Data-Free Generalized Zero-Shot Learning

Bowen Tang[1]*, Jing Zhang[1]†*, Long Yan[1]*, Qian Yu[1], Lu Sheng[1], Dong Xu[2]

[1] School of Software, Beihang University
[2] Department of Computer Science, The University of Hong Kong
tbw.broHan@qq.com, {zhang_jing, yanlong, qianyu, lsheng}@buaa.edu.cn, dongxu@hku.hk

Abstract

Deep learning models have the ability to extract rich knowledge from large-scale datasets. However, the sharing of data has become increasingly challenging due to concerns regarding data copyright and privacy. Consequently, this hampers the effective transfer of knowledge from existing data to novel downstream tasks and concepts. Zero-shot learning (ZSL) approaches aim to recognize new classes by transferring semantic knowledge learned from base classes. However, traditional generative ZSL methods often require access to real images from base classes and rely on manually annotated attributes, which presents challenges in terms of data restrictions and model scalability. To this end, this paper tackles a challenging and practical problem dubbed as data-free zero-shot learning (DFZSL), where only the CLIP-based base classes data pre-trained classifier is available for zero-shot classification. Specifically, we propose a generic framework for DFZSL, which consists of three main components. Firstly, to recover the virtual features of the base data, we model the CLIP features of base class images as samples from a von Mises-Fisher (vMF) distribution based on the pre-trained classifier. Secondly, we leverage the text features of CLIP as low-cost semantic information and propose a feature-language prompt tuning (FLPT) method to further align the virtual image features and textual features. Thirdly, we train a conditional generative model using the well-aligned virtual image features and corresponding semantic text features, enabling the generation of new classes features and achieve better zero-shot generalization. Our framework has been evaluated on five commonly used benchmarks for generalized ZSL, as well as 11 benchmarks for the base-to-new ZSL. The results demonstrate the superiority and effectiveness of our approach. Our code is available in https://github.com/ylong4/DFZSL.

Introduction

The power of deep learning models lies in their ability to extract rich knowledge, including visual features and semantic information, from large-scale datasets. However, the sharing of data across different companies, institutions, and countries has become increasingly challenging and sensitive. Concerns related to data copyright and privacy, particularly in sensitive domains such as health and security, pose significant obstacles to the seamless transfer of knowledge from large-scale datasets to novel downstream tasks and concepts. These challenges impede the widespread utilization of deep learning models and limit their potential impact in various fields.

Inspired by the mounting concerns regarding data and model privacy issues, particularly in the context of knowledge transfer to new concepts, this paper addresses the problem of data-free zero-shot learning without access to any real data. Zero-shot learning (ZSL) addresses the challenge of recognizing new classes by leveraging semantic knowledge transferred from base classes. Despite the notable advancements in ZSL, most ZSL methods often require access to labeled images from base classes, either for aligning visual-semantic embeddings or training conditional generative models (Xian et al. 2018a,b; Narayan et al. 2020). Unfortunately, obtaining real data from base classes is often impractical in real-world applications due to privacy or copyright restrictions. Moreover, existing approaches heavily rely on manually annotated attributes, which present challenges in terms of scalability and the difficulty of annotation (Farhadi et al. 2009; Gan et al. 2015; Shigeto et al. 2015; Romera-Paredes and Torr 2015).

The recent progress in large-scale pre-trained vision-language models, such as CLIP (Radford et al. 2021), have demonstrated impressive zero-shot generalization abilities. These models achieve this capability through extensive training on vast collections of image-caption pairs without the requirement of manually annotated attributes. However, effectively transferring the knowledge from these models, which are trained on weakly aligned image-caption pairs, to downstream fine-grained zero-shot classification tasks remains challenging and sub-optimal. This is primarily due to the discrepancy in class granularity between the pre-trained models and the specific classification tasks at hand. Propmt tuning deals with this issue by adding learnable prompts to the inputs. However the recent prompt tuning methods (Zhou et al. 2022b; Bahng et al. 2022; Zhou et al. 2022a) still suffer from single-side alignment and rely on the access to the real images.

To this end, this paper addresses a challenging and practical problem dubbed as data-free zero-shot learning (DFZSL). In this setting, the only available resource for zero-shot classification is a pre-trained base classes classi-

*These authors contributed equally.

†Corresponding author.

fier based on CLIP features. Notably, we do not have access to any real data from either the base or new classes, and manual attribute annotations are not required. Our setting is closely related to Absolute Zero-Shot Learning (Gao et al. 2022). However, their method still relies on manual attribute annotations and performs poorly in both conventional and generalized ZSL.

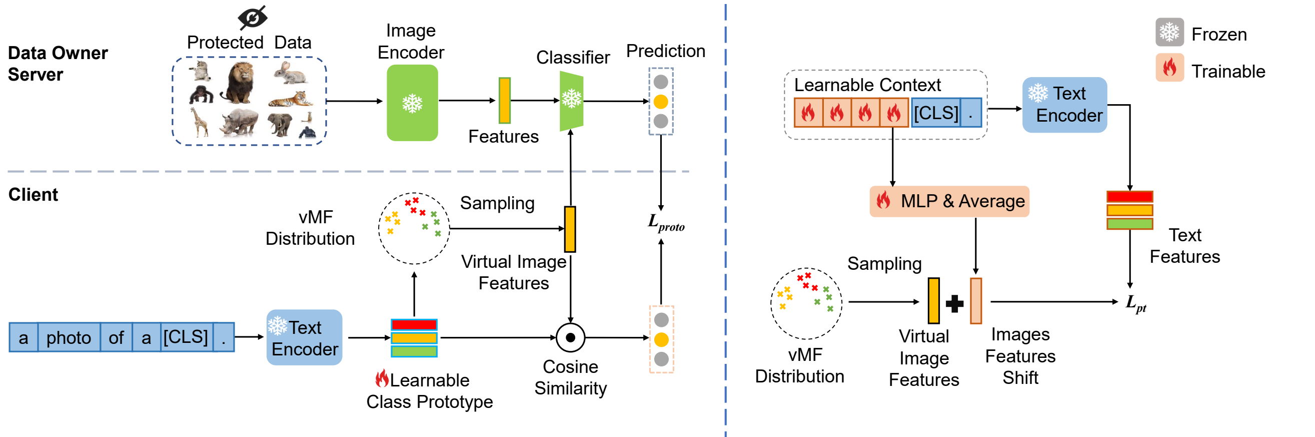
The proposed framework consists of three main components. Firstly, to recover the base class data, we model the CLIP features of base class images as samples from a von Mises-Fisher (vMF) distribution, with learnable mean ($\boldsymbol{\mu}$) and proper concentration (κ) parameters based on the pretrained classifier. This allows us to recover the virtual features of the base data by sampling from the distribution. It is important to note that our method does not recover the original images. Instead, our focus is on recovering the high-level image feature vectors, which is more efficient and avoids the privacy and copyright concerns. Secondly, to bridge the base and new classes, we leverage the text encoder of CLIP to obtain low-cost semantic information in the form of generalizable text features, which eliminates the need for manual attribute annotations. Our framework is generic, and any vision-language foundation models can be potentially used. In order to enhance the adaptation to downstream fine-grained zero-shot classification tasks, we introduce a feature-language prompt tuning method. This method aims to further align the virtual image features of base classes with their corresponding text features by tuning both visual features and textual inputs. Thirdly, we train a conditional generative model using the well-aligned virtual image features and corresponding semantic text features, which enables us to generate labeled data for new classes. And then zero-shot classification is achieved through supervised learning. Our framework has been evaluated on five commonly used benchmarks for generalized ZSL, as well as 11 benchmarks for the base-to-new generalization. The results demonstrate the superiority of our approach.

