Visual-Augmented Dynamic Semantic Prototype for Generative Zero-Shot Learning

Wenjin Hou[1,4], Shiming Chen[2*], Shuhuang Chen[1], Ziming Hong[3], Yan Wang[4], Xuetao Feng[4], Salman Khan[2,5], Fahad Shahbaz Khan[2,6], Xinge You[1*]
[1]Huazhong University of Science and Technology (HUST), China
[2]Mohamed bin Zayed University of AI [3]The University of Sydney
[4]Alibaba Group [5]Australian National University [6]Linköping University
{houwj17, gchenshiming}@gmail.com youxg@mail.hust.edu.cn

Abstract

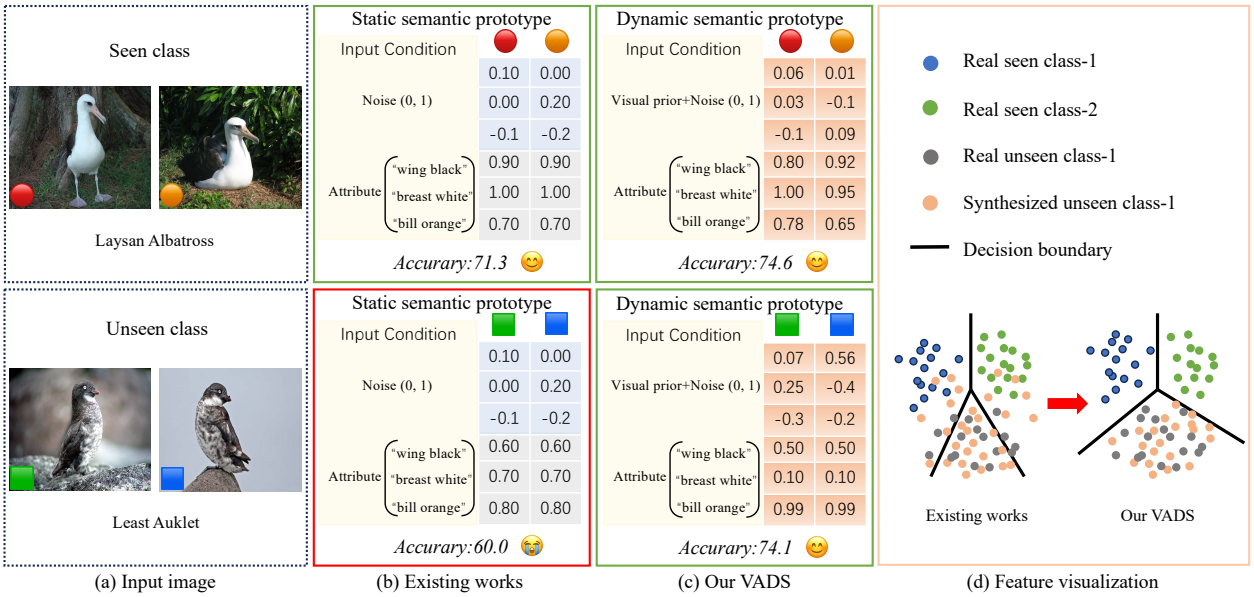
Generative Zero-shot learning (ZSL) learns a generator to synthesize visual samples for unseen classes, which is an effective way to advance ZSL. However, existing generative methods rely on the conditions of Gaussian noise and the predefined semantic prototype, which limit the generator only optimized on specific seen classes rather than characterizing each visual instance, resulting in poor generalizations (e.g., overfitting to seen classes). To address this issue, we propose a novel Visual-Augmented Dynamic Semantic prototype method (termed VADS) to boost the generator to learn accurate semantic-visual mapping by fully exploiting the visual-augmented knowledge into semantic conditions. In detail, VADS consists of two modules: (1) Visual-aware Domain Knowledge Learning module (VDKL) learns the local bias and global prior of the visual features (referred to as domain visual knowledge), which replace pure Gaussian noise to provide richer prior noise information; (2) Vision-Oriented Semantic Updation module (VOSU) updates the semantic prototype according to the visual representations of the samples. Ultimately, we concatenate their output as a dynamic semantic prototype, which serves as the condition of the generator. Extensive experiments demonstrate that our VADS achieves superior CZSL and GZSL performances on three prominent datasets and outperforms other state-of-the-art methods with averaging increases by 6.4%, 5.9% and 4.2% on SUN, CUB and AWA2, respectively.

1. Introduction

Zero-shot learning [33], which transfers knowledge from seen classes to unseen classes, has garnered much attention recently. By establishing interactions between visual features and semantic prototypes (also referred to as attribute vectors, side information, or semantic embeddings [47]), generative ZSL methods exhibit impressive performance, demonstrating the potential of feature synthesis. One of the most successful frameworks is the conditional generative adversarial network (GAN) [17]. The main idea of generative ZSL methods is to align semantic prototypes and visual features to synthesize feature of unseen classes. Recent emerging studies have either designed more effective frameworks [7, 19, 20, 31, 46] or addressed more specific issues related to visual-semantic alignment [4, 8, 12, 43, 53]. These methods have achieved significant improvements.

However, these methods rely on the conditions of Gaussian noise and the predefined semantic prototype (referred to as the static semantic prototype), which limit the generator only optimized on specific seen classes rather than characterizing each visual instance, resulting in poor generalizations (*e.g.*, overfitting to seen classes). Fig. 1 illustrates these issues: (1) The noise is sampled from a Gaussian distribution $\mathcal{N}(0, 1)$, which lacks the dataset-specific visual prior knowledge (*e.g.*, global visual information "flying" and "still" and background information "sky" and "grass"). As a result, the domain knowledge shared between seen and unseen classes cannot be utilized for feature synthesis of unseen classes, limiting the knowledge transfer. (2) The predefined semantic prototype fails to characterize each instance well. For example, the attributes "wing black", "breast white" and "bill orange" of the Laysan Albatross are not fixed on different images. Due to these limitations, the visual features synthesized by existing works [4, 6, 21, 24, 31, 43, 46] struggle to represent the distribution of real features, leading to poor generalization to unseen classes, as shown in Fig. 1(b). More intuitively, as shown in Fig. 1(d), the features of unseen classes synthesized by these methods are confusing, resulting in the decision boundary overfitting to the seen classes.

*Corresponding authors



(a) Input image (b) Existing works (c) Our VADS (d) Feature visualization

Figure 1. An illustration of the core idea of our method. (a) The semantic prototype (i.e., attribute) of different images of the same category is not fixed, so the predefined semantic prototype is inaccurate in characterizing each instance. (b) Most existing works utilize Gaussian noise and the predefined semantic prototype as conditions to train a semantic→visual generator on seen classes, which fails to generalize to unseen classes. (c)(d) Our method incorporates rich visual prior with an updated semantic prototype to construct a visual-augmented dynamic semantic prototype of each instance, empowering the generator to synthesize features that faithfully represent the real distribution of unseen classes. Thus, our method achieves better generalization on seen and unseen classes than existing works (*e.g.*, CLSWGAN [46]).

