alignment, we introduce cross-domain constraints for both visual and semantic adapters. Specifically, given the latent representation z_i and z_t, we reconstruct the representation by another decoders:

$$\mathcal{L}_{cr} = \mathbb{E}\left[||t - D_t(z_i)||_2\right] + \mathbb{E}\left[||x - D_i(z_t)||_2\right], \tag{6}$$

where $D_t(z_i)$ and $D_i(z_t)$ are regarded as cross-domain representations.

Although Eqs. (3) (4) (5) (6) align latent representations in a shared space by SVAE and cross-domain constraints, they still suffer from the lack of any class-wise knowledge, which could lead to feature confusion. To reduce the feature confusion, we propose to adopt dual classification loss (DCL):

$$\mathcal{L}_{cls} = \underbrace{\mathbb{E}_t[\log p_{\phi^{pl}_{cls}}(y|x)]}_{\mathcal{L}^{pl}_{cls}} + \underbrace{\mathbb{E}[\log p_{\phi^{la}_{cls}}(y|z_t)] + \mathbb{E}[\log p_{\phi^{la}_{cls}}(y|z_i)]}_{\mathcal{L}^{la}_{cls}}, \tag{7}$$

where $\mathcal{L}^{pl}_{cls}$ and $\mathcal{L}^{la}_{cls}$ tend to perform classification constraints in instance-level and latent-level, respectively. For the instance-level, the proposed PVA learns task-specific knowledge with training samples $p_{\phi^{pl}_{cls}}(y|\boldsymbol{x}) = \frac{\exp(\cos(\mathcal{I}(\boldsymbol{x}), \mathcal{T}(t_y))/\sigma)}{\sum \exp(\cos(\mathcal{I}(\boldsymbol{x}), \mathcal{T}(t_n))/\sigma)}$, where y and $\mathcal{T}(t_y)$ represent the class labels and semantic features mentioned in Eq. (2). Based on the end-to-end training paradigm, the discrimination of text features can be preserved. For the latent level, the softmax classifier ϕ^{la}_{cls} is trained using latent features and corresponding labels to ensure the decision boundary is clear enough for the final classification task.

As previously illustrated, we develop the overall training process with SVAE, cross-domain, and DCL classification loss:

$$\mathcal{L} = \mathcal{L}_{SVAE} + \alpha\mathcal{L}_{dis} + \beta\mathcal{L}_{cr} + \gamma\mathcal{L}_{cls}, \tag{8}$$

where $\mathcal{L}_{SVAE} = \mathcal{L}^{svae}_{visual} + \mathcal{L}^{svae}_{text}$ and α, β, γ are hyper parameters.

Separate Novel Classes from Base Classes. Following the training stage, both the visual and semantic latent features are well represented. Thus, novel samples entangled with base samples can be easily separated from the test space using distance comparisons. Specifically, as shown on the right side of Fig. 2, we measure the distance between test samples and learned prompts equipped with the [CLASS] token by:

$$distance = cos(z^{te}_i, z_t), \tag{9}$$

where $cos(,)$ denotes the cosine similarity. $z_i^{te} = E_i(x^{te})$ denotes the latent features encoded from test visual features, which may come from base or novel domains. $z_t = [z_t^b, z_t^n]$ contain the latent text features of all the classes (*i.e., base and novel*), where [,] denotes concatenate operation for the latent prompts of base and novel domains. Intuitively, if the test sample belongs to a novel domain, the distance $cos(z_i^{te}, z_t^n)$ would be closer than $cos(z_i^{te}, z_t^b)$.

By leveraging the distance metric, we can assign domain labels by:

$$L = \begin{cases} Base, & \max\left\{cos\left(z_i^{te}, z_t^b\right) \mid \forall t \in \mathcal{A}_s\right\} \geq threshold \\ Novel, & \max\left\{cos\left(z_i^{te}, z_t^b\right) \mid \forall t \in \mathcal{A}_s\right\} < threshold \end{cases}. \tag{10}$$

Here, we alternate the hand-crafted threshold with the base knowledge to fit different datasets and reduce complexity. In practice, we propose to automatically calculate the *threshold* by τ, where $\tau = \frac{count(d^j > threshold)}{N}$. d^j is the maximum distance between j-th training samples and learned prompts, and N is the number of all training samples. The larger the value τ is, the smaller the *threshold* becomes, meaning that more samples are divided into base classes. Then we only need to choose a suitable $\tau \in [0, 1]$ to balance the data separation.