Related Work

Traditional Zero-Shot Learning. Zero-shot learning (ZSL) is a research area that explores the generalizability of deep learning models. Specifically, it focuses on training a classifier that can recognize samples from the new classes that are unseen during training. It is broadened to generalized zero-shot learning (GZSL) where both base and new classes should be recognized during the testing phase. Embedding based methods and generative-model based methods are the two mainstream methodologies for GZSL. An embedding based approach aims to learn a mapping function that maps visual features and semantic information into a unified space (Romera-Paredes and Torr 2015; Jiang et al. 2019a; Reed et al. 2016; LeCun et al. 1998; Elman 1990). The absence of new class data makes it prone to overfitting so that the test samples of the new classes are easy to be incorrectly classified into a base class. To mitigate the data imbalance problem, recent studies prefer generative-model based methods because they can convert the challenging ZSL problem into a fully-supervised recognition task by synthesizing the absent samples of the new classes. Most generative-model based methods use GAN or VAE for generation (Ye et al. 2019; Han et al. 2021; Chen et al. 2021b; Ye et al. 2022), and some studies have explored the combination of them which we termed as VAEGAN (Larsen et al. 2016; Xian et al. 2019a). A significant drawback in both embedding based methods and generative-model based methods is that they rely on a large amount of real image data to learn the shared embedding space or train the generator. This requirement raises concerns related to copyright infringement and privacy issues. Moreover, these methods often necessitate experts attribute annotations, which are labor-intensive and expensive.

Fine-Tuning for Vision-Language Models. Recently, the contrastive trained vision-language model CLIP (Radford et al. 2021) shows impressive zero-shot performance on recognition tasks. When we directly infer with the pretrained CLIP, which is denoted as Zero-Shot CLIP, the performance is still limited on some downstream datasets. It is because of the domain shift between the pre-training dataset and the downstream datasets for specific tasks, especially when the task is fine-grained. Adapter methods focus on further learning a mapping network for the output features. One of them is CLIP-Adapter (Romera-Paredes and Torr 2015), which uses a residual connected MLP after the last vision layer and the last text layer. Prompt tuning methods introduce parameters to the input. CoOp (Zhou et al. 2022b) establishes a set of learnable vectors on the textual side to learn a generic prompt template. CoCoOp (Zhou et al. 2022a) leverages information from the visual side and builds instance-level prompt templates to achieve base-to-new generalization. VP (Bahng et al. 2022) and VPT (Derakhshani et al. 2022) design the tunable visual prompts either on the images or on the patch tokens. MaPLe (Khattak et al. 2023) inserts learnable tokens inside the visual encoder and text encoder for deep fine-tuning, but it still relies on images and requires more training. Unlike existing prompt tuning methods that focus on learning single-modality prompts or rely on the original images, our approach involves tuning prompts for both visual features (without need of the original images) and textual inputs.

Data-Free Transfer Learning. Data plays a significant role in the development of artificial intelligence. The value of data is more and more appreciative so real data is always not accessible due to copyright or privacy issues nowadays. Thus, data-free transfer learning, which just needs the source model for training but leaves the source data protected, receives increased attention. Existing research under the Data-Free setting can be divided into three categories based on the level of the guards of the source model parameters: available source model parameters (Gao et al. 2022; Tian et al. 2021), inaccessible parameters but enable gradient propogation (Chen et al. 2019), and a black-box service, where the source model just exposes an API for the client to request predictions (Gao et al. 2022). Despite their exhaustive experiments with various guard levels, the results are not satisfactory in the most realistic black-box scenario.



(a) Base-class Data Recovery

(b) Feature-Language Prompt Tuning

Figure 1: The proposed framework is based on vision-language pre-trained models, such as CLIP. (a) Stage 1: Model the distribution of base class image features properly and then sample virtual image features. (b) Stage 2: Align the obtained virtual image features with the extracted text features via FLPT (Feature-Language Prompt Tuning).

Methodology

Problem Formulation and Overview

In this paper, we propose and address a novel problem called data-free zero-shot learning (DFZSL). In the data-free setting, data from base classes are protected considering privacy or copyrights and cannot be directly used for training zero-shot learning methods.

Formally, consider a data owner as the server and there are images of base classes on the server. The image encoder of pre-trained CLIP can be used to extract image features $\boldsymbol{x}^{base} \in X^{base}$. The corresponding labels for base classes are $y^{base} \in Y^{base}$. A classifier for base classes can be trained with image features extracted from real images by the pre-trained CLIP model. We denote the trained base classifier as $f^{base} : X^{base} \to Y^{base}$. In a white-box scenario, the classifier weights are available. While in a black-box scenario, the server does not leak the classifier weights and just exposes an API for the client to request the prediction.