Drawing inspiration from image captioning [36], which highlights the generalization of instance-conditional learning, we aim to fully leverage the visual-augmented knowledge into semantic conditions to tackle the aforementioned challenges. On the one hand, we can exploit rich domain visual prior knowledge, serving as a prior noise, to enhance the adaptation and generalization of models [18, 51, 54]. On the other hand, we can update the predefined semantic prototype to align visual representations based on visual features. As such, incorporating richer and more accurate visual information acts as the semantic condition to train an instance-conditional generative model, which is optimized to characterize each instance (more robust to class shift) rather than to serve only for specific classes. Accordingly, the generative model can synthesize features of unseen classes closer to the real ones, facilitating the classifier in learning an appropriate decision boundary (see the right of Fig. 1(d)).

In this paper, we propose an approach called **V**isual-**A**ugmented **D**ynamic **S**emantic prototype (**VADS**) to improve generative ZSL methods. Specifically, VADS consists of two learnable modules: a Visual-aware Domain Knowledge Learning module (VDKL) and a Vision-Oriented Semantic Updation module (VOSU). The VDKL explores domain visual prior knowledge derived from visual information, which provides richer information for representing instances. The VOSU predicts instance-level semantics through visual→semantic mapping, guiding the updation of the predefined semantic prototype and promoting accurate semantic prototype learning. Finally, the extracted visual prior and the updated semantic prototype are concatenated as a visual-augmented dynamic semantic prototype, which serves as the condition of the generator during training and feature synthesis, as illustrated in Fig. 1(c). Extensive experiments demonstrate the effectiveness of our VADS.

Our contributions can be summarized as follows:

- We introduce a Visual-Augmented Dynamic Semantic prototype (VADS) to enhance the generalization of generative ZSL methods, facilitating substantial knowledge transfer.
- We devise the VDKL to leverage domain visual prior knowledge from visual features and design the VOSU to dynamically update the predefined semantic prototype. Their outputs together serve as the generator's conditions, providing richer and more accurate visual information.
- We conduct extensive experiments on AWA2 [47], SUN [34] and CUB [44] datasets. The comprehensive results demonstrate that visual prior knowledge significantly improves the generalization of generative ZSL methods, *i.e.*, average improvements of the harmonic mean over existing generative methods (*e.g.*, f-CLSWGAN [46], TF-VAEGAN [31] and FREE [7]) 6.4%, 5.9% and 4.2% on SUN, CUB and AWA2, respectively.

2. Related Work

Embedding-based Zero-Shot Learning. Embedding-based ZSL methods are one of the mainstream branches that project visual information into semantic space to align with semantic prototypes. Earlier works [25, 39, 48] directly mapped global visual features to semantic space, failing to capture local discriminative representation, resulting in sub-optimal ZSL performance. Also, embeddings are learned only in seen classes, leading to inevitable seen class bias. In this regard, some studies [9, 10, 22] have attempted to use calibration loss to balance the prediction results between seen and unseen classes. Recently, attention mechanisms [41] have emerged with surprising localization abilities, so semantic-guided methods [9, 11, 29, 30, 32, 42, 48, 49] learn to discover attribute-related local regions, providing more accurate inter-class discrimination. Among these methods, APN [49] proposed an attribute prototype network to learn local features, and DPPN [42] updated attribute and category prototypes. Inspired by their work, we introduce a dynamic semantic prototype for generative ZSL methods.

Generative Zero-Shot Learning. Generative ZSL methods learn semantic→visual mapping to synthesize unseen class features, effectively alleviating the lack of unseen class data. Consequently, the quality of synthesized features, which preserves visual-semantic correspondence, plays a crucial role in classification. Thus, TF-VAEGAN [31] forced semantic alignment at all stages, and FREE [7] fine-tuned visual features to address cross-dataset biases. CE-GZSL [20] and ICCE [24] projected visual features into the latent space for classification. However, these methods constructed projection spaces on seen classes, resulting in inferior generalization ability on unseen classes. Moreover, they uniformly utilize the predefined semantic prototype as a condition, making it difficult to achieve accurate visual-semantic alignment. The method most related to ours is DSP [12], which updates the prototype by simply adding the evolved and predefined semantic prototype.

Large-Scale Vision-Language Models Generalization. Vision-language models like CLIP [35], pre-trained on large-scale image-text pairs, have demonstrated significant potential for downstream tasks. When performing zero-shot recognition, the class prompts are input into the text encoder to obtain the classification weights, and the cosine similarity between the test image and the weights determines the resulting classification score. It is different from the classical ZSL methods [9, 25, 39, 42, 48, 49]. Recent research has focused on improving the generalization to unseen classes, with several previous works proposing prompt learning [1, 51, 54]. Motivated by optimizing visual conditional prompts, we introduce visual-aware domain knowledge learning into generative ZSL methods, facilitating knowledge transfer to unseen classes.

3. Visual-Augmented Dynamic Semantic Prototype Method

Fig. 2 shows the framework of our VADS. Next, we first present the problem formulation and briefly review the generative ZSL model. Then, we introduce the detailed design of our method.

Problem Formulation. Conventional zero-shot learning (CZSL) recognizes unseen classes in the inference stage. Generalized zero-shot learning (GZSL) recognizes both seen and unseen classes. Both settings generalize from seen data $\mathcal{D}^s$ to unseen domains $\mathcal{D}^u$. $\mathcal{D}^s = \{(x_i^s, y_i^s)|x_i^s \in \mathcal{X}^s, y_i^s \in \mathcal{Y}^s\}_{i=1}^{N_s}$, where N_s is the sample number of seen classes, x_i^s is a feature vector in $\mathcal{X}^s$ and y_i^s is the corresponding lable from $\mathcal{Y}^s$. The $\mathcal{D}^s$ is split into a training set $\mathcal{D}_{tr}^s$ and a testing set $\mathcal{D}_{te}^s$ following Xian et al. [47]. Similarly, $\mathcal{D}^u = \{(x_i^u, y_i^u)|x_i^u \in \mathcal{X}^u, y_i^u \in \mathcal{Y}^u\}_{i=1}^{N_u}$, where x_i^u is a feature vector in $\mathcal{X}^u$ and y_i^u is the lable from $\mathcal{Y}^u$. $\mathcal{Y}^s$ and $\mathcal{Y}^u$ are disjoint. Define attribute semantic prototypes $\mathcal{A} = \mathcal{A}^s \cup \mathcal{A}^u$, corresponding to each category, as a bridge to transfer knowledge from seen classes to unseen classes. In this paper, we dynamically update $\mathcal{A}$ to learn accurate visual-semantic alignment.