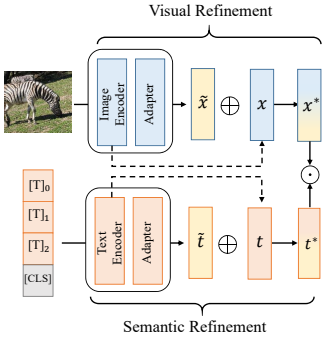


Fig. 3: The architecture of the Visual and Semantic Refinement modules.

Fine-tuning for GZSL Classification. When detecting based on the proposed distance, the domain for these test samples is accurately assigned. Although the challenging GZSL task can be converted into a traditional ZSL and a supervised task, how to recognize novel classes is also a tricky problem. To improve the transfer of general knowledge for novel classes, we further refine the output feature of the learned adapters with the original zero-shot features via a residual connection. As shown in the Fig. 2 and Fig. 3, we refine the final output of adapters by $t^* = \epsilon \cdot \tilde{t} + (1 - \epsilon) \cdot t$ and $x^* = \epsilon \cdot \tilde{x} + (1 - \epsilon) \cdot x$. Since CLIP is obtained by massive image-text pairs that are not limited to the current samples, the features of the original encoder are easier to transfer. Based on the

complement of stored knowledge from CLIP and task-specific knowledge from adapters, we can thus improve the transferability of the PVA.

4 Experiments

4.1 Setup

Datasets and Setting. For the GZSC and CDT tasks, the datasets include two coarse-grained datasets, *i.e.*, AwA2 [12] and SUN [30], and two fine-grained datasets *i.e.*, CUB [36] and FLO [29], containing 50, 717, 200, and 102 categories, respectively. It is worth noting that due to space limitations, we only report the results for popular AwA2, CUB, and SUN in Table 1. The performance is evaluated by the harmonic mean of the average per class Top-1 accuracy: $\mathbf{H} = 2 * \mathbf{B} * \mathbf{N}/(\mathbf{B} + \mathbf{N})$, where $\mathbf{B}$ and $\mathbf{N}$ represent the performance of base and novel classes, respectively. To further explore the knowledge transferability of the proposed method, we follow CoCoOP [44] to test the generalization performance of the trained model for other datasets. Specifically, we first train the PVA on AwA2, which is regarded as a source dataset. Then we compare the performance which is tested on the remaining three datasets without fine-tuning.

Implementation details. The proposed PVA is comprised of two sub-models: a visual adapter and a text adapter. Each adapter consists of an encoder and a decoder implemented by two-layer FC networks with 256 hidden units and ReLU. The input of the adapters comes from the frozen CLIP (ViT-B/16 [11,23]) encoder with 512 dimensions. The latent feature dimension is set to 16 for all these datasets. For the prompts, the learnable parameters length $M = 4$. We train our proposed PVA by Adam optimizer with a 1e-4 learning rate. We select $\alpha = 0.1$, $\beta = 1$, and $\gamma = 1$ as hyperparameters, while τ and ϵ are discussed in the following section.

4.2 Comparisons with SOTA

Generalized Zero-Shot Classification. We partition existing methods into three paradigms to compare the GZSL performance in Table. 1. Specifically, our method achieves the best harmonic mean of 90.7% and 65.8% on AwA2 and SUN, respectively. Although we observe a slightly lower result on CUB than PSVMA [23], our method still achieves competitive results without requiring manually designed attributes. Compared to the popular VLMs-based method CoOP [45], our method achieves a 15.3% improvement on the $\mathbf{N}$ metric of AwA2, indicating the effectiveness of the proposed PVA for reducing prediction bias. CoOP achieves worse $\mathbf{N}$ than the original CLIP, which indicates that existing fine-tuned methods indeed suffer from feature confusion of novel classes, resulting in worse transferability. Notably, the proposed PVA not only improves the performance of novel classes but also ensures the discriminative property of base classes by

Table 1: GZSC performance (%) comparisons on three popular benchmarks.**E**, **G** and **C** represent Embedding, Generative, and Calibration methods, respectively. The best results and second best results are marked by **bold** and underline. PVA+R denotes the proposed method equipped with refinement modules. "†" denotes that the methods are implemented by ViT [11] networks.