The client side does not have access to any images that can be used for training in the data-free setting. The image features and labels of new class data are denoted as $\boldsymbol{x}^{new} \in X^{new}$ and $y^{new} \in Y^{new}$, respectively. The classes of base data and new data are disjoint so that $Y^{base} \cap Y^{new} = \emptyset$. Now we have $X = X^{base} \cup X^{new}$ and $Y = Y^{base} \cup Y^{new}$.

In data-free zero-shot learning, the objective is to classify test images from both base and new classes on the client side, utilizing the assistance of a base classifier located on the server, without directly accessing the training data.

Overview. To solve the proposed data-free zero-shot learning problem, we propose a generic framework consisting of three main stages. The overall pipeline is shown in Fig. 1. Firstly, we propose to recover the image features of base classes from the base classifier. Secondly, we leverage the CLIP text features as the semantic side information, and propose a feature-language prompt tuning method to extract well-aligned high-quality image features and text features. Thirdly, with both, we follow traditional generative model-based methods to train the generator and classifier for the downstream fine-grained zero-shot classification tasks.

Base-Class Data Recovery

While we want to transfer from base classes to disjoint new classes on downstream datasets via traditional generative methods, we have to prepare the training data. Therefore, we propose a novel method to recover the image features of base classes from the base classifier located on the server in both white-box and black-box setting. The general idea is to model the distribution of the original data properly and then sample virtual data from it.

The most commonly used softmax classifier in deep learning is in the form of a matrix that consists of weights for each class and is trained with cross-entropy loss. During the training process, these weights are forced to be closer to the image features of the corresponding class and keep away from other classes. Especially for CLIP, cosine similarity is used as the metric for prediction. As a result, the L2-normed weights are good representations of class prototypes.

To recover the base-class data, we hypothesise that the L2-normed image features follow a distribution on a hypersphere, where features belonging to the same class tend to cluster together. Therefore, we approximate the real image features of base classes using the von Mises-Fisher (vMF) distribution that is defined on the surface of a unit hypersphere. A vMF distribution is determined by two parameters, the mean direction $\boldsymbol{\mu}$ and the concentration κ. We establish a set of class prototypes: $M = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_{|Y^{base}|}\}$ as the mean direction of each class. In the white-box scenario, the base classifier weights $W = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_{|Y^{base}|}\}$ are available, and we directly set $M = W$. In the black-box scenario, we turn the class prototypes M into learnable parameters and initialize them by the text features $T = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_{|Y^{base}|}\}$ of corresponding classes. The text features are extracted by CLIP's text encoder $\mathcal{T}_{encoder}$ from the prompt "a photo of a [CLS].", where [CLS] is a class name.

With the mean direction, virtual image features can be sampled surrounding the center and the concentration controls the spread of sampled features. While we choose a proper concentration, the underlying principle is that if sampled features from the distribution belong to different classes, they should be sufficiently separable. We assume an isotropic covariance for simplicity. In statistics, vMF distribution is a close approximation to the wrapped Gaussian distribution. According to the empirical rules of the Gaussian distribution, approximately 99.73% of the sampled data deviates from the mean within a range of 3 standard deviations 3σ. Therefore, if we aim for the sampled virtual data from the two classes to be discriminative, the arc length between their respective prototypes should be at least greater than 6σ. Then the concentration parameter can be derived from the definition of the von Mises-Fisher (vMF) distribution, expressed as $\kappa = \frac{1}{\sigma^2}$. Thus, we set the concentration that caters to all class pairs in the downstream datasets with the class prototypes:

$$\kappa_{text} = \max_{\forall y,y' \in Y^{base}, y \neq y'} \left\{ \left[\frac{1}{6} \arccos \left(\frac{\boldsymbol{\mu}_y}{|\boldsymbol{\mu}_y|} \cdot \frac{\boldsymbol{\mu}_{y'}}{|\boldsymbol{\mu}_{y'}|} \right) \right]^{-2} \right\} \quad (1)$$

The virtual image features can be considered to follow the vMF distribution determined by prototypes M, λ, and κ_{text}:

$$\tilde{\boldsymbol{x}} \sim \mathrm{vMF}\left(M, \lambda \cdot \kappa_{text}\right), \quad (2)$$

where λ is a hyper-parameter to refine the concentration (which is nomally set to $\lambda = 1$ and a greater λ produces a more concentrated distribution within each class).

We can now sample virtual image features of base classes $\tilde{\boldsymbol{x}}^{base} \in \tilde{X}^{base}$ from the vMF distribution. However, for the black-box scenario, we further tune the initial class prototypes to mitigate the modality gap between image and text. We first upload the sampled virtual image features as test samples to the base classifier on the server and obtain the prediction scores s^{base}. Subsequently, we treat the learnable class prototypes as a base classifier on the client and predict scores for these virtual image features. The objective is to align the two sets of scores and make the prototypes closely resemble the classifier weights protected on the server:

$$L_{proto}(M) = \frac{1}{|Y^{base}|} \sum_{y \in Y^{base}} \left(\boldsymbol{s}^{base} - \cos\left(\tilde{\boldsymbol{x}}^{base}, M \right) \right)^2 \quad (3)$$

Feature-Language Prompt Tuning

In addition to the recovered virtual image features, another crucial component for achieving the transfer from base to new classes is the semantic information. We use the text features extracted by the pre-trained CLIP model. However, due to the discrepancy in class granularity between the pre-trained models and the downstream fine-grained zero-shot classification tasks at hand, the semantic information produced by the CLIP models trained on weakly aligned image-caption pairs is sub-optimal.

