Distinguishing Unseen from Seen for Generalized Zero-shot Learning

Hongzu Su[1] Jingjing Li[12*] Zhi Chen[3] Lei Zhu[4] Ke Lu[1]
[1]University of Electronic Science and Technology of China
[2] Institute of Electronic and Information Engineering of UESTC in Guangdong
[3]The University of Queensland [4]Shandong Normal University

Abstract

Generalized zero-shot learning (GZSL) aims to recognize samples whose categories may not have been seen at training. Recognizing unseen classes as seen ones or vice versa often leads to poor performance in GZSL. Therefore, distinguishing seen and unseen domains is naturally an effective yet challenging solution for GZSL. In this paper, we present a novel method which leverages both visual and semantic modalities to distinguish seen and unseen categories. Specifically, our method deploys two variational autoencoders to generate latent representations for visual and semantic modalities in a shared latent space, in which we align latent representations of both modalities by Wasserstein distance and reconstruct two modalities with the representations of each other. In order to learn a clearer boundary between seen and unseen classes, we propose a two-stage training strategy which takes advantage of seen and unseen semantic descriptions and searches a threshold to separate seen and unseen visual samples. At last, a seen expert and an unseen expert are used for final classification. Extensive experiments on five widely used benchmarks verify that the proposed method can significantly improve the results of GZSL. For instance, our method correctly recognizes more than 99% samples when separating domains and improves the final classification accuracy from 72.6% to 82.9% on AWA1.

1. Introduction

Conventional visual classification tasks deal with the same object categories in training and testing stage, i.e., samples in the training set and testing set have the same label space. Generally, methods for these tasks cannot correctly recognize samples which did not appear in training categories. Unfortunately, unseen categories are often involved in many real-world applications since the training dataset is finite. Zero-shot learning (ZSL) [10, 11, 14, 26, 32, 43] aims to handle unseen or novel instances by leveraging shared representations of visual and semantic modalities. In conventional ZSL [1, 25], a model recognizes samples only from unseen domain. Generalized zero-shot learning (GZSL) is a more challenging task which handles visual samples from both seen and unseen domains.

Early ZSL methods focus on embedding visual and semantic representations into a shared space [2, 3, 14, 17, 36, 38], e.g., mapping visual features into semantic space, or vice versa, and measuring similarity between two modalities. Recently, generating synthetic unseen visual features is widely adopted [8,27,28,33,37,41,43]. Generative methods firstly train a generative model such as GAN [18] or VAE [23] and synthesize a batch of features with unseen semantic attributes. Then a classifier is trained with seen samples and synthesized unseen samples to distinguish different classes. Since GZSL involves both seen and unseen categories, separating seen and unseen domains [5,7,12,31] is a reasonable solution. Once seen and unseen domains are separated, the GZSL problem is decomposed to a conventional zero-shot learning task and arbitrary seen and unseen experts can be adopted to accomplish classification.

Although separating seen and unseen domains is promising, it is quite challenging to distinguish seen and unseen visual features. For instance, as illustrated in Figure 1, $killerwhale$ and $humpbackwhale$ are categories can be accessed during the training phase, $dolphin$ is the unseen category for testing. These three species share a large number of common visual features and semantic attributes. Then generator is prone to generate features of $dolphin$ highly similar to $killerwhale$ and $humpbackwhale$ for lacking visual prior of category $dolphin$ in the training phase. Therefore, in the testing phase, a sample of $dolphin$ is prone to be wrongly recognized as $killerwhale$ or $humpbackwhale$. This phenomenon results in low classification accuracy on unseen classes. Since the supervision information of training is from seen categories, it is a disaster for unseen categories with similar samples in seen domains. Therefore, distinguishing seen and unseen samples is essential to promote GZSL performance.