Methods	Paradigm	AwA2			CUB			SUN		
		B	N	H	B	N	H	B	N	H
DVBE [26]		70.8	63.6	67.0	60.2	53.2	56.5	37.2	45.0	40.7
GEM [24]		77.5	64.8	70.6	77.1	64.8	70.4	35.7	38.1	36.9
TransZero [4]		82.3	61.3	70.2	68.3	69.3	68.8	33.4	52.6	40.8
MSDN [5]		74.5	62.0	67.7	67.5	68.7	68.1	34.2	52.2	41.3
I2MVFomer [27]		79.6	75.7	77.6	59.9	42.5	49.7	-	-	-
DUET [9]†	E	84.7	63.7	72.7	72.8	62.9	67.5	45.8	45.7	45.8
PSVMA [23]†		77.3	73.6	75.4	77.8	**70.1**	**73.8**	45.3	61.7	52.3
CLIP [31]†		92.9	86.6	89.6	55.1	54.9	55.0	40.2	49.4	44.3
CoOP [45]†		**95.3**	72.7	82.5	63.8	49.2	55.6	49.3	53.5	51.3
SHIP [37]†		94.4	84.1	89.0	58.9	55.3	57.1	-	-	-
f-VAEGAN [40]		70.6	57.6	63.5	60.1	48.4	53.6	38.0	45.1	41.3
TF-VAEGAN [28]		75.1	59.8	66.6	64.7	52.8	58.1	40.7	45.6	43.0
CEZSL [15]		78.6	63.1	70.0	66.8	63.9	65.3	38.6	48.8	43.1
FREE [6]	G	75.4	60.4	67.1	59.9	55.7	57.7	37.2	47.4	41.7
HSVA [7]		76.6	59.3	66.8	58.3	52.7	55.3	39.0	48.6	43.3
ICCE [20]		82.3	65.3	72.8	65.5	67.3	66.4	-	-	-
CS [3]		77.8	5.9	11.0	63.8	12.6	21.0	-	-	-
COSMO [2]		-	-	-	60.5	41.0	48.9	40.2	35.3	37.6
OOD [8]		75.9	55.6	64.2	50.2	49.5	49.8	33.9	41.7	37.0
GatingAE [21]	C	81.3	60.3	69.3	58.1	55.4	56.7	38.1	45.3	41.4
PVA (**Ours**)		92.7	88.0	90.3	78.5	62.5	69.6	**63.4**	66.3	64.8
PVA+R (**Ours**)		93.6	**88.0**	**90.7**	**79.6**	65.7	72.0	62.5	**69.5**	**65.8**

separating novel and base domains. For instance, our method achieves the best **B** of 79.6% and 63.4% on the CUB and SUN datasets, respectively.

In addition, the calibration-based method, *i.e.,* GatingAE [21] focuses on data separation while ignoring the model's transferability. In contrast, we model the entire process in an adapter-style tuning fashion to consider knowledge transfer. Intuitively, it can be seen from the last two rows that the proposed refinement networks further improve the **N** results of 3.2% on both CUB and SUN. This indicates that general knowledge should indeed be introduced by refinement networks to enhance the transferability of the model.

Cross-dataset Transfer Learning. To further evaluate the generalization ability of the model in an open-world setting, we report cross-dataset transfer learning performance in Table 2. Specifically, we train the model on AwA2 base classes and then test the performance on three target datasets. Since all target samples are test-only, we rename the original splits for base and novel classes to Subset1 (Sub1) and Subset2 (Sub2), respectively. As shown in the table, CoOP [45] is even worse than CLIP as it suffers from an overfitting problem for base classes. In contrast, by combining the prompt learning and adapter

Table 2: Cross-dataset transfer performance (%) comparisons on three popular benchmarks. The model is trained on AwA2 and then applied to target datasets. Sub1 and Sub2 represent data splits with different classes. △ denotes PVA's gain over CoOP.