To further enhance the quality of the semantic information, we propose a feature-language prompt tuning method. On the visual side, our method requires only the features of the image, not the original picture. Considering that the recovered virtual image features only cover base classes and we have to transfer to new classes, we tune the prompts in a class-agnostic way. Particularly, we replace the embeddings of class-agnostic prefix "a photo of a" used in CLIP with learnable parameters $P = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3, \boldsymbol{p}_4\}$. In addition to the text prompts, we also introduce a shift term $\boldsymbol{x}_{shift}$ as class-agnostic learnable parameters added to the image features. To accumulate class-agnostic generalizable knowledge, we establish a connection between the text prompts and the image shift using a light mapping network F_Θ parameterized by Θ. It serves as a bridge between the textual information and the image transformation, allowing for the integration of both modalities. The $\boldsymbol{x}_{shift}$ is then defined as:

$$\boldsymbol{x}_{shift} = \frac{1}{4} \sum_{i=1}^{4} F_\Theta\left(P\right). \quad (4)$$

We try to tune these parameters to better align the features of two modalities and mitigate the domain shift between CLIP's pre-training dataset and the downstream datasets. Then we have the enhanced image features which can help us easily generalize to different classes:

$$\hat{\boldsymbol{x}}^{base} = \tilde{\boldsymbol{x}}^{base} + \alpha \boldsymbol{x}_{shift}, \quad (5)$$

where α is the trade-off parameter that controls how much we add the shift term. The enhanced text features are:

$$\hat{\boldsymbol{t}} = \mathcal{T}_{encoder}\left(\{\boldsymbol{p}_1\ \boldsymbol{p}_2\ \boldsymbol{p}_3\ \boldsymbol{p}_4\ [CLS]\,.\}\right), \quad (6)$$

where $[CLS]$ represents the embedding of class name. The parameters are optimized by lower the cross-entropy when classifying the image features by the text features:

$$L_{pt}(\Theta, P) = -\log \frac{\exp\left(\cos\left(\hat{\boldsymbol{x}}^{base}, \hat{\boldsymbol{t}}_y\right)/\tau\right)}{\sum_{y' \in Y^{base}} \exp\left(\cos\left(\hat{\boldsymbol{x}}^{base}, \hat{\boldsymbol{t}}_{y'}\right)/\tau\right)}, \quad (7)$$

and τ is the temperature used in CLIP which equals to 0.01.

New Class Features Generation for Zero-shot Classification

After we recovered the base class image features and conducted the feature-language prompt tuning, we have already prepared high-quality training data for traditional generative zero-shot learning methods.

Generate Data of New Classes. We choose to train a suitable generative model based on the enhanced base data. The loss function is termed as follows:

$$L_{generator}(\Phi) = \ell\left(\hat{\boldsymbol{x}}^{base}, G_\Phi\left(\boldsymbol{z}, \hat{\boldsymbol{t}}^{base}\right)\right), \quad (8)$$

where ℓ can be GAN loss or other loss defined by the chosen generative model. G is the generator parameterized by Φ and $\boldsymbol{z}$ represents the random noise. Then we condition the generator with text features of new classes and generate the new class image features:

$$\hat{\boldsymbol{x}}^{new} = G_\Phi\left(\boldsymbol{z}, \hat{\boldsymbol{t}}^{new}\right). \quad (9)$$

Supervised Image Classification. With the enhanced virtual image features of base classes $\hat{\boldsymbol{x}}^{base}$ and the generated virtual image features of new classes $\hat{\boldsymbol{x}}^{new}$, the generalized zero-shot learning problem is converted into a fully-supervised image classification problem. Moreover, we initialize the weights of the final classifier by the enhanced text features to speed up the training process.

Experiments

Datasets and Implementation Details

Datasets. We evaluate our method in two different tasks: generalized zero-shot learning and base-to-new generalization. For generalized zero-shot learning, we follow the same setting as (Xian et al. 2018a). Our framework is evaluated on five datasets: Attribute Pascal and Yahoo (APY) (Farhadi et al. 2009), CaltechUCSD-Birds (CUB) (Welinder et al. 2010), Oxford Flowers (FLO) (Nilsback and Zisserman 2008), SUN Attribute (SUN) (Patterson and Hays 2012), and Animals with Attributes2 (AWA2) (Xian et al. 2018a), which contains 32, 200, 102, 717 and 50 classes, respectively. As for the base-to-new generalization, we follow the setting proposed in CoCoOp (Zhou et al. 2022a). We evaluate the performance of our framework on 11 different image classification datasets which covers a wide range of recognition tasks. This includes a large-scale visual dataset, ImageNet (Deng et al. 2009); a generic-objects datasets, Caltech101 (Fei-Fei 2004); five fine-grained image recognition datasets, OxfordPets (Parkhi et al. 2012), StanfordCars (Krause et al. 2013), Flowers102 (Nilsback and Zisserman 2008), Food101 (Bossard, Guillaumin, and Van Gool 2014) and FGVCAircraft (Maji et al. 2013); a satellite-view topographic image dataset EuroSAT (Helber et al. 2019); an action recognition dataset UCF101 (Soomro, Zamir, and Shah 2012); a texture dataset DTD (Cimpoi et al. 2014); and a scene recognition dataset SUN397 (Xiao et al. 2010). These datasets will be detailed in the appendix.

Implementation Details. Our proposed framework consists of three stages: recover virtual image features of base classes, utilize FLPT to enhance both the virtual image features and semantic text features, and finally adopt traditional generative ZSL methods. We recover the base class data in the data-free setting at the first stage. In the white-box scenario, where the base classifier weights are accessible, we directly apply them as the class prototypes. In the black-box scenario, the class prototypes are initialized by text features, and λ is set to 1. We apply the Adam optimizer and the learning rate is set to 0.0003. In the second prompt tuning stage, we implement the light mapping network with a single-hidden-layer MLP activated by GELU. For the third stage, we use off-the-shelf generative-model based methods. The setup stays the same as what they proposed in their papers. All experiments are performed on an NVIDIA GeForce RTX3090, except for the ImageNet, which is performed on an NVIDIA A100.

Results of Generalized Zero-Shot Learning

Setup. We follow the splits and evaluation protocols proposed in (Xian et al. 2018a), train on base classes and then evaluated on test set that mixes the base classes and the new classes. Differently, the only input information of our base classes is the base classifier on the server. We do not use the attribute vectors provided in these benchmarks, but extract text features by CLIP with “a photo of a [CLS].” instead. The mean per-class top-1 accuracy is reported on base and new classes and the harmonic mean is computed to demonstrate the balanced performance of our framework.

Baselines. We choose f-CLSWGAN (Xian et al. 2018b), Cycle-WGAN (Felix et al. 2018), LisGAN (Li et al. 2019), TCN (Jiang et al. 2019b), f-VAEGAN (Xian et al. 2019b), TF-VAEGAN (Narayan et al. 2020), GCMCF (Yue et al. 2021), HSVA (Chen et al. 2021a) MSDN (Chen et al. 2022b) AZSL (Gao et al. 2022) and SHIP+CoOp (Wang et al. 2023) as our baseline methods.