3.1. Generative ZSL Model

The goal of the generative ZSL methods is to learn a semantic→visual generative model (G) on seen classes and then use it to synthesize samples of unseen classes to train a classifier. Existing methods use Gaussian noise and the predefined semantic prototype as input conditions to supervise G synthetic features (*i.e.*, $\mathcal{A} \times \mathcal{Z} \rightarrow \hat{X}$). In our method, G represents an off-the-shelf CLSWGAN [46], which contains a generator and a discriminator. We develop the dynamic semantic prototype as a condition, allowing G to characterize more accurate visual-semantic relationships.

3.2. Visual-aware Domain Knowledge Learning (VDKL)

Drawing inspiration from previous prompt learning [51, 54], we exploit the rich information in visual features to assist in synthesizing features. VDKL is a data-efficient module allowing the visual features to be used to improve generalization. As shown in Fig. 2, we design a Visual Encoder (VE) and a Domain Knowledge Learning network (DKL). First, the VE encodes visual features into a latent feature l and a latent code z. The latent feature enables inter-class alignment of visual features, and latent code is subsequently confined to a prior distribution $\boldsymbol{Z}$. The optimization of the VE is achieved via contrastive loss [13] and evidence-lower bound given by the equation as follows:

$$\mathcal{L}_{con} = \mathbb{E}[\log \frac{\exp(l_i^T l^+/\tau)}{\exp(l_i^T l^+/\tau) + \sum_{k=1}^{K} \exp(l_i^T l_k^-/\tau)}], \quad (1)$$

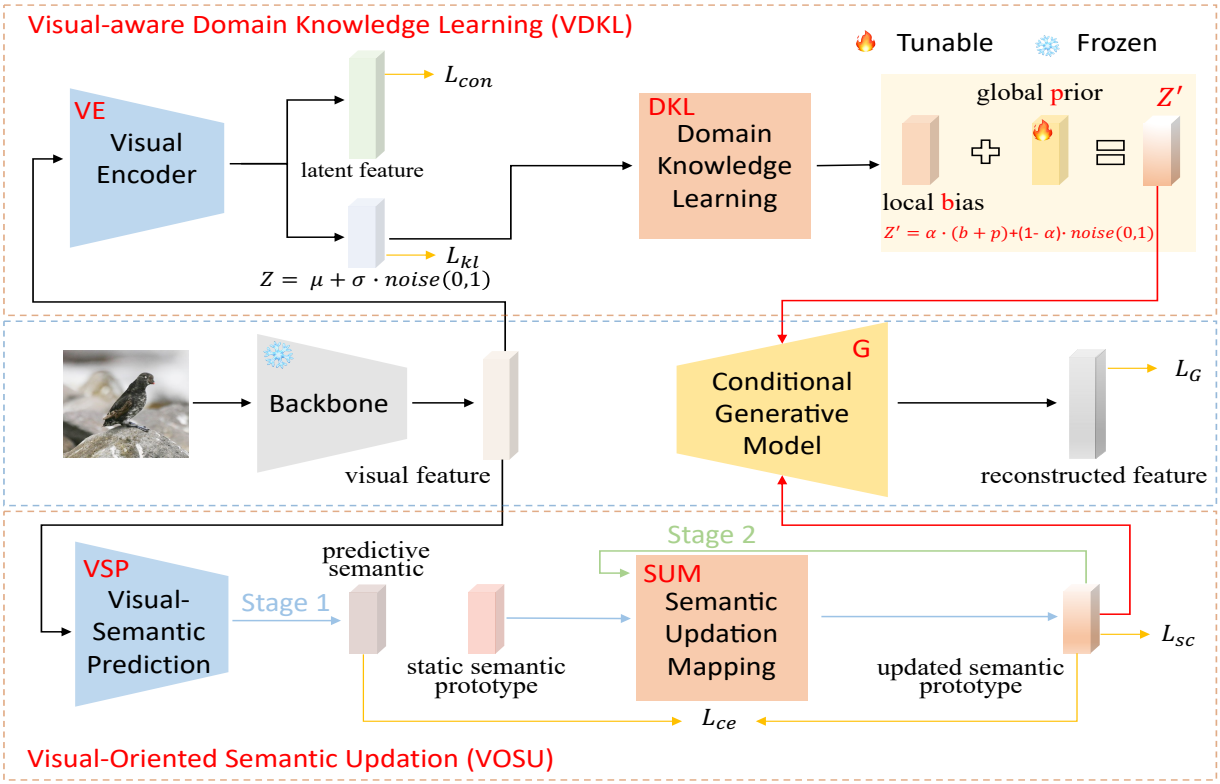


Figure 2. The architecture of our proposed VADS. It consists of two learnable modules: a Visual-Oriented Semantic Updation module (VOSU) and a Visual-aware Domain Knowledge Learning module (VDKL). First, we obtain the prior distribution $\boldsymbol{Z}$ by the Visual Encoder (VE). Following this, the Domain Knowledge Learning network (DKL) transforms $\boldsymbol{Z}$ into a local bias $\boldsymbol{b}$, which is subsequently added to global learnable prior vectors ($\boldsymbol{p}$) to construct the domain visual prior noise (*i.e.*, $\boldsymbol{Z}^{'}$). At the bottom, VOSU notably updates the semantic prototype in two stages (depicted by the blue and green arrows). Finally, the visual prior noise and the updated semantic prototype together form a dynamic semantic prototype, used for the reconstruction of features by the generator.

$$\mathcal{L}_{kl} = KL(VE(x)||p(z)), \tag{2}$$

where l^{+} and l_k^{-} represent positive and negative latent features, τ is a temperature parameter set as 0.15, K is the class numbers, KL denotes the Kullback-Leibler divergence and $p(z)$ is a prior distribution that is assumed to be $\mathcal{N}(0,1)$.

To further utilize visual prior knowledge during the training and synthesis stages, we propose a Domain Knowledge Learning network (DKL) to obtain a local bias $\boldsymbol{b}$ of visual features (*i.e.*, $\boldsymbol{b} = DKL(\boldsymbol{Z})$). Additionally, we employ a learnable prior vector $\boldsymbol{p}$ to capture global visual information ($\boldsymbol{p}$ is randomly initialized). Subsequently, we obtain domain-specific visual prior noise as follows:

$$\boldsymbol{Z}^{'} = \alpha \cdot (\boldsymbol{b} + \boldsymbol{p}) + (1 - \alpha) \cdot \boldsymbol{noise(0,1)}, \tag{3}$$

where the $\boldsymbol{noise(0,1)}$ represents Gaussian noise aimed at enhancing diversity in synthesis, α is the combination coefficient set as 0.9. Through this operation, we argue that $\boldsymbol{Z}^{'}$ includes rich domain visual knowledge and feeds it into the generator to provide instance conditions, promoting the generator learning and utilizing it for feature synthesis of unseen classes. Note that unseen class samples are unavailable in the feature synthesis stage, so we randomly sample Gaussian noise input to DKL, transferring the domain knowledge acquired from seen classes to unseen classes.