In this paper, we focus on accurately separating seen and

*Jingjing Li is the corresponding author. Email: lijin117@yeah.net

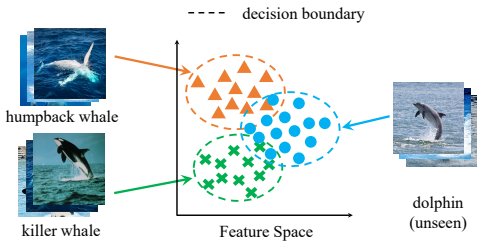


Figure 1. Illustration of similar species in AWA2 (best viewed in color). Samples of dolphin locate near in killer-whale cluster and humpback whale cluster.

unseen categories in cases when similar categories exist. To this end, we propose to generate a special class, called fictitious class, to separate similar visual features in a latent space. In our method, the latent representations of both visual and semantic modalities are embedded class-wisely on a latent space. Then we analyze the embedding boundary of each class and search a threshold to split seen and unseen samples. Specifically, we deploy hyperspherical VAE [13] models for both visual and semantic modalities and align the latent representations of the two modalities at the category-level. To leverage fictitious class, we propose a two-stage training scheme. Specifically, we firstly train both visual and semantic VAE models with seen samples and corresponding semantic attributes. Then we generate fictitious classes and train semantic VAE with fictitious samples and unseen semantic attributes. We measure similarity between the latent representations of two modalities and search a threshold to distinguish seen and unseen domains. By this, seen and unseen samples can be successfully distinguished. Further, we propose an unseen expert which is regularized by attention mechanism to classify unseen visual samples.

To summarize, the main contributions of this paper are threefold: (1) We propose a novel method to distinguish seen and unseen domains for GZSL. We design a two-stage training scheme which significantly improves the model performance by leveraging both seen and unseen semantic attributes. (2) We propose to leverage a novel fictitious class to separate similar visual representations. With fictitious class, we can successfully separate indistinguishable seen and unseen samples. In addition, we propose an unseen expert with attention mechanism to recognize unseen samples. It is worth noting that the unseen expert can be trained less than a minute in all tested datasets. (3) We conduct extensive experiments on five open benchmarks. The results verify that the proposed method can significantly improve the result of previous state-of-the-art approaches.

2. Related Work

Zero-shot learning. Conventional zero-shot learning (ZSL) [15, 26, 36, 43, 47] aims to classify categories which do not appear in the training set. In this paper, we focus on a more realistic yet challenging setting of ZSL named generalized zero-shot learning (GZSL). Different from conventional ZSL, the testing set of GZSL contains both seen and unseen categories.

In GZSL, a majority of previous studies focus on classification tasks with visual samples. The solutions of this task can mainly be divided into three categories, i.e., embedding methods [2, 3, 14, 30, 38, 48], generative methods [9,27,33,37,43,45] and domain-aware methods [7,31]. With respect to embedding methods, Ivan et al. [38] propose a method that maps semantic descriptions into visual space with class-wise normalization and significantly outperforms other methods. Alternatively, generative methods synthesize samples for unseen categories with corresponding semantic attributes, then inject these samples into training data. For instance, f-CLSWGAN [43] is a representative GAN-based method that deploys a Wasserstein GAN [4] to generate visual features. Domain-aware methods aim to explicitly distinguish seen and unseen domains. For instance, DVBE [31] learns to distinguish seen and unseen visual features in a semantic-free space and semantic-aligned space, respectively. DVBE embeds semantic descriptions into visual space to distinguish both seen and unseen categories with a single model. COSMO [5] designs a confidence-based gating to separate seen and unseen samples. This model consists of three parts, i.e., a seen classifier, an unseen classifier and a gating binary classifier. The gating classifier takes predictions of seen and unseen classifier as input and predicts gating scores for the two classifiers. Another representative method is proposed by Chen et al. [7], which embeds visual features and semantic descriptions in a latent space to split seen and unseen samples.

Our method falls into the domain-aware group. Specifically, we employ two VAE models to align visual features and semantic descriptions in a latent space and conduct a two-stage training strategy. Meanwhile, we learn a classifier in the latent space to distinguish seen and unseen representations. Notice that although our method shares the similar idea with OOD [7], the formulations of our method and OOD are significantly different. On one hand, our method conducts a two-stage training and generates synthetic features with unseen semantic attributes. However, OOD cannot leverage unseen knowledge for lacking unseen visual features. On the other hand, we propose to leverage a novel fictitious class to distinguish similar seen and unseen visual representations, which is shown capable of distinguishing seen and unseen samples accurately even they share quite similar visual and semantic characteristics. Furthermore, we propose an unseen expert with attention mechanism to classify unseen classes while OOD directly adopts an unseen classifier trained in f-CLSWGAN [43]. In the experiments, we will show that our method can significantly outperform OOD on all evaluated datasets, and much more de-