Main results. The results of generalized zero-shot learning on the five commonly used benchmarks are shown in Table 1. Firstly, in comparison to traditional generalized zero-shot learning approaches that utilize the ImageNet-1k pre-trained ResNet-101 as the backbone, the baseline data-free ZSL method (AZSL (Gao et al. 2022)) exhibits significantly inferior performance. This highlights the inherent challenge of the data-free ZSL task, as it performs notably worse than ZSL methods that have access to real base data. Secondly, our method improves both traditional generative ZSL methods using CLIP features and the state-of-the-art prompt-tuning methods even without access to the real data in both black-box (Data-Free) and white-box (Data-Free*) settings. For example, when compared with the state-of-the-art SHIP+CoOp method, our framework improves the harmonic mean accuracy by 4.5%, 6.3%, and 10.8% on the three standard benchmarks of AWA2, CUB and FLO, respectively. This verify the effectiveness of the proposed framework. Thirdly, it is worth noting that the performance gain by our method does not solely originate from the base classes but primarily from the new classes. This observation validates the generalization ability of the proposed method.

Results of Base-to-New Generalization

Setup. We follow CoCoOp (Zhou et al. 2022a) to makes a half-and-half split on 11 datasets, which turns out to divide them into two non-overlapping subsets: the base classes and the new classes. The base-to-new generalization task requires the model to train on base classes and then separately test on base classes and new classes. It is just the same as the conventional ZSL when the model is tested on new classes.

Baselines. We take several recent prompt-tuning methods as baselines, including CoOp, CoCoOp, MaPLe (Khattak et al. 2023), CLIP-Adapter (Gao et al. 2021), VPT (Derakhshani et al. 2022) and SHIP (Wang et al. 2023).

Main Results. As shown in Table 2, most of the prompt-tuning methods improve the performance of CLIP in the base classes, while they demonstrate limited performance gain or even degradation for new classes. This may be due to the fact that training only on the base classes leads the model to overfit the base classes. To mitigate overfitting to the base classes, SHIP generates the new-class data via the pre-trained CLIP encoders. However, the CLIP features may not be optimal due to the domain gap between the pre-training data and the downstream task data, leading to sub-optimal results. By contrast, our method generates new-class data based on the proposed FLTP method that further align the visual and textual features through multi-modal prompt-tuning strategy. The promising results achieved by FLPT+TFVAEGAN show that the further aligned features are better for training an effective conditional generator.

			AWA2			APY			CUB			SUN			FLO		
			Base	**New**	**H**	**Base**	**New**	**H**	**Base**	**New**	**H**	**Base**	**New**	**H**	**Base**	**New**	**H**
Resnet-101	Real-Data	HSVA	79.8	56.7	66.3	-	-	-	58.3	52.7	55.3	48.6	39.0	43.3	-	-	-
		MSDN	74.5	62.0	67.7	-	-	-	67.5	**68.7**	68.1	34.2	52.2	41.3	-	-	-
		f-CLSWGAN	61.4	57.9	59.6	-	-	-	57.7	3.7	49.7	36.6	42.6	39.4	73.8	59.0	65.6
		Cycle-WGAN	63.4	59.6	59.8	-	-	-	59.3	47.9	53.0	33.8	47.2	39.4	69.2	61.6	65.2
		LisGAN	76.3	52.6	62.3	-	-	-	57.9	46.5	51.6	37.8	42.9	40.2	83.8	57.7	68.3
		f-VAEGAN	70.6	57.6	63.5	-	-	-	60.1	48.4	53.6	38.0	45.1	41.3	74.9	56.8	64.6
		TCN	65.8	61.2	63.4	64.0	24.1	35.1	52.0	52.6	52.3	37.3	31.2	34.0	-	-	-
		GCM-CF	75.1	60.4	67.0	56.8	37.1	44.9	59.7	61.0	60.3	37.8	47.9	42.2	-	-	-
		TF-VAEGAN	75.1	59.8	66.6	61.5	31.7	41.8	64.7	52.8	58.1	40.7	45.6	43.0	84.1	62.5	71.7
	Data-Free	AZSL	3.7	3.5	3.6	4.0	6.8	5.1	-	-	-	-	-	-	-	-	-
	Data-Free*	AZSL	44.3	27.3	33.7	52.5	17.9	26.7	-	-	-	-	-	-	-	-	-
CLIP	Real-Data	f-VAEGAN	95.9	61.2	74.7	-	-	-	82.2	22.5	35.3	-	-	-	**97.6**	11.1	20.0
		TF-VAEGAN	**96.3**	43.7	60.1	71.7	22.3	34.0	**84.4**	21.1	34.0	51.4	61.4	55.9	97.2	37.4	54.0
		CoOp	95.3	72.7	82.5	85.4	76.1	80.5	63.8	49.2	55.6	61.3	61.8	61.6	85.8	52.2	64.9
		SHIP+CoOp	94.4	84.1	89.0	-	-	-	58.9	55.3	57.1	-	-	-	76.3	69.0	72.4
	Data-Free	CLIP*	93.0	88.2	90.6	81.6	75.8	78.6	56.3	56.1	56.2	51.2	55.9	53.5	69.4	67.9	69.6
	Data-Free	FLPT+TF-VAEGAN	93.9	93.2	93.5	84.2	**81.1**	82.6	66.1	**60.9**	63.4	60.9	**65.6**	63.2	89.0	78.2	83.2
	Data-Free*	FLPT+TF-VAEGAN	93.9	**93.6**	**93.7**	**84.4**	81.0	**82.7**	70.4	60.8	**65.2**	**63.8**	62.8	**63.3**	89.7	**79.5**	**84.3**

Table 1: Generalized zero-shot learning. The model is trained on the base classes and is evaluated on the **mixture** of base classes and new classes. 'Base' indicates the base-class results, 'New' indicates the new-class results, and 'H' is the harmonic mean. 'Data-Free' represents the black-box scenario. 'Data-Free*' means the white-box scenario. 'CLIP*' means that the hand-crafted prompt templates are used.