3.3. Visual-Oriented Semantic Updation (VOSU)

We observe that the predefined semantic prototype struggles to represent each visual sample accurately, so we propose a Visual-Oriented Semantic Updation module (VOSU), optimizing the semantic prototype dynamically. Our semantic prototype updation involves a two-stage process. In the first stage, we feed visual features x_s into the Visual-Semantic Prediction network (VSP) to generate a predictive semantic $\hat{a}$ that explicitly captures specific visual patterns of the target image. Then, the predefined semantic prototype is input into a Semantic Updation Mapping network (SUM) to learn an updated semantic $\dot{a}$. This mapping can be expressed as:

$$\dot{a} = SUM(a). \tag{4}$$

To maintain the attribute information of the prototype and integrate the visual information, we jointly optimize them by the cross-entropy loss $\mathcal{L}_{ce}$. $\mathcal{L}_{ce}$ is defined as:

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(VSP(x_i)^T\dot{a}^y)}{\sum_{\bar{c}\in C^{s\cup u}}\exp(VSP(x_i)^T\dot{a}^{\bar{c}})}, \quad (5)$$

where N is the batch numbers. Accordingly, the updated semantic prototype incorporates rich visual information. In the second stage, we employ SUM to update $\dot{a}$ during the conditional generative model training and use it as a condition to learn together with the generator G and the discriminator D. This implementation facilitates dynamic updation and accurate visual-semantic matching. To this end, we propose the semantic consistency loss $\mathcal{L}_{sc}$ as follows:

$$\mathcal{L}_{sc} = \mathbb{E}\left[\|SUM(\dot{a}) - \dot{a}\|_1\right]. \quad (6)$$

In summary, the first stage leverages visual features to assist semantic updation, and the second stage dynamically updates the prototype of each sample. Then, we concatenate the updated semantic prototype with the visual prior noise $Z^{'}$, called the dynamic semantic prototype, which serves as the condition for the generator, as depicted in Fig. 2.

3.4. Overall Objective and Inference

VADS Objective Loss Function. Overall, the objective loss function of VADS is:

$$\mathcal{L}_{total} = \mathcal{L}_G + \lambda_{con}\mathcal{L}_{con} + \lambda_{kl}\mathcal{L}_{kl} + \lambda_{sc}\mathcal{L}_{sc}, \quad (7)$$

where $\mathcal{L}_G$ is the loss of conditional generative model G, λ_{con}, λ_{kl} and λ_{sc} are the hyper-parameter to balance each loss term. To fully validate our method, by using this loss, we train on various mainstream generative models (*e.g.*, CLSWAGN [46], TFVAEGAN [31], and FREE [7]). Next, we illustrate feature synthesis and classifier training.

Visual-Augmented Feature Synthesis for Unseen Classes. To fully utilize the visual knowledge and accurate semantic prototypes, we sample Gaussian noise input DKL to obtain the prior noise $\boldsymbol{Z}^{'}$ (*i.e.,* Eq. (3))and use SUM to update the semantic prototypes of unseen classes (*i.e.*, $\dot{a}_u = SUM(\hat{a}_u)$). They serve as conditions to synthesize visual samples closer to the real features for training the classifier. The form can be written as:

$$\hat{x}_u = \mathrm{G}(\boldsymbol{Z}^{'}, \dot{a}_u), \quad (8)$$

where $\dot{a}_u$ is updated semantic prototype of the unseen classes and $\hat{x}_u$ is the synthesized features of unseen classes.

ZSL Classifier Training and Inference. After synthesizing features, we input the seen class training features and synthesized unseen class features into VE to extract latent features and concatenate them to enhance the original features, alleviating cross-dataset bias [7]. Then, we train a CZSL classifier using enhanced-synthetic features (i.e., $f_{CZSL} : \mathcal{X} \rightarrow \mathcal{Y}^u$) and train a GZSL classifier using enhanced seen class training features and enhanced-synthetic features (i.e., $f_{GZSL} : \mathcal{X} \rightarrow \mathcal{Y}^s \cup \mathcal{Y}^u$). Finally, we perform inference using the test sets $\mathcal{D}^s_{te}$ and $\mathcal{D}^u$.

4. Experiments

4.1. Experimental Setup

Benchmark Datasets. We conduct extensive experiments on three prominent ZSL benchmark datasets: Animals with Attributes 2 (AWA2 [47]), SUN Attribute (SUN [34]) and Caltech-USCD Birds-200-2011 (CUB [44]). We follow the Proposed Split (PS) setting [47] to split each dataset into seen and unseen classes, as detailed in Tab. 2.

Evaluation Protocols. During inference (*i.e.*, performing CZSL and GZSL classification), we follow the evaluation protocols in [47]. In the CZSL setting, we calculate the average per-class Top-1 accuracy of unseen classes, denoted as $\boldsymbol{Acc}$. For the GZSL scenario, we measure the Top-1 accuracy of the seen and unseen classes, defined as $\boldsymbol{S}$ and $\boldsymbol{U}$, respectively. We also compute the harmonic mean, defined as $\boldsymbol{H} = (2 \times \boldsymbol{S} \times \boldsymbol{U})/(\boldsymbol{S} + \boldsymbol{U})$.

Implementation Details. We follow PSVMA [26] using the ViT-Base Backbone [15] without fine-tuning as the feature extractor, obtaining 768-dimensional visual features for all samples. The global prior $\boldsymbol{p}$ has the same dimension as the semantic prototype. We set the mini-batch to 64, 128 and 128 for AWA2, SUN and CUB, respectively. We use the Adam optimizer [23] with β_1 = 0.5, β_2 = 0.999, and set the initial learning rate to 0.0001. We synthesize 5600, 100, and 400 samples for each class on AWA2, SUN and CUB. Our experiments are based on the PyTorch and implemented on a NVIDIA GeForce RTX 3090 GPU.