Methods	AwA2 → CUB			AwA2 → SUN			AwA2 → FLO		
	Sub1	Sub2	H	Sub1	Sub2	H	Sub1	Sub2	H
CLIP [31]	54.8	55.2	55.0	40.2	**41.4**	40.8	67.9	65.6	66.7
CoOP [45]	50.8	53.5	52.1	44.6	37.5	40.7	68.1	66.3	67.2
PVA+R (**Ours**)	**56.8**	**55.5**	**56.1**	**44.9**	38.0	**41.1**	**71.2**	**68.6**	**69.8**
△	+6.0	+2.0	+4.0	+0.3	+0.5	+0.4	+3.1	+2.3	+2.6

Table 3: Ablation study of the proposed PVA. We highlight the PVA+R variants of our methods with a light blue background.

$\mathcal{L}_{dis}$	$\mathcal{L}_{cls}$	$\mathcal{L}_{cr}$	ϵ	AwA2			CUB		
				B	N	H	B	N	H
✓				89.2	52.9	66.4	67.7	13.5	22.5
✓	✓			90.1	83.4	86.6	72.1	61.0	66.1
✓	✓	✓		92.7	88.0	90.3	78.5	62.5	69.6
✓	✓	✓	0.1	91.7	88.9	90.3	73.7	65.0	69.0
✓	✓	✓	0.5	92.4	88.9	90.6	72.2	**65.9**	68.9
✓	✓	✓	0.9	**93.6**	88.0	**90.7**	**79.6**	64.7	**71.4**

Table 4: Ablation study of the domain detection. FPR denotes the False Positive Rate (in %) on the threshold that yields 95% TPR.

$\mathcal{L}_{dis}$	$\mathcal{L}_{cls}$	$\mathcal{L}_{cr}$	CUB		SUN	
			FPR ↓	AUC ↑	FPR ↓	AUC ↑
✓			77.1	76.2	89.0	62.6
✓	✓		10.5	98.0	66.3	87.2
✓	✓	✓	**2.6**	**99.2**	**11.3**	**97.8**
	MAX-3 [17]		79.6	73.4	92.3	61.0
	OOD [8]		85.0	71.2	88.8	63.1
	COSMO [2]		72.0	82.0	77.5	77.7

tuning, PVA+R outperforms CoOP by 4.0%, 0.4%, and 2.6% on CUB, SUN, and FLO, respectively. Additionally, it is more interesting to look backward to Table. 1, the performance on CUB observed by traditional generative methods, such as F-VAEGAN [40] and HSVA [7] is even inferior to PVA+R transferred from AwA2. This also demonstrates that our prompt-adapter hybrid framework can be extended to a variety of downstream tasks.

4.3 Ablation Study

Different Constraints. To give a clear insight into each constraint, we conduct a series of ablation experiments. As shown in Table. 3, the performance gradually improves as we incorporate $\mathcal{L}_{dis}$, $\mathcal{L}_{cls}$ and $\mathcal{L}_{cr}$ into the original baseline. It is worth noting that we observe significant improvements when $\mathcal{L}_{cls}$ is added. This is because classification constraints can help models learn discriminative and reasonable prompts while ensuring class-wise alignments for visual and semantic branches. Unlike the changing trend of the above constraints, PVA+R achieves more stable yet high results across various ϵ, indicating that refinement processes balance general and task-specific knowledge to improve overall performance.

Data Separation. To further explore the results of feature separation, we report the binary classification results. As shown in Table. 4, we observe that the AUC raises with the help of $\mathcal{L}_{cls}$ and $\mathcal{L}_{cr}$, maintaining the same trend as **H** shown in Table. 3. In addition, PVA exhibits much stronger domain detection

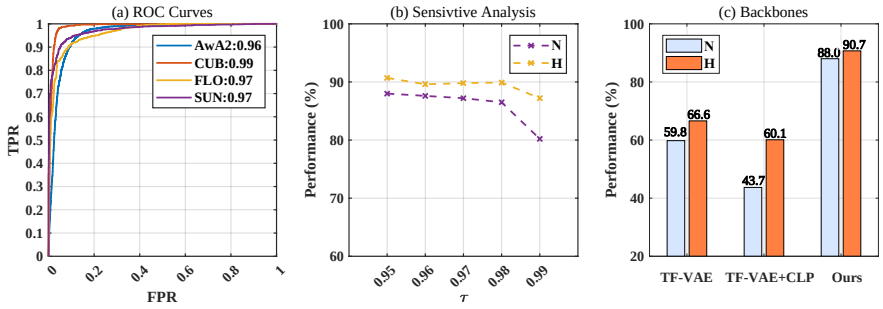


Fig. 4: (a) ROC curves on four benchmarks. (b) GZSL performance (N and H) on AwA2 with different τ. (c) Comparisons between different feature extractors on AwA2. We replace original Res101 [16] based TF-VAE [28] with CLIP features, termed TF-VAE+CLIP.