		11 Dataset Average			ImageNet			Caltech101			OxfordPets		
		Base	**New**	**H**	**Base**	**New**	**H**	**Base**	**New**	**H**	**Base**	**New**	**H**
Real-Data	CoOp	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
	CoCoOp	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43
	MaPLe	82.28	75.14	78.55	76.66	70.54	**73.47**	97.74	94.36	96.02	95.43	97.76	96.58
	CLIP-Adapter	83.05	65.20	73.05	75.74	68.21	71.78	98.13	92.19	95.39	91.55	90.10	90.82
	CoOp + VPT	71.98	74.76	73.34	74.73	**70.60**	72.60	95.47	93.80	94.62	90.77	97.83	94.16
	SHIP + CoOp	80.03	73.69	76.73	75.87	69.95	72.79	97.55	95.20	96.36	95.37	97.87	96.61
	SHIP + CLIP-Adapter	83.14	67.77	74.67	76.00	69.32	72.51	97.68	95.09	96.37	92.19	93.85	93.01
Data-Free	CLIP*	69.34	74.22	71.70	72.43	68.14	70.22	96.84	94.00	95.40	91.17	97.26	94.12
	FLPT	78.08	75.46	76.85	74.06	68.74	71.30	97.87	**96.29**	97.07	95.59	97.76	96.66
	FLPT+TFVAEGAN	**83.91**	**76.21**	**79.71**	**76.99**	68.22	72.34	**98.64**	96.18	**97.40**	**96.49**	**98.21**	**97.34**

		StanfordCars			Flowers102			Food101			FGVCAircraft		
		Base	**New**	**H**	**Base**	**New**	**H**	**Base**	**New**	**H**	**Base**	**New**	**H**
Real-Data	CoOp	78.12	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
	CoCoOp	70.49	73.59	72.01	94.87	71.75	81.71	90.70	91.29	90.99	33.41	23.71	27.74
	MaPLe	72.94	74.00	73.47	95.92	72.46	82.56	90.71	92.05	91.38	37.44	35.61	36.50
	CLIP-Adapter	**79.16**	59.49	67.93	**98.29**	64.68	78.02	88.24	88.33	88.29	42.14	25.67	31.91
	CoOp + VPT	65.27	**75.97**	70.21	72.97	75.90	74.40	90.37	91.67	91.01	29.57	33.80	31.54
	SHIP + CoOp	68.57	73.90	71.14	94.02	74.40	83.06	90.54	91.03	90.78	34.27	32.33	33.28
	SHIP + CLIP-Adapter	78.51	62.52	69.61	98.20	65.89	78.86	88.63	87.07	87.84	42.26	30.05	35.13
Data-Free	CLIP*	63.37	74.89	68.65	72.08	77.80	74.83	90.10	91.22	90.66	27.19	**36.29**	31.09
	FLPT	65.24	75.74	70.10	88.79	76.52	82.20	90.74	92.09	91.41	32.89	33.17	33.03
	FLPT+TFVAEGAN	77.06	75.41	**76.23**	94.11	**78.65**	**85.69**	**91.71**	**92.10**	**91.90**	**45.26**	32.81	**38.04**

		SUN397			DTD			EuroSAT			UCF101		
		Base	**New**	**H**	**Base**	**New**	**H**	**Base**	**New**	**H**	**Base**	**New**	**H**
Real-Data	CoOp	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
	CoCoOp	79.74	76.86	78.27	77.01	56.00	64.85	87.49	60.04	71.21	82.33	73.45	77.64
	MaPLe	80.82	**78.70**	**79.75**	80.36	59.18	68.16	94.07	73.23	82.35	83.00	**78.66**	**80.77**
	CLIP-Adapter	79.44	66.81	72.58	**81.94**	39.49	53.30	93.45	54.41	68.78	85.42	67.77	75.58
	CoOp + VPT	73.77	77.90	75.77	57.67	58.70	58.18	67.97	71.63	69.75	73.23	74.63	73.92
	SHIP + CoOp	79.54	75.27	77.35	74.88	56.88	64.65	88.62	66.87	76.22	81.08	76.85	78.91
	SHIP + CLIP-Adapter	79.86	66.52	72.58	81.60	46.38	59.14	93.05	57.15	70.81	**86.61**	71.61	78.40
Data-Free	CLIP*	69.36	75.35	72.23	53.24	59.90	56.37	56.48	64.05	60.03	70.53	77.50	73.85
	FLPT	77.84	76.25	77.04	70.37	62.68	66.30	86.00	79.54	82.64	79.52	75.66	77.55
	FLPT+TF-VAEGAN	**82.23**	76.23	79.12	80.32	**64.37**	**71.47**	**95.74**	**79.97**	**87.15**	84.50	76.21	80.13

Table 2: Base-to-new generalization. The model is trained on the base classes and is evaluated on the base classes and new classes independently. 'Base' indicates the base-class results, 'New' indicates the new-class results, and 'H' is the harmonic mean. 'Data-Free' represents the black-box scenario. 'Data-Free*' means the white-box scenario. 'CLIP*' means that the hand-crafted prompt templates are used.