4.2. Comparison with State-of-the-Art Methods

We report the performance of our proposed VADS using CLSWGAN as a generative model compared to state-of-the-art methods. Tab. 1 shows the results, including the embedding-based and generative ZSL methods. In the CZSL scenario, our method notably outperforms the sub-optimal results by 8.4%, 10.3%, and 8.4%, achieving the best results on AWA2, SUN, and CUB. The results confirm that our method incorporating dynamic semantic prototypes is more generalizable to unseen classes than static semantic prototypes. In the GZSL scenario, our method obtains the best harmonic mean $\boldsymbol{H}$ on all datasets (*i.e.*, AWA2 ($\boldsymbol{H}$ = 79.3), SUN ($\boldsymbol{H}$ = 55.7) and CUB ($\boldsymbol{H}$ = 74.3)). Our method significantly outperforms CLSWGAN+DSP [12], which proposed an evolved semantic prototype, indicat-

Table 1. Compared our VADS with the state-of-the-art on AWA2, SUN and CUB benchmark datasets in the CZSL and GZSL settings. The best and second-best results are marked in **Red** and **Blue**, respectively. Symbol "–" denotes no results are reported.

Type	Methods	Venue	Backbone	AWA2				SUN				CUB			
				CZSL	GZSL			CZSL	GZSL			CZSL	GZSL		
				Acc	U	S	H	Acc	U	S	H	Acc	U	S	H
Embedding	PREN [52]	CVPR'19	ResNet-101	**74.1**	32.4	88.6	47.4	62.9	35.4	27.2	30.8	66.4	35.2	55.8	43.1
	DAZLE [22]	CVPR'20	ResNet-101	67.9	60.3	75.7	67.1	59.4	52.3	24.3	33.2	66.0	56.7	59.6	58.1
	DVBE [28]	CVPR'20	ResNet-101	–	63.6	70.8	67.0	–	45.0	37.2	40.7	–	53.2	60.2	56.5
	CN [38]	ICLR'21	ResNet-101	–	60.2	77.1	67.6	–	44.7	41.6	43.1	–	49.9	50.7	50.3
	GEM-ZSL [27]	CVPR'21	ResNet-101	67.3	64.8	77.5	70.6	62.8	38.1	35.7	36.9	77.8	64.8	77.1	70.4
	ViT-ZSL [3]	IMVIP'21	ViT-Large	–	51.9	**90.0**	65.8	–	44.5	**55.3**	49.3	–	67.3	75.2	71.0
	IEAM-ZSL [2]	DGAM'21	ViT-Large	–	53.7	89.9	67.2	–	48.2	**54.7**	51.3	–	68.6	73.8	71.1
	DUET [14]	AAAI'23	ViT-Base	69.9	63.7	84.7	72.7	64.4	45.7	45.8	45.8	72.3	62.9	72.8	67.5
	PSVMA [26]	CVPR'23	ViT-Base	–	**73.6**	77.3	**75.4**	–	**61.7**	45.3	**52.3**	–	**70.1**	**77.8**	**73.8**
Generative	f-VAEGAN-D2 [47]	CVPR'19	ResNet-101	71.1	57.6	70.6	63.5	64.7	45.1	38.0	41.3	61.0	48.4	60.1	53.6
	TF-VAEGAN [31]	ECCV'20	ResNet-101	72.2	59.8	75.1	66.6	**66.0**	45.6	40.7	43.0	64.9	52.8	64.7	58.1
	FREE [7]	ICCV'21	ResNet-101	–	60.4	75.4	67.1	–	47.4	37.2	41.7	–	55.7	59.9	57.7
	HSVA [8]	NeurIPS'21	ResNet-101	–	56.7	79.8	66.3	–	48.6	39.0	43.3	–	52.7	58.3	55.3
	CE-GZSL [20]	CVPR'21	ResNet-101	70.4	63.1	78.6	70.0	63.3	48.8	38.6	43.1	77.5	63.9	66.8	65.3
	FREE+ESZSL [5]	ICLR'22	ResNet-101	–	51.3	78.0	61.8	–	48.2	36.5	41.5	–	51.6	60.4	55.7
	ICCE [24]	CVPR'22	ResNet-101	72.7	65.3	82.3	72.8	–	–	–	–	**78.4**	67.3	65.5	66.4
	TDCSS [16]	CVPR'22	ResNet-101	–	59.2	74.9	66.1	–	–	–	–	–	44.2	62.8	51.9
	CDL + OSCO [4]	TPAMI'23	ResNet-101	–	48.0	71.0	57.1	–	32.0	65.0	42.9	–	29.0	69.0	40.6
	CLSWGAN+DSP [12]	ICML'23	ResNet-101	–	60.0	86.0	70.7	–	48.3	43.0	45.5	–	51.4	63.8	56.9
	TFVAEGAN+SHIP [43]	ICCV'23	ViT-Base	–	61.2	**95.9**	74.7	–	–	–	–	–	22.5	**82.2**	35.3
	VADS (Ours)	–	ViT-Base	**82.5**	**75.4**	83.6	**79.3**	**76.3**	**64.6**	49.0	**55.7**	**86.8**	**74.1**	74.6	**74.3**

Table 2. A detailed illustration of the ZSL benchmark datasets. s and u are the number of seen and unseen classes. N_A refers to semantic prototype dimensions. We follow CE-GZSL [20] using 1024-dimensional semantic prototypes generated from textual descriptions [37] on CUB.

Datasets	# images	# classes (s | u)	# N_A
AWA2 [47]	37322	50 (40 | 10)	85
SUN [34]	14340	717 (645 | 72)	102
CUB [44]	11788	200 (150 | 50)	1024

ing that our design is reasonable. Meanwhile, our method is also competitive compared to the recent PSVMA [26], DUET [14], and TFVAEGAN+SHIP [43] methods using the ViT Backbone. Furthermore, our method achieves optimal results in unseen class accuracy $\boldsymbol{U}$, demonstrating that the features synthesized by the generator are closer to the real features of unseen classes, effectively alleviating the over-fitting problem. Noted that TFVAEGAN+SHIP [43] using CLIP Encoder and ViT-ZSL [2] using ViT-Large achieve the best and second-best accuracy for seen classes, but they fail to generalize well to unseen classes. These results consistently demonstrate that our method synthesizes reliable features of unseen classes to facilitate classifier learning, resulting in superior ZSL performance.

4.3. Ablation Study

Component Analysis. In this section, we perform a series of experiments to analyze the effectiveness of significant components. Tab. 3 summarizes the results of ablation studies on CUB and AWA2. We first use ViT-Base to extract visual features to train CLSWAGN as the baseline. Compared to the baseline (*i.e.*, configuration (1)), our VADS (configuration (7)) performance improves significantly (*i.e.*, the $\boldsymbol{Acc/H}$ increases by 6.7%/9.1% on CUB and 11.5%/10% on AWA2). Configurations (2) and (3) correspond to the effectiveness of the VDKL and VOSU modules. When there is no VDKL, that is, no visual prior is introduced to generate samples, the performance drops by 1.9%/1.5% and 4.4%/0.8% for CUB and AWA2, indicating that visual prior is beneficial for transferring knowledge to unseen classes. When without VOSU, the performance drops most severely, verifying visual-semantic alignment is crucial for learning an accurate generator G. Next, without VE enhancing classification features, the results of configuration (4) drop slightly in both CZSL and GZSL settings, which shows that feature enhancement mitigates cross-dataset bias. Lastly, we conduct experiments to study the impact of the loss terms on performance. Configuration (5) without contrastive loss $\mathcal{L}_{con}$, $\boldsymbol{Acc/H}$ decreases

Table 3. Ablation study of VADS on modules, feature enhancement and loss terms on CUB and AWA2. We use CLSWGAN [45] as a generative model. The best result is marked in **boldface**.