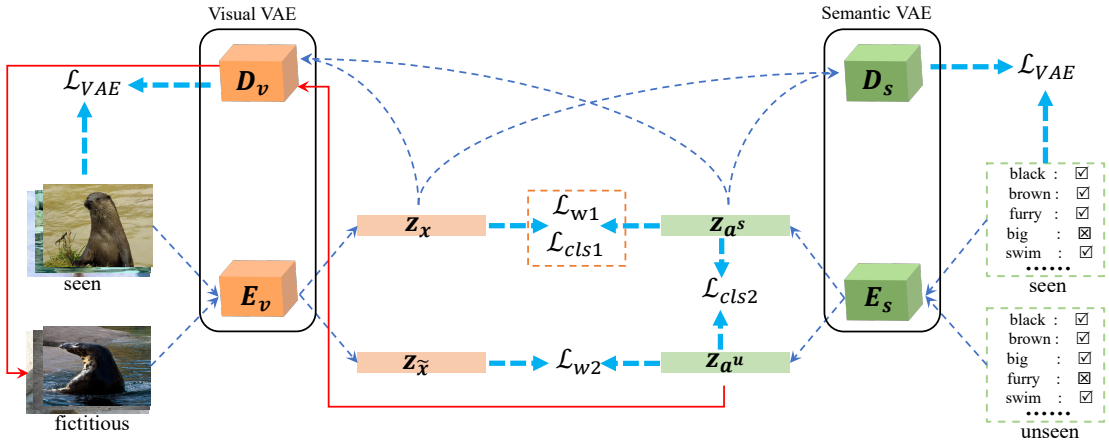


Figure 2. Illustration of our framework. E_v, D_v, E_s and D_s refer to visual encoder, visual decoder, semantic encoder and semantic decoder, respectively. Notations $z_x, z_{a^s}, z_{\tilde{x}}$ and z_{a^u} denote latent representations of seen visual samples, seen semantic descriptions, fictitious visual samples and unseen semantic descriptions, respectively. Notice that fictitious classes are generated with semantic encoder and visual decoder in our method. Red lines indicate the generation of fictitious classes.

tails will be illustrated in Section 4.

Attention mechanism has been widely adopted since it was proposed by Vaswani et al [40]. Methods in the literature [12,20–22,49] have verified that zero-shot learning can also benefit from attention mechanism. For instance, Huynh et al. [20] apply dense attention to capture visual features of specific regions and then embed them with corresponding semantic attributes. In another research of Huynh et al. [21], they propose a multi-attention method to matchup shared representations between visual and semantic modalities. Different from them, our unseen expert applies self-attention on the visual modality with an embedding model to enhance the principle representations of visual features.

3. Method

3.1. Problem Setting

In this paper, we focus on GZSL where both seen images and unseen images are used for evaluation. In GZSL, we are given a dataset $S = \{X^s, Y^s, A^s\}$ with N^s seen classes and another dataset $U = \{X^u, Y^u, A^u\}$ with N^u unseen classes. The labels for seen visual samples X^s and unseen visual samples X^u are denoted as Y^s and Y^u, respectively. A^s and A^u are the corresponding semantic descriptions of seen classes and unseen classes. Notice that the seen dataset S and the unseen dataset U are disjoint. Following the widely used setting in GZSL, the seen dataset is further split into a training set S^{tr} and a testing set S^{te}.

3.2. Overall Idea of Our Method

As illustrated in Figure 2, our method learns two variational auto-encoders, one for each modality for zero-shot learning. The training process can be split into two stages. At the first stage, we map visual features and semantic descriptions into a shared latent space, in which we align the latent representations of the two modalities and learn a classifier for seen classes. After this stage, the representations encoded by our model is expected to be modality-invariant for these seen classes. At the second stage, we explicitly leverage the semantic descriptions of unseen classes to synthesize artificial samples that form the *fictitious class* in the visual space, where we exploit the modality-invariant features for unseen classes as in the first stage. We deploy another classifier to separate fictitious classes and seen classes in the latent space. After the two-stage training, latent representations of both modalities are aligned class-wisely and a boundary of each class can be easily found. By analyzing these boundaries of seen categories, we compute a threshold that can separate seen and unseen samples. Once seen and unseen classes are separated, arbitrary seen or unseen experts can be adopted to carry out visual classification.