Table 5: Effectiveness analysis of variational adapter and alternatives. w/ and w/o represent with and without respectively. DS is the abbreviation of Data Separation. △ represents the gain over the second-best results.

Methods	AwA2			CUB			SUN			FLO		
	B	N	H	B	N	H	B	N	H	B	N	H
MLP w/o DS	**95.3**	76.8	85.0	60.3	54.1	57.0	51.2	52.7	51.9	71.8	68.9	70.3
MLP w/ DS	92.8	69.0	79.1	70.3	62.1	66.6	53.1	66.8	59.1	88.2	**72.8**	79.8
VAE w/ DS	93.6	**88.0**	**90.7**	**79.6**	**65.7**	**72.0**	**62.5**	**69.5**	**65.8**	**92.1**	72.1	**80.9**
△	-1.7	+11.2	+5.7	+9.3	+3.6	+5.4	+9.4	+2.7	+6.7	+3.9	-0.7	+1.1

ability than other related methods, leading to pleasing performance. We draw the ROC curves in Fig. 4(a) to report detection results for these datasets, where all AUCs are greater than 96%. This also indicates that novel samples can be effectively separated from entangled base samples by well-trained latent features of adapters. In addition, we believe that the performance of data separation is important because if the domain cannot be accurately distinguished, it will directly affect the classification results and even have negative effects.

Furthermore, we report quantitative analysis between other related methods and the proposed PVA in the lower half of Table. 4. It can be seen that these methods fail to accurately classify entangled features at the domain level, which is consistent with the results in Table. 1.

Analysis of Variational Adapter. We evaluate the effects of different adapter structures. As shown in Table. 5, we first adopt MLPs as the adapter to fine-tune on four datasets. It can be seen that the performance of the novel class decreases significantly in MLPs w/o DS. This is because MLPs fail to learn transferable knowledge due to the simple structures. Although MLP w/ DS shows considerable improvements of **H** on CUB(9.6%), SUN(7.2%), and FLO (9.5%) with domain detection, it still suffers from unstable performance. For example, the harmonic mean drops from 85.0% to 79.1% on AwA2 after domain detection.

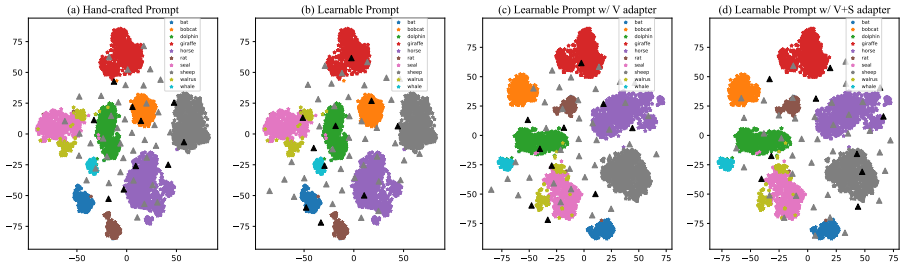


Fig. 5: t-SNE visualizations [25] of learnable prompts and visual features with (w/) Visual and Semantic adapters. Grey and black triangles denote the base and novel prototype, respectively. Color stars denote novel visual features. (Best viewed in color)

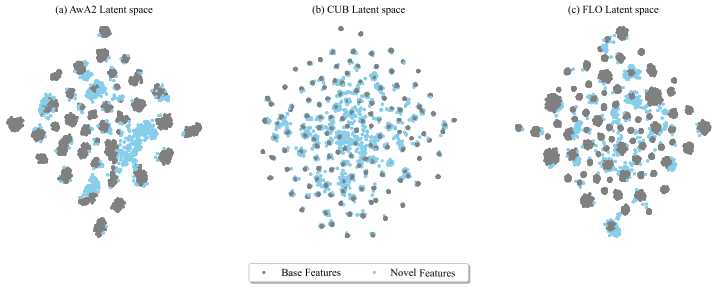


Fig. 6: t-SNE visualizations of latent features for AwA2, CUB and FLO. Grey dots represent base features and blue dots represent novel ones.