Configurations	CUB		AWA2	
	Acc	H	Acc	H
(1) VADS w/o VDKL & VOSU	80.1	65.2	71.3	69.3
(2) VADS w/o VDKL	84.9	72.8	78.1	78.5
(3) VADS w/o VOSU	83.8	70.4	75.4	77.0
(4) VADS w/o enhancement	85.1	73.4	81.8	79.1
(5) VADS w/o $\mathcal{L}_{con}$ (*i.e.*, Eq. (1))	85.3	73.1	79.4	78.6
(6) VADS w/o $\mathcal{L}_{sc}$ (*i.e.*, Eq. (6))	86.0	73.5	79.5	78.7
(7) **VADS (full)**	**86.8**	**74.3**	**82.5**	**79.3**

Table 4. Evaluation of VADS with multiple generative ZSL models on three prominent datasets using ViT-Base Backbone. Each row pair shows the effect of adding VADS to a particular generative ZSL model. We use the same hyperparameter optimization policy in all cases to make results comparable.

Generative ZSL Methods	AWA2				SUN				CUB			
	CZSL	GZSL			CZSL	GZSL			CZSL	GZSL		
	Acc	U	S	H	Acc	U	S	H	Acc	U	S	H
CLSWGAN [46]	71.3	66.2	72.6	69.3	66.0	46.9	42.6	44.6	80.1	60.0	71.3	65.2
CLSWGAN + **VADS**	82.5$^{+11.5}$	75.4	83.6	79.3$^{+10.0}$	76.3$^{+10.3}$	64.6	49.0	55.7$^{+11.1}$	86.8$^{+6.7}$	74.1	74.6	74.3$^{+9.1}$
TFVAEGAN [31]	78.2	66.7	87.1	75.6	73.1	60.6	48.6	54.0	81.6	64.8	74.6	69.3
TFVAEGAN + **VADS**	80.2$^{+2.0}$	75.7	83.3	79.3$^{+3.7}$	76.3$^{+3.2}$	61.9	51.0	55.9$^{+1.9}$	83.6$^{+2.0}$	70.1	70.9	70.5$^{+1.2}$
FREE [7]	70.6	62.9	85.9	72.6	71.7	45.4	50.4	47.8	84.3	68.7	73.5	70.9
FREE + **VADS**	79.4$^{+8.8}$	70.1	84.6	76.6$^{+4.0}$	75.0$^{+3.3}$	57.6	50.7	53.9$^{+6.1}$	85.5$^{+1.2}$	70.9	75.4	73.1$^{+3.2}$

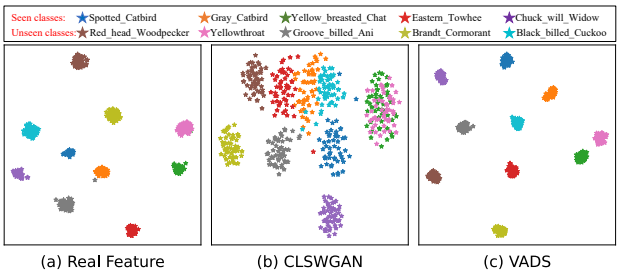


Figure 3. t-SNE visualizations on CUB. The 10 different colors refer to the 5 seen classes and 5 unseen classes that are randomly selected. Please zoom in for a better view.

by 1.5%/1.2% and 3.1%/0.7% on CUB and AWA2, respectively. The main reason is that contrastive loss achieves feature alignment, without which the network may learn category-agnostic redundant information. For the lack of semantic consistency loss $\mathcal{L}_{sc}$ (*i.e.*, configuration (6)), there is no guarantee that the semantic prototype maintains the original definition when training on seen classes, thus hurting the performance of seen classes. Note that the cross-entropy loss $\mathcal{L}_{ce}$ and KL loss $\mathcal{L}_{kl}$ are necessary for single network training, so we did not perform ablation.

Qualitative Evaluation. To further demonstrate that our method synthesizes reliable unseen class features for classifier training, we project various visual features into two principal components via t-SNE [40]. Fig. 3 shows the t-SNE visualization on CUB. Each color of '$\star$' represents a class, and we randomly select 5 seen classes and 5 unseen classes. From left to right, they represent the real visual features extracted by ViT-Base [15], the features synthesized by CLSWGAN, and the features synthesized by our VADS. We observe that the real visual features are inter-class dispersion and intra-class aggregation. In Fig. 3(b), Gaussian noise and the predefined semantic prototype serve as conditions to synthesize samples by CLSWGAN. There are two apparent phenomena: first, the feature distributions are scattered, which cannot truly reflect each class; second, the synthesized seen and unseen class features are confusing (*e.g.*, "$\star$" and "$\star$", which denote seen class "Yellow breasted Chat" and unseen class "Yellowthroat", respectively). Therefore, the decision boundary of the CZSL/GZSL classifier trained with these features is unclear, consistent with the motivation of Fig. 2(d). In contrast, the features generated with our VADS are closer to the real features and are inter-class separated, as shown in Fig. 3(c). On the one hand, we analyze that visual-augmented dynamic semantic prototype motivates the generator to learn accurate semantic→visual mapping. On the other hand, synthesized unseen class features are more reliable, leading to learning appropriate classification boundaries of unseen classes.

4.4. Generative ZSL Models with VADS

To further evaluate VADS as a generic technology to improve generative ZSL, we integrate it into three prevalent generative ZSL frameworks: CLSWGAN [45], TF-VAEGAN [31], and FREE [7]. We use the official repository to reproduce the results and then insert our module to verify our method's effectiveness. Note that TF-VAEGAN and FREE contain a visual encoder, so we maintain their design. When inserting our modules, we keep the hyper-parameters unchanged to make the results comparable. The results on the three datasets are presented in Tab. 4. In terms of $\boldsymbol{Acc}$ and $\boldsymbol{H}$, we observe varying degrees of performance improvements (*e.g.*, a maximum of 11.5 points and a minimum of 1.2 points). Average growth is 7.4%/5.9%, 5.6%/6.4% and 3.3%/4.2% for $\boldsymbol{Acc}/\boldsymbol{H}$ on AWA2, SUN and CUB. Overall, the consistent improvement over competitive benchmarks validates the effectiveness of our proposed method.