3.3. Aligning Seen Latent Representations

Inspired by some recently works in ZSL [7, 13, 37], we align both semantic and visual representations in a shared hyperspherical space on top of SVAE [13]. SVAE is first proposed by Davidson et al. [13], which replaces the hyperplane with a hypersphere as the latent space since hyperspherical representations can better explain data types such as directional data [13]. And von Mises-Fisher distribution is adopted to build the hyperspherical latent space. SVAE has verified the effectiveness of hyperspherical space with low dimensional data. Lately, Chen *et al*. [7] adopt this structure and achieve significant improvement in zero-shot learning. Given its effectiveness, we adopt the SAVE in our method to learn latent representations.

We deploy a specific SVAE for each modality and align latent representations of both modalities by minimizing the Wasserstein distance between them. The advantage of Wasserstein distance is that it can work well even if two dis-

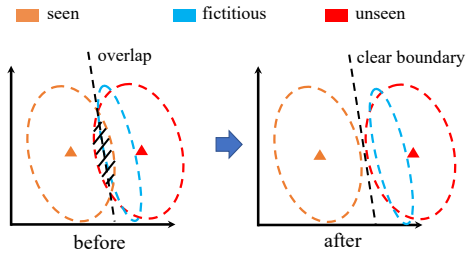


Figure 3. Illustration of fictitious class. Orange, red and blue circles indicate boundaries of seen class, unseen class and fictitious class, respectively. Red and orange triangles denote the class center of unseen class and seen class, respectively.

tributions are not overlapped [4]. Formally, the Wasserstein distance is defined as

$$\mathcal{L}_{W1} = \inf_{\gamma \in \Pi\left(P_{z_x}, P_{z_{a^s}}\right)} \mathbb{E}_{(z_x, z_{a^s}) \sim \gamma}[\|z_x - z_{a^s}\|], \quad (1)$$

where z_x and z_{a^s} denote the visual and the semantic latent representations of seen categories, respectively. P_{z_x} and $P_{z_{a^s}}$ are marginal distributions of z_x and z_{a^s}, respectively. $\Pi\left(P_{z_x}, P_{z_{a^s}}\right)$ denotes all possible joint distributions of z_x and z_{a^s}.

To further encourage the model to learn modality-invariant representations, we impose the following cycle-consistent reconstruction loss on latent representations,

$$\mathcal{L}_{cr} = \mathbb{E}\left[|a^s - D_s(z_x)| + |x - D_v(z_{a^s})|\right], \quad (2)$$

where z_x and z_{a^s} are latent representations of visual and semantic modalities, respectively. The two VAE models are trained with following loss:

$$\begin{aligned}\mathcal{L}_{\mathrm{VAE}} = & \mathbb{E}_{q_{\phi_1}(z_x|x)}\left[\log p_{\theta_1}\left(x \mid z_x\right)\right] - \lambda \mathrm{KL}\left(q_{\phi_1}\left(z_x \mid x\right) \| p\left(z_x\right)\right) \\ & + \mathbb{E}_{q_{\phi_2}(z_{a^s}|a^s)}\left[\log p_{\theta_2}\left(a^s \mid z_{a^s}\right)\right] - \lambda \mathrm{KL}\left(q_{\phi_2}\left(z_{a^s} \mid a^s\right) \| p\left(z_{a^s}\right)\right), \end{aligned} \quad (3)$$

where λ is the penalty coefficient of KL-divergence.

In the first stage, we train both visual and semantic VAE models with samples from seen categories. To distinguish seen categories and learn more discriminative latent representations, a classifier is introduced in our model. The loss of this classifier is defined as

$$\mathcal{L}_{cls1} = -\mathbb{E}[p_{z_x} \log q_{z_x}] - \mathbb{E}[p_{z_{a^s}} \log q_{z_{a^s}}], \quad (4)$$

where p_{z_x} and $p_{z_{a^s}}$ are label vectors of z_x and z_{a^s}, respectively. q_{z_x} and $q_{z_{a^s}}$ are the predictions of the classifier.

3.4. Generating Fictitious Classes

The visual VAE model trained with seen samples is prone to project unseen samples into the same region of seen samples in the latent space. Therefore, the model is easy to get confused by similar features from different categories. However, this long-standing headache for previous methods serves as the direct motivation of our work. In our method, we propose to explicitly generate fictitious classes in the latent space for these unseen classes with their semantic attributes. Then, we deploy a classifier to distinguish samples from both seen classes and fictitious classes. At last, the classifier is able to identify real seen and unseen samples in the test phase.