		AWA2			APY			CUB			SUN			FLO		
		Base	**New**	**H**	**Base**	**New**	**H**	**Base**	**New**	**H**	**Base**	**New**	**H**	**Base**	**New**	**H**
Data-Free	CLIP*	93.04	88.21	90.57	81.63	75.76	78.58	56.29	56.12	56.21	51.20	55.90	53.45	69.39	67.86	69.62
	CoOp	94.35	85.27	89.58	82.21	78.47	80.30	51.90	51.73	51.82	55.78	53.19	54.45	76.82	65.30	70.59
	FLPT	93.57	92.92	93.24	83.00	80.11	81.82	57.26	57.75	57.50	56.94	59.72	58.30	70.54	72.64	71.57
Real-Data	FLPT	93.84	93.60	93.72	84.50	79.95	82.16	62.22	59.44	60.80	57.64	64.10	60.69	71.61	73.20	72.40
Data-Free	CLIP*+ZLAP	94.05	92.79	93.42	81.97	76.90	79.36	63.13	61.52	62.31	67.52	51.39	58.36	87.62	68.39	76.82
	CoOp+ZLAP	94.50	88.76	91.54	83.11	79.69	81.36	64.43	60.76	62.54	66.09	52.85	58.73	89.64	68.57	77.70
	FLPT+ZLAP	94.37	93.69	94.03	83.99	80.59	82.25	65.80	60.70	63.14	65.89	62.29	64.04	91.53	77.28	83.80
Real-Data	FLPT+ZLAP	92.66	96.05	94.32	83.54	81.91	82.72	66.15	62.80	64.43	59.53	79.38	68.04	96.51	74.27	83.94
Data-Free	CLIP*+SDGZSL	93.14	93.69	93.41	82.02	76.37	79.09	59.14	54.51	56.73	58.10	59.79	58.93	94.73	62.68	75.45
	CoOp+SDGZSL	93.43	92.15	92.79	82.14	79.31	80.70	57.04	53.20	55.05	58.99	61.18	60.07	94.37	66.23	77.83
	FLPT+SDGZSL	92.47	95.30	93.87	84.10	80.20	82.10	63.18	65.38	64.26	59.50	68.26	63.58	92.40	74.12	82.26
Real-Data	FLPT+SDGZSL	93.94	93.03	93.48	83.72	81.20	82.44	74.46	62.09	67.72	64.11	76.32	69.68	96.34	74.17	83.81
Data-Free	CLIP*+TF-VAEGAN	93.88	89.23	91.49	81.93	78.22	80.04	65.15	53.14	58.54	63.06	62.85	62.95	90.61	65.96	76.34
	CoOp+TF-VAEGAN	94.51	88.50	91.41	83.01	79.39	81.16	63.04	56.72	59.71	60.31	61.46	60.88	89.63	67.81	77.21
	FLPT+TF-VAEGAN	93.86	93.16	93.51	84.18	81.12	82.62	66.06	60.92	63.39	60.93	65.56	63.16	88.98	78.15	83.22
Real-Data	FLPT+TF-VAEGAN	93.84	93.59	93.71	84.56	80.75	82.61	73.25	58.45	65.02	62.87	70.97	66.67	94.11	77.58	85.05

Table 3: Ablation Study. Comparison results of different prompt-tuning methods, different generative models, and 'Data-Free' v.s. 'Real-Data' settings. 'CLIP*' means the hand-crafted prompt templates are used.

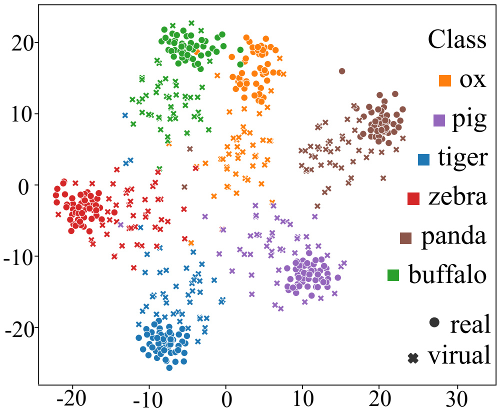


Figure 2: Visualization of tSNE of the real source data and recovered virtual data.

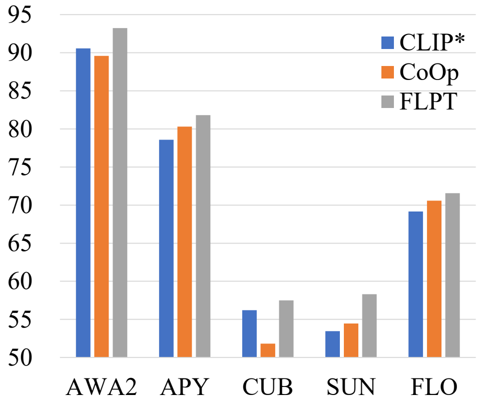


Figure 3: Comparisons of the proposed FLPT method with the baseline prompt tuning methods.

Ablation Study

We conducted ablation studies under the setting of generalized zero-shot learning to demonstrate the validity and generalizability of our proposed framework.

Quality of the Recovered Virtual Data. We visualized the recovered virtual base data and the real base data of 6 classes in AWA2 in Figure 2. It can be seen that the recovered data exhibits a distribution similar to that of the real data and possesses sufficient class discriminative qualities. To further validate the quality of the recovered image features and make a fairer comparison with baselines, we performed the experiments on real and virtual data, respectively. As shown in Table 3, It can be seen that the performance gap between the real data and virtual data is small, especially for AWA2, APY and FLO, which validate the effectiveness of the proposed base class data recovery method. The gap is slightly larger for CUB and SUN datasets, which is due to their fine-grained and challenging nature.

Comparisons of Different Prompt-tuning Methods. To evaluate the proposed FLPT that enhances the image features and text features, we compare FLPT to hand-crafted prompts in CLIP and the learning-based CoOp in Figure 3 and Table 3. It can be seen that FLPT outperforms the other two on all the five GZSL benchmarks. Furthermore, the performance is further improved after applying the enhanced features to the generative-model based methods. Compared to hand-crafted prompts, FLPT is a data-driven method, which costs less and is more effective. While compared to CoOp, FLPT makes a link between two modalities and aligns both image features and text features simultaneously.

Different Generative-Model-Based Methods. The proposed FLPT method can be combined with any generative models to further improve the ZSL performance. We evaluate our method with three generative methods: GAN-based ZLAP (Chen et al. 2022a), VAE-based SDGZSL (Chen et al. 2021c), and VAEGAN-based TF-VAEGAN (Narayan et al. 2020). As shown in Table 3, when integrating FLPT with the three types of generative models, all of them exhibit improved performance compared to FLPT alone.

Conclusion

This paper addresses a challenging and practical problem dubbed as data-free zero-shot learning (DFZSL). In DFZSL, the use of real images from both the base classes and the new classes is not necessary, thereby effectively preserving data copyright and privacy. To tackle DFZSL, we propose a CLIP-based framework, which consists of three main stages. Firstly, the virtual base-class data are recovered via a modeled von Mises-Fisher distribution based on the pre-trained CLIP classifier. Secondly, we propose a feature-language prompt tuning method to further align the virtual image features and textual features. Thirdly, to achieve better zero-shot classification, we generate the new-class data by training a conditional generative model based on the well aligned base-class multi-modal features. Extensive experiments on both base-to-new ZSL and generalized ZSL demonstrate the effectiveness of the proposed framework.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2021YFB1714300), National Natural Science Foundation of China (No.62006012, No.62132001, No.62002012), in part by the Hong Kong Research Grants Council General Research Fund (17203023), in part by The Hong Kong Jockey Club Charities Trust under Grant 2022-0174, in part by the Startup Funding and the Seed Funding for Basic Research for New Staff from The University of Hong Kong, and in part by the funding from UBTECH Robotics.