4.5. Generalization Analysis of Visual Prior

In our method, we learn a local bias and a fixed global prior vector representing domain visual prior knowledge to generalize to unseen classes. Therefore, we investigate the impact of different forms of prior knowledge on performance. The results are detailed in Tab. 5. "Random" refers to the prior knowledge sampled from Gaussian distribution, "VGSE" [50] indicates that we take the semantic knowledge extracted from the visual representation as the prior, and "Other domain" means using the global vector learned

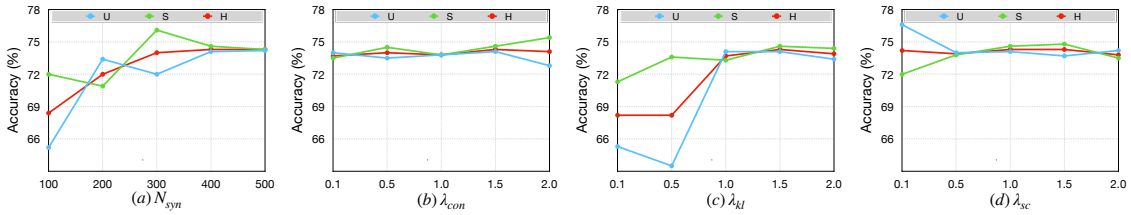


Figure 4. Effect of (a) synthesized samples N_{syn}, (b) loss weights λ_{con}, (c) loss weights λ_{kl}, and (d) loss weights λ_{sc} on CUB.

Table 5. We evaluate different forms of prior knowledge on CUB. The best result is marked in **boldface.**

Condition	Learnable	CUB			
		Acc	U	S	H
w/o prior		84.9	69.1	75.5	72.2
Random		85.6	72.7	74.7	73.4
VGSE [50]	✓	85.5	69.2	72.3	70.7
Other domain	✓	84.0	66.7	72.0	69.3
VADS (Ours)	✓	**86.8**	74.1	74.6	**74.3**

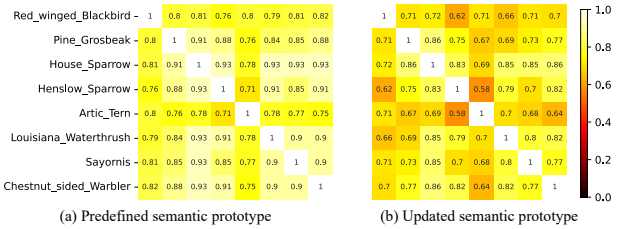


Figure 5. Visualization of the heatmap of semantic prototype similarity. We randomly select 10 classes on CUB.

from SUN to transfer to CUB. The results indicate that our method utilizes the global prior and the local bias, which yields the best performance. Knowledge transfer from SUN to CUB suffers from a negative impact, underscoring the importance of dataset-specific global information. VGSE captures semantic side information from seen classes, limiting the performance of unseen classes.

4.6. Hyper-parameters Analysis

We study the impact of different hyper-parameters of our VADS on the CUB dataset. Fig. 4(a) shows the results of synthesizing different numbers per unseen class. The unseen class accuracy varies with the number of synthesized samples, and when $N_{syn} = 400$, the performance is optimal. This result demonstrates that the features synthesized by our method alleviate the lack of unseen class data. Next, we evaluate the influence of individual loss weights (*i.e.*, λ_{con} of the $\mathcal{L}_{con}$, λ_{kl} of the $\mathcal{L}_{kl}$ and λ_{sc} of the $\mathcal{L}_{sc}$). The results are presented in Fig. 4(b)(c)(d). The performance of $\boldsymbol{S}$ and $\boldsymbol{U}$ changes slightly as $\mathcal{L}_{con}$ increases. When $\mathcal{L}_{con}$ is set to 1.5, $\boldsymbol{H}$ obtains the maximum value. For $\mathcal{L}_{kl}$ loss, we find that larger λ_{kl} achieves better performance because the prior distribution assumption of the data is crucial. Finally, $\mathcal{L}_{sc}$ loss forces the semantic update process to maintain semantic consistency. Its weight λ_{sc} is insensitive to performance.

4.7. Predefined Semantic Prototype vs Updated Semantic Prototype

To give a clearer insight into the predefined semantic prototypes and the semantic prototypes updated by our method, we randomly select 10 classes on CUB and calculate their cosine similarity. Then, we visualize them in Fig. 5. We observe that the similarity between the predefined semantic prototypes is very high. Our method dynamically refines the visual-semantic relationships of each instance based on visual information, making the updated semantic prototype easier to distinguish between categories and achieving more accurate semantic-visual alignment.

5. Conclusion

In this work, we propose a novel Visual-Augmented Dynamic Semantic prototype method (VADS) to boost the generator to synthesize reliable features of unseen classes. Considering that rich visual knowledge can effectively generalize to unseen classes, our proposed VADS fully leverages visual information. Specifically, we design a Visual-aware Domain Knowledge Learning module (VDKL) to acquire visual prior and a Vision-Oriented Semantic Updation module (VOSU) to dynamically update the predefined semantic prototype. Ultimately, we concatenate their output to form a dynamic semantic prototype, serving as the condition of the generator to learn accurate semantic-visual mapping and synthesize features of unseen classes. Extensive experiments demonstrate remarkable results in both CZSL and GZSL scenarios. In summary, our study provides a timely insight into reliable feature synthesis, improving the generalization to unseen classes. Additionally, tasks related to knowledge transfer can draw inspiration from this concept.

Acknowledgements

This work is partially supported by National Key R&D Program of China (2022YFC3301000).

References

[1] Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[2] Faisal Alamri and Anjan Dutta. Implicit and explicit attention for zero-shot learning. In *DAGM German Conference on Pattern Recognition*, pages 467–483. Springer, 2021. 6

[3] Faisal Alamri and Anjan Dutta. Multi-head self-attention via vision transformer for zero-shot learning. *arXiv preprint arXiv:2108.00045*, 2021. 6

[4] Jacopo Cavazza, Vittorio Murino, and Alessio Del Bue. No adversaries to zero-shot learning: Distilling an ensemble of gaussian feature generators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 6

[5] Samet Cetin. Closed-form sample probing for training generative models in zero-shot learning. Master's thesis, Middle East Technical University, 2022. 6

[6] Dubing Chen, Yuming Shen, Haofeng Zhang, and Philip HS Torr. Zero-shot logit adjustment. *arXiv preprint arXiv:2204.11822*, 2022. 1

[7] Shiming Chen, Wenjie Wang, Beihao Xia, Qinmu Peng, Xinge You, Feng Zheng, and Ling Shao. Free: Feature refinement for generalized zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 122–131, 2021. 1, 2, 3, 5, 6, 7