Specifically, we first consider a specific unseen class which has samples prone to be classified into seen ones. It is easy to be observed that the visual representations of this unseen class is located between the corresponding semantic representation and the visual representations of those seen classes. As illustrated in Figure 3, these samples can be referred to as the overlap between seen and unseen categories which contains both seen and unseen latent representations. Therefore, we can separate the seen class and the unseen class by minimizing the overlapped area between them.

Since the goal of our method is to recognize seen and unseen domains, we can separate seen and unseen visual representations by classifying them into corresponding categories. However, the visual representations of unseen classes are unavailable during training. In this paper, we propose to generate these representations explicitly with unseen semantic attributes, which can be expressed as follows,

$$z_{\tilde{x}} = E_v(\tilde{x}),\ \tilde{x} = D_v(E_s(a^u)), \quad (5)$$

where $a^u \in A^u$ denotes unseen attributes, E_v, D_v and E_s denote visual encoder, visual decoder and semantic encoder, respectively. A fictitious class is composed of latent representations $z_{\tilde{x}}$ corresponding to an unseen category.

To separate representations of seen classes and fictitious classes, we explicitly train a classifier on them. The process can be expressed as minimizing the following objective

$$-\mathbb{E}[p_{z_x} \log q_{z_x}] - \mathbb{E}[p_{z_{\tilde{x}}} \log q_{z_{\tilde{x}}}], \quad (6)$$

where z_x and $z_{\tilde{x}}$ denote the latent representations of seen classes and fictitious classes, respectively.

It is worth noting that the number of generated samples can be infinite. Thus, it is uncertain to decide on which samples the classifier should be trained. In this paper, we turn to exploit the invariant side of ZSL and replace visual representations with corresponding semantic representations to learn a robust classifier. For this purpose, we first explicitly align representations of generated fictitious classes and corresponding semantic attributes to enhance the consistency between them. Then, we directly replace visual representations in Eq.(6) with corresponding semantic representations. At last, we convert Eq.(6) to

$$\mathcal{L}_{cls2} = -\mathbb{E}[p_{z_{a^s}} \log q_{z_{a^s}}] - \mathbb{E}[p_{z_{a^u}} \log q_{z_{a^u}}], \quad (7)$$

where z_{a^s} and z_{a^u} denote the latent representations of seen and unseen semantic descriptions, respectively.

We conduct the aforementioned alignment by leveraging both semantic and visual VAEs in our work, yet we are not intended to train visual VAE in this stage for lacking real unseen visual features. Therefore, we only optimize semantic VAE model with

$$\begin{aligned}\mathcal{L}_s = \mathbb{E}_{q_{\phi_2}(z_{a^u}|a^u)}\left[\log p_{\theta_2}\left(a^u \mid z_{a^u}\right)\right] \\ -\lambda \mathrm{KL}\left(q_{\phi_2}\left(z_{a^u} \mid a^u\right) \| p\left(z_{a^u}\right)\right).\end{aligned} \tag{8}$$

At last, we formulate the optimizing objective for the alignment as

$$\mathcal{L}_{W2} = \inf_{\gamma \in \Pi\left(P_{z_{\tilde{x}}}, P_{z_{a^u}}\right)} \mathbb{E}_{(z_{\tilde{x}}, z_{a^u})\sim\gamma}[\|z_{\tilde{x}} - z_{a^u}\|]. \tag{9}$$

3.5. Overall Training Strategy

As illustrated before, our model is trained by two stages, one is based on seen visual samples and seen attributes, and the other is based on unseen attributes and fictitious classes. With all above formulations in Section 3.3, the overall loss at the first stage is

$$\mathcal{L}_1 = \mathcal{L}_{VAE} + \lambda_{rc}\mathcal{L}_{cr} + \lambda_{cls}\mathcal{L}_{cls1} + \lambda_w \mathcal{L}_{W1}, \tag{10}$$

where λ_{rc}, λ_{cls} and λ_w are used to balance loss penalty.