We speculate that plain MLPs can not sufficiently align vision-language modality in latent space, resulting in poor domain classification. However, the proposed variational adapters capture discriminative yet transferable information for both latent and output features to ensure the performance of domain detection and classification.

Sensitivity and Backbones. Fig. 4(b) shows the sensitivity of τ. The performance of $\mathbf{N}$ maintains stable at first and then decreases sharply when $\tau \geq 0.99$. The reason for this observation is that the larger τ tends to classify more samples into base classes, resulting in lower $\mathbf{N}$. The descent speed of $\mathbf{H}$ is less than $\mathbf{N}$, for the performance of $\mathbf{B}$ improves as τ increases. In our experiments, we set $\tau = 0.95$ for all the datasets.

We evaluate the influence of different backbones for fair comparisons. As shown in Fig. 4(c), when we applied CLIP features to existing generative methods such as TF-VAE [28], we did not observe improvements, which are brought by high-quality representations. We speculate that the features extracted from VLMs are discriminative enough for classification, while they do not require additional complex optimization. Different from simple combinations between off-the-shelf methods and CLIP features, we fine-tune the VLMs in a prompt-

adapter hybrid fashion to guide feature optimization, leading to superior performance.

4.4 Visualization

Visualizations for Visual and Semantic Features. To show the distribution, we visualize them by t-SNE [25] in Fig. 5. By comparing Fig. 5(a) and Fig. 5(b), we can find that the learnable prompts exhibit a higher degree of dispersion. These learnable prompts capture task-specific knowledge from training samples to improve classification performance. Further, we explore the visual and semantic embedding fine-tuned by the proposed variational adapter. From Fig. 5(c) to Fig. 5(d), we can see that visual adapters tend to learn a more intensive distribution for each class to preserve intra-class information, while semantic adapter refines the corresponding relationship between vision and semantic modalities. By incorporating prompt learning and adapter-style tuning into a unified framework, PVA gives a reasonable solution for downstream task generalization.

Visualizations for latent Features. We visualize the latent features of both base and novel classes in Fig. 6. It can be seen that blue dots are located in gaps of gray dots with clear boundaries. Thus, we can separate these novel samples by cosine similarity. Although most of the novel samples can be accurately separated, some samples are also misclassified due to feature confusion. This is coincident with the results of Fig. 4(a).

5 Conclusion

In this paper, we provide a Prompt-Adapter hybrid method PVA to handle the feature confusion problem caused by unbalanced fine-tuning. In general, we fine-tune the corresponding variational adapters for visual and language branches by aligning latent representations of them, where the input of the language adapter is the learnable textual tokens. In addition, we refine adapter output features with global features stored in the original CLIP to improve knowledge transfer. Once trained, we can build two independent classification tasks by similarity metrics. We show that the proposed PVA can achieve state-of-the-art performance on GZSC and challenging CDT tasks. Considering limitations, although PVA achieves remarkable results, it requires additional training costs, which we aim to alleviate in the future.

Acknowledgment

This research was supported in part by NSFC (12326608, U19B2043, 62441602), National Science Foundation for Distinguished Young Scholars under Grant 62225605, Zhejiang Provincial Natural Science Foundation of China under Grant LD24F020016, the Key R&D Program of Zhejiang Province, China 2023C01043, Science and Technology Innovation of Ningbo (2023Z236, 2024Z294), and the Fundamental Research Funds for the Central Universities.

References

1. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for attribute-based classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 819–826 (2013)
2. Atzmon, Y., Chechik, G.: Adaptive confidence smoothing for generalized zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11671–11680 (2019)
3. Chao, W.L., Changpinyo, S., Gong, B., Sha, F.: An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In: Proceedings of the European Conference on Computer Vision. pp. 52–68. Springer (2016)
4. Chen, S., Hong, Z., Liu, Y., Xie, G.S., Sun, B., Li, H., Peng, Q., Lu, K., You, X.: Transzero: Attribute-guided transformer for zero-shot learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 330–338 (2022)
5. Chen, S., Hong, Z., Xie, G.S., Yang, W., Peng, Q., Wang, K., Zhao, J., You, X.: Msdn: Mutually semantic distillation network for zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7612–7621 (2022)
6. Chen, S., Wang, W., Xia, B., Peng, Q., You, X., Zheng, F., Shao, L.: Free: Feature refinement for generalized zero-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 122–131 (2021)
7. Chen, S., Xie, G., Liu, Y., Peng, Q., Sun, B., Li, H., You, X., Shao, L.: Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. Proceedings of the Advances in Neural Information Processing Systems **34**, 16622–16634 (2021)
8. Chen, X., Lan, X., Sun, F., Zheng, N.: A boundary based out-of-distribution classifier for generalized zero-shot learning. arXiv preprint arXiv:2008.04872 (2020)
9. Chen, Z., Huang, Y., Chen, J., Geng, Y., Zhang, W., Fang, Y., Pan, J.Z., Chen, H.: Duet: Cross-modal semantic grounding for contrastive zero-shot learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 405–413 (2023)
10. Davidson, T.R., Falorsi, L., De Cao, N., Kipf, T., Tomczak, J.M.: Hyperspherical variational auto-encoders. arXiv preprint arXiv:1804.00891 (2018)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1778–1785 (2009)

13. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision pp. 1–15 (2023)
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 2672–2680 (2014)
15. Han, Z., Fu, Z., Chen, S., Yang, J.: Contrastive embedding for generalized zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2371–2381 (2021)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
17. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)
18. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of the International Conference on Machine Learning. pp. 4904–4916 (2021)
19. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113–19122 (2023)
20. Kong, X., Gao, Z., Li, X., Hong, M., Liu, J., Wang, C., Xie, Y., Qu, Y.: En-compactness: Self-distillation embedding & contrastive generation for generalized zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9306–9315 (2022)
21. Kwon, G., Al Regib, G.: A gating model for bias calibration in generalized zero-shot learning. IEEE Transactions on Image Processing (2022)
22. Li, X., Xu, Z., Wei, K., Deng, C.: Generalized zero-shot learning via disentangled representation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 1966–1974 (2021)
23. Liu, M., Li, F., Zhang, C., Wei, Y., Bai, H., Zhao, Y.: Progressive semantic-visual mutual adaption for generalized zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15337–15346 (2023)
24. Liu, Y., Zhou, L., Bai, X., Huang, Y., Gu, L., Zhou, J., Harada, T.: Goal-oriented gaze estimation for zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3794–3803 (2021)
25. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(11) (2008)
26. Min, S., Yao, H., Xie, H., Wang, C., Zha, Z.J., Zhang, Y.: Domain-aware visual bias eliminating for generalized zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12664–12673 (2020)
27. Naeem, M.F., Khan, M.G.Z.A., Xian, Y., Afzal, M.Z., Stricker, D., Van Gool, L., Tombari, F.: I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15169–15179 (2023)
28. Narayan, S., Gupta, A., Khan, F.S., Snoek, C.G., Shao, L.: Latent embedding feedback and discriminative features for zero-shot classification. In: Proceedings of the European Conference on Computer Vision. pp. 479–495. Springer (2020)

29. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729 (2008)
30. Patterson, G., Hays, J.: Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2751–2758 (2012)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021)
32. Schonfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., Akata, Z.: Generalized zero- and few-shot learning via aligned variational autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8247–8255 (2019)
33. Su, H., Li, J., Chen, Z., Zhu, L., Lu, K.: Distinguishing unseen from seen for generalized zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7885–7894 (2022)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. proceedings of the Advances in Neural Information Processing Systems (2017)
35. Villani, C., et al.: Optimal transport: old and new, vol. 338. Springer (2009)
36. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
37. Wang, Z., Liang, J., He, R., Xu, N., Wang, Z., Tan, T.: Improving zero-shot generalization for clip with synthesized prompts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3032–3042 (2023)
38. Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., Schiele, B.: Latent embeddings for zero-shot classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 69–77 (2016)
39. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5542–5551 (2018)
40. Xian, Y., Sharma, S., Schiele, B., Akata, Z.: f-vaegan-d2: A feature generating framework for any-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10275–10284 (2019)
41. Xu, B., Zeng, Z., Lian, C., Ding, Z.: Generative mixup networks for zero-shot learning. IEEE Transactions on Neural Networks and Learning Systems (2022)
42. Zhang, L., Xiang, T., Gong, S.: Learning a deep embedding model for zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2021–2030 (2017)
43. Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free adaption of clip for few-shot classification. In: Proceedings of the European Conference on Computer Vision. pp. 493–510. Springer (2022)
44. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022)
45. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)