[8] Shiming Chen, GuoSen Xie, Yang Liu, Qinmu Peng, Baigui Sun, Hao Li, Xinge You, and Ling Shao. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. *Advances in Neural Information Processing Systems*, 34:16622–16634, 2021. 1, 6

[9] Shiming Chen, Ziming Hong, Wenjin Hou, Guo-Sen Xie, Yibing Song, Jian Zhao, Xinge You, Shuicheng Yan, and Ling Shao. Transzero++: Cross attribute-guided transformer for zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[10] Shiming Chen, Ziming Hong, Yang Liu, Guo-Sen Xie, Baigui Sun, Hao Li, Qinmu Peng, Ke Lu, and Xinge You. Transzero: Attribute-guided transformer for zero-shot learning. In *AAAI*, page 3, 2022. 3

[11] Shiming Chen, Ziming Hong, Guo-Sen Xie, Wenhan Yang, Qinmu Peng, Kai Wang, Jian Zhao, and Xinge You. Msdn: Mutually semantic distillation network for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7612–7621, 2022. 3

[12] Shiming Chen, Wenjin Hou, Ziming Hong, Xiaohan Ding, Yibing Song, Xinge You, Tongliang Liu, and Kun Zhang. Evolving semantic prototype improves generative zero-shot learning. *arXiv preprint arXiv:2306.06931*, 2023. 1, 3, 5, 6

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[14] Zhuo Chen, Yufeng Huang, Jiaoyan Chen, Yuxia Geng, Wen Zhang, Yin Fang, Jeff Z Pan, and Huajun Chen. Duet: Cross-modal semantic grounding for contrastive zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 405–413, 2023. 6

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5, 7

[16] Yaogong Feng, Xiaowen Huang, Pengbo Yang, Jian Yu, and Jitao Sang. Non-generative generalized zero-shot learning via task-correlated disentanglement and controllable samples synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9346–9355, 2022. 6

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[18] Jiayi Guo, Chaofei Wang, You Wu, Eric Zhang, Kai Wang, Xingqian Xu, Shiji Song, Humphrey Shi, and Gao Huang. Zero-shot generative model adaptation via image-specific prompt learning. In *CVPR*, pages 11494–11503, 2023. 2

[19] Akshita Gupta, Sanath Narayan, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Joost van de Weijer. Generative multi-label zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[20] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2371–2381, 2021. 1, 3, 6

[21] Ziming Hong, Shiming Chen, G Xie, Wenhan Yang, Jian Zhao, Yuanjie Shao, Qinmu Peng, and Xinge You. Semantic compression embedding for generative zero-shot learning. *IJCAI, Vienna, Austria*, 7:956–963, 2022. 1

[22] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4483–4493, 2020. 3, 6

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[24] Xia Kong, Zuodong Gao, Xiaofan Li, Ming Hong, Jun Liu, Chengjie Wang, Yuan Xie, and Yanyun Qu. 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*, pages 9306–9315, 2022. 1, 3, 6

[25] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):453–465, 2013. 3

[26] Man Liu, Feng Li, Chunjie Zhang, Yunchao Wei, Huihui Bai, and Yao Zhao. Progressive semantic-visual mutual adaption for generalized zero-shot learning. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15337–15346, 2023. 5, 6

[27] Yang Liu, Lei Zhou, Xiao Bai, Yifei Huang, Lin Gu, Jun Zhou, and Tatsuya Harada. Goal-oriented gaze estimation for zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3794–3803, 2021. 6

[28] Shaobo Min, Hantao Yao, Hongtao Xie, Chaoqun Wang, Zheng-Jun Zha, and Yongdong Zhang. Domain-aware visual bias eliminating for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12664–12673, 2020. 6

[29] Muhammad Ferjad Naeem, Yongqin Xian, Luc V Gool, and Federico Tombari. I2dformer: Learning image to document attention for zero-shot image classification. *Advances in Neural Information Processing Systems*, 35:12283–12294, 2022. 3

[30] Muhammad Ferjad Naeem, Muhammad Gul Zain Ali Khan, Yongqin Xian, Muhammad Zeshan Afzal, Didier Stricker, Luc Van Gool, and Federico Tombari. 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*, pages 15169–15179, 2023. 3

[31] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *European Conference on Computer Vision*, pages 479–495. Springer, 2020. 1, 2, 3, 5, 6, 7

[32] Sanath Narayan, Akshita Gupta, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Mubarak Shah. Discriminative region-based multi-label zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8731–8740, 2021. 3

[33] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. *Advances in neural information processing systems*, 22, 2009. 1

[34] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758. IEEE, 2012. 2, 5, 6

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[36] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849, 2023. 2

[37] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 49–58, 2016. 6

[38] Ivan Skorokhodov and Mohamed Elhoseiny. Class normalization for (continual)? generalized zero-shot learning. *arXiv preprint arXiv:2006.11328*, 2020. 6

[39] Jie Song, Chengchao Shen, Yezhou Yang, Yang Liu, and Mingli Song. Transductive unbiased embedding for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1024–1033, 2018. 3

[40] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 7

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[42] Chaoqun Wang, Shaobo Min, Xuejin Chen, Xiaoyan Sun, and Houqiang Li. Dual progressive prototype network for generalized zero-shot learning. *Advances in Neural Information Processing Systems*, 34:2936–2948, 2021. 3

[43] Zhengbo Wang, Jian Liang, Ran He, Nan Xu, Zilei Wang, and Tieniu Tan. Improving zero-shot generalization for clip with synthesized prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3032–3042, 2023. 1, 6

[44] P. Welinder, S. Branson, T. Mita, C. Wah, Florian Schroff, Serge J. Belongie, and P. Perona. Caltech-ucsd birds 200. *Technical Report CNS-TR-2010-001, Caltech,*, 2010. 2, 5, 6

[45] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 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, 2018. 6, 7

[46] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551, 2018. 1, 2, 3, 5, 7

[47] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10275–10284, 2019. 1, 2, 3, 5, 6

[48] Guo-Sen Xie, Li Liu, Xiaobo Jin, Fan Zhu, Zheng Zhang, Jie Qin, Yazhou Yao, and Ling Shao. Attentive region embedding network for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9384–9393, 2019. 3

[49] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. *Advances in Neural Information Processing Systems*, 33:21969–21980, 2020. 3

[50] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Vgse: Visually-grounded semantic embeddings for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9316–9325, 2022. 7, 8

[51] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023. 2, 3

[52] Meng Ye and Yuhong Guo. Progressive ensemble networks for zero-shot recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11728–11736, 2019. 6

[53] Sheng Zhang, Muzammal Naseer, Guangyi Chen, Zhiqiang Shen, Salman Khan, Kun Zhang, and Fahad Shahbaz Khan. S3a: Towards realistic zero-shot classification via self structural semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7278–7286, 2024. 1

[54] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 2, 3