According to Section 3.4, the overall loss of the second stage is given by

$$\mathcal{L}_2 = \lambda_w \mathcal{L}_{W2} + \lambda_{cls}\mathcal{L}_{cls2} + \mathcal{L}_s. \tag{11}$$

In summary, the two-stage training forms a mutually reinforcing cycle in our method. Our model learns latent representations of two modalities and aligns them class-wisely in the first stage. Then, we separate similar representations with fictitious classes generated with unseen attributes. Therefore, latent representations of both modalities are aligned more compactly, which lead to a clearer boundary between seen and unseen domains.

3.6. Domain Distinguishment

Firstly, we calculate cosine similarities between all training samples and corresponding semantic descriptions in the latent space. We collect these cosine similarities into set C. Then, we search a cosine similarity γ which is lower than most of the values in C as follows

$$\eta = \frac{|\{\gamma \leq c \mid c \in C\}|}{|C|}, \tag{12}$$

where $|C|$ denotes the number of elements in set C. Threshold η denotes percentage of similarities greater than γ. With similarity γ, we identify the domain of a visual samples using

$$D = \left\{ \begin{array}{ll} seen, & \max\left\{cos\left(z_{te}, z_{a^s}\right) \mid \forall a \in \mathcal{A}_s\right\} \geq \gamma \\ unseen, & \max\left\{cos\left(z_{te}, z_{a^s}\right) \mid \forall a \in \mathcal{A}_s\right\} < \gamma \end{array} \right\}, \tag{13}$$

where z_{te} denotes latent representations of tested sample.

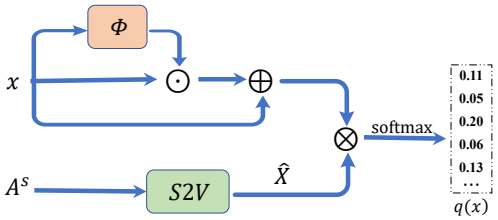


Figure 4. Illustration of unseen expert (best viewed in color). S2V denotes a Semantic-to-Visual module. ϕ denotes attention model. $\odot$, $\oplus$, $\otimes$ denote Hadamard product, add operation and matrix multiplication, respectively.

3.7. Zero-shot Classification

Once seen and unseen samples are separated, a GZSL problem is decomposed into a conventional ZSL problem and arbitrary seen and unseen expert can be deployed to recognize test samples. In our method, we directly adopt the classifier trained with VAEs to classify seen categories and propose an unseen expert to distinguish unseen samples.

The architecture of our unseen expert is illustrated in Figure 4. In this model, visual feature x is regularized with self-attention to learn principle representations as follows

$$\hat{x} = x + x \odot softmax(\Phi(x)), \tag{14}$$

where $\Phi(\cdot)$ denotes the attention module, $\odot$ denotes the Hadamard product. Our unseen expert embeds semantic descriptions into visual space. All seen attributes A^s are mapped into the visual space with a Semantic-to-Visual (S2V) model to learn visual representations $\hat{X}$ in the training phase. We measure the cosine similarity between $\hat{X}$ and $\hat{x}$ to classify visual features. The loss objective used to train this model is defined as

$$\mathcal{L} = -\sum_x l(x) \log\left(softmax(cos(\hat{X}, \hat{x}))\right), \tag{15}$$

where $l(x)$ is the one-hot labels of features x. Once the model is trained, we can directly replace A^s with unseen attribute A^u to recognize unseen samples.

4. Experiments

4.1. Dataset Descriptions

For fair comparisons, we use the same dataset as that in [5, 7]. We evaluate our model on five widely used datasets, including Caltech-UCSD Birds-200-2011 (CUB) [2], Oxford Flowers (FLO) [34], SUN Attribute (SUN) [35], Animals with Attributes 1 (AwA1) [24] and Animals with Attributes 2 (AwA2) [42]. The first three datasets are fine-grained datasets and the others are conventional datasets. Specifically, CUB is consists of 200 bird categories and FLO contains 102 different flower species. SUN is composed of 717 scenes and up to 14K samples. AwA1 is a large-scale dataset that contains 30,475 images and 50 animal categories. AWA2 consists of 37,322 samples divided into 50 categories from public sources.

Table 1. Comparison with the state-of-the-arts. **"U"** and **"S"** denote the top-1 accuracy of unseen and seen classes, respectively. **"H"** denotes the harmonic mean accuracy. The best results are shown in **Bold**. "Ours + f-CLSWGAN" denotes the results of a ZSL classifier trained by f-CLSWGAN [43]. "Ours" denotes the results of the proposed unseen expert. The methods after LMILR are published recently.