References

Bahng, H.; Jahanian, A.; Sankaranarayanan, S.; and Isola, P. 2022. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, 446–461. Springer.

Chen, D.; Shen, Y.; Zhang, H.; and Torr, P. H. 2022a. Zero-shot logit adjustment. *arXiv preprint arXiv:2204.11822*.

Chen, H.; Wang, Y.; Xu, C.; Yang, Z.; Liu, C.; Shi, B.; Xu, C.; Xu, C.; and Tian, Q. 2019. Data-free learning of student networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3514–3522.

Chen, S.; Hong, Z.; Xie, G.-S.; Yang, W.; Peng, Q.; Wang, K.; Zhao, J.; and You, X. 2022b. Msdn: Mutually semantic distillation network for zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7612–7621.

Chen, S.; Xie, G.; Liu, Y.; Peng, Q.; Sun, B.; Li, H.; You, X.; and Shao, L. 2021a. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. *Advances in Neural Information Processing Systems*, 34: 16622–16634.

Chen, Z.; Luo, Y.; Qiu, R.; Wang, S.; Huang, Z.; Li, J.; and Zhang, Z. 2021b. Semantics disentangling for generalized zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8712–8720.

Chen, Z.; Luo, Y.; Qiu, R.; Wang, S.; Huang, Z.; Li, J.; and Zhang, Z. 2021c. Semantics disentangling for generalized zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8712–8720.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Derakhshani, M. M.; Sanchez, E.; Bulat, A.; da Costa, V. G. T.; Snoek, C. G.; Tzimiropoulos, G.; and Martinez, B. 2022. Variational prompt tuning improves generalization of vision-language models. *arXiv preprint arXiv:2210.02390*.

Elman, J. L. 1990. Finding structure in time. *Cognitive science*, 14(2): 179–211.

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. 2009. Describing objects by their attributes. In *2009 IEEE conference on computer vision and pattern recognition*, 1778–1785. IEEE.

Fei-Fei, L. 2004. Learning generative visual models from few training examples. In *Workshop on Generative-Model Based Vision, IEEE Proc. CVPR, 2004*.

Felix, R.; Reid, I.; Carneiro, G.; et al. 2018. Multi-modal cycle-consistent generalized zero-shot learning. In *Proceedings of the European conference on computer vision (ECCV)*, 21–37.

Gan, C.; Lin, M.; Yang, Y.; Zhuang, Y.; and Hauptmann, A. G. 2015. Exploring semantic inter-class relationships (sir) for zero-shot action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

Gao, R.; Wan, F.; Organisciak, D.; Pu, J.; Wang, J.; Duan, H.; Zhang, P.; Hou, X.; and Long, Y. 2022. Absolute Zero-Shot Learning. *arXiv preprint arXiv:2202.11319*.

Han, Z.; Fu, Z.; Chen, S.; and Yang, J. 2021. Contrastive embedding for generalized zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2371–2381.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226.

Jiang, H.; Wang, R.; Shan, S.; and Chen, X. 2019a. Transferable contrastive network for generalized zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9765–9774.

Jiang, H.; Wang, R.; Shan, S.; and Chen, X. 2019b. Transferable contrastive network for generalized zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9765–9774.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19113–19122.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 554–561.

Larsen, A. B. L.; Sønderby, S. K.; Larochelle, H.; and Winther, O. 2016. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, 1558–1566. PMLR.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, J.; Jing, M.; Lu, K.; Ding, Z.; Zhu, L.; and Huang, Z. 2019. Leveraging the invariant side of generative zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7402–7411.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Narayan, S.; Gupta, A.; Khan, F. S.; Snoek, C. G.; and Shao, L. 2020. Latent embedding feedback and discriminative features for zero-shot classification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, 479–495. Springer.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, 722–729. IEEE.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, 3498–3505. IEEE.

Patterson, G.; and Hays, J. 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE conference on computer vision and pattern recognition*, 2751–2758. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Reed, S.; Akata, Z.; Lee, H.; and Schiele, B. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 49–58.

Romera-Paredes, B.; and Torr, P. 2015. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, 2152–2161. PMLR.

Shigeto, Y.; Suzuki, I.; Hara, K.; Shimbo, M.; and Matsumoto, Y. 2015. Ridge regression, hubness, and zero-shot learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I 15*, 135–151. Springer.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Tian, J.; Zhang, J.; Li, W.; and Xu, D. 2021. VDM-DA: Virtual domain modeling for source data-free domain adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6): 3749–3760.

Wang, Z.; Liang, J.; He, R.; Xu, N.; Wang, Z.; and Tan, T. 2023. Improving zero-shot generalization for clip with synthesized prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3032–3042.

Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; and Perona, P. 2010. Caltech-UCSD birds 200.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2018a. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9): 2251–2265.

Xian, Y.; Lorenz, T.; Schiele, B.; and Akata, Z. 2018b. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5542–5551.

Xian, Y.; Sharma, S.; Schiele, B.; and Akata, Z. 2019a. f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10275–10284.

Xian, Y.; Sharma, S.; Schiele, B.; and Akata, Z. 2019b. f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10275–10284.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.

Ye, Z.; Hu, F.; Lyu, F.; Li, L.; and Huang, K. 2022. Disentangling Semantic-to-Visual Confusion for Zero-Shot Learning. *IEEE Transactions on Multimedia*, 24: 2828–2840.

Ye, Z.; Lyu, F.; Li, L.; Fu, Q.; Ren, J.; and Hu, F. 2019. SR-GAN: Semantic rectifying generative adversarial network for zero-shot learning. In *2019 IEEE international conference on multimedia and expo (ICME)*, 85–90. IEEE.

Yue, Z.; Wang, T.; Sun, Q.; Hua, X.-S.; and Zhang, H. 2021. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15404–15414.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.