Methods	AwA1			AwA2			CUB			SUN			FLO		
	U	S	H	U	S	H	U	S	H	U	S	H	U	S	H
ALE [2]	16.8	76.1	27.5	14	81.8	23.9	23.7	62.8	34.4	21.8	33.1	26.3	13.3	61.6	21.9
SJE [3]	11.3	74.6	19.6	8.0	73.9	14.4	23.5	59.2	33.6	14.7	30.5	19.8	13.9	47.6	21.5
DeViSE [17]	13.4	68.7	22.4	17.1	74.7	27.8	23.8	53.0	32.8	16.9	27.4	20.9	9.9	44.2	16.2
ESZSL [36]	6.6	75.6	12.1	5.9	77.8	11.0	12.6	63.8	21.0	11.0	27.9	15.8	11.4	56.8	19.0
Ivan*et al.* [38]	63.1	73.4	67.8	60.2	77.1	67.6	49.9	50.7	50.3	44.7	41.6	43.1	-	-	-
f-CLSWGAN [43]	57.9	61.4	59.6	52.1	68.9	59.4	43.7	57.7	49.7	42.6	36.6	39.4	59.0	73.8	65.6
cyc-CLSWGAN [16]	56.9	64.0	60.2	-	-	-	45.7	61.0	52.3	49.4	33.6	40.0	59.2	72.5	65.1
LisGAN [26]	52.6	76.3	62.3	-	-	-	46.5	57.9	51.6	42.9	37.8	40.2	57.7	83.8	68.3
CADA-VAE [37]	57.3	72.8	64.1	55.8	75.0	63.9	51.6	53.5	52.4	47.2	35.7	40.6	-	-	-
LMILR [28]	61.5	75.0	67.6	57.5	83.9	68.2	52.4	57.9	55.0	47.9	36.4	41.4	-	-	-
TF-VAEGAN [33]	-	-	-	59.8	75.1	66.6	52.8	64.7	58.1	45.6	40.7	43.0	62.5	84.1	71.7
APN [44]	-	-	-	62.2	69.5	65.6	65.7	74.9	70.0	49.4	39.2	43.7	-	-	-
OOD [7]	59.0	94.3	72.6	55.9	94.9	70.3	53.8	94.6	68.6	57.8	95.1	71.9	61.9	91.7	73.9
SDGZSL [9]	-	-	-	64.6	73.6	68.8	59.9	66.4	63.0	48.2	36.1	41.3	62.2	79.3	69.8
GCM-CF [46]	-	-	-	60.4	75.1	67.0	61.0	59.7	60.3	47.9	37.8	42.2	-	-	-
AGZSL [12]	-	-	-	65.1	78.9	71.3	41.4	49.7	45.2	29.9	40.2	34.3	-	-	-
GEM-ZSL [29]	-	-	-	64.8	77.5	70.6	64.8	77.1	70.4	38.1	35.7	36.9	-	-	-
CE-GZSL [19]	65.3	73.4	69.1	63.1	78.6	70.0	63.9	66.8	65.3	48.8	38.6	43.1	69.0	78.7	73.5
FREE [6]	62.9	69.4	66.0	60.4	75.4	67.1	55.7	59.9	57.7	47.4	37.2	41.7	67.4	84.5	75.0
Ours + f-CLSWGAN	66.7	98.9	79.7	63.8	98.8	77.5	54.6	99.3	**70.5**	57.9	99.3	73.1	62.1	99.3	**76.4**
Ours	71.3	98.9	**82.9**	67.3	98.8	**80.1**	54.5	99.3	70.3	60.6	99.3	**75.2**	59.4	99.3	74.3

Table 2. Results of distinguishing unseen from seen. **H, FPR, AUC** denote the harmonic mean accuracy, False-Positive-Rate and Area-Under-Curve, respectively.

Methods	AwA1			CUB			SUN		
	H	AUC	FPR	H	AUC	FPR	H	AUC	FPR
COSMO [5]	56.6	91.2	39.8	44.8	80.5	70.7	40.1	72.2	82.5
OOD [7]	70.1	95.0	12.5	67.7	99.4	2.5	71.0	99.5	1.6
Ours (0.95)	**81.5**	**99.9**	**0.4**	**69.3**	**99.9**	**0.0**	**74.0**	**99.9**	**0.1**
Ours (0.99)	**82.9**	**99.9**	**0.7**	**70.3**	**99.9**	**0.0**	**75.2**	**99.9**	**0.1**

Table 3. Comparison cosine similarities with OOD method. The higher, the better.

Methods	AwA1	AwA2	CUB	SUN	FLO
OOD (0.95) [7]	0.995	0.990	0.961	0.901	0.993
Ours (0.95)	**0.999**	**0.998**	**0.987**	**0.975**	**0.996**

4.2. Experimental Protocols

Evaluation Metrics. For the results of GZSL, we follow the widely-used metric per-class top-1 accuracy to evaluate our model. Specifically, it is adopted to evaluate our model on both seen classes and unseen classes, denoted as S and U in the results, respectively. The harmonic mean of S and U, denoted as $H = (2 \times U \times S)/(U + S)$, is adopted to evaluate the GZSL performance.

For the results of spliting seen and unseen domains, we measure the True-Positive-Rate (TPR), False-Positive-Rate (FPR) and Area-Under-Curve (AUC) following Atzmon et al. [5]. The TPR and FPR indicate the percentage of classifying seen samples into seen domains and classifying unseen samples into seen domain, respectively. The AUC is measured with the threshold searched by Eq. (12).

Implementation Details. Following Xian et al. [43], we extract the features of the visual data from a ResNet-101 backbone pre-trained on ImageNet. Our model is implemented by PyTorch and trained on NVIDIA RTX 2080Ti GPUs. Both of the visual and semantic encoders are a three-layer MLP with a hidden-layer of 512 units, by which visual features and semantic attributes are mapped into 64D latent representations. The visual and semantic decoders follow a similar structure, i.e., a three-layer MLP with 512 units hidden-layer which maps 64D latent representations into original features or attributes. The classifier follows a widely used Linear LogSoftmax structure and is fed with 64D latent representations. We adopt Adam optimizer [23] to update the parameters of networks from scratch. The hyper-parameters for Adam are set as $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We set $\lambda_{rc} = 1.0$, $\lambda_{cls} = 1.0$ and $\lambda_w = 0.1$. It is worth noting that the threshold in Eq. (12) is set to 0.99.

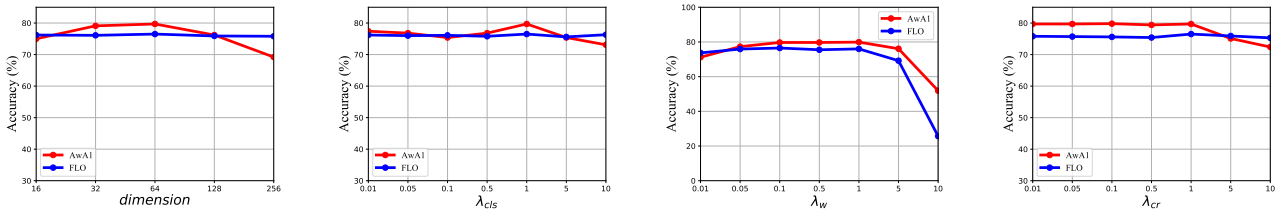
The attention module of our unseen expert is a fully-connected layer with a softmax. The S2V mapping is a four-layer MLP with two hidden layers of 1024 units. Besides, Class Normalization [38] is added to every hidden layer of the S2V module. We adopt Adam optimizer [23] to train the unseen expert with learning rate 0.005.

4.3. Results of GZSL

We compare our method with other approaches on five benchmark datasets and report the results in Table 1. From the results, we can observe that our method is able to significantly outperform previous state-of-the-art approaches. Compared with the second-best results in Table 1, our method achieves improvements of 10.3%, 8.8%, 3.3%,

Table 4. Results of ablation study. In the table, "CLS", "CR", "WD", "FC" denote classification loss, cycle-consistent reconstruction, Wassertein distance minimization and fictitious class, respectively. "w/o" is short for "without". "DUS-VAE" denotes our method. We do not report the accuracy results of "DUS-VAE w/o CLS" because the classifier is necessary for classification. The results are reported with corresponding threshold of 0.99.

Settings	AwA1			AwA2			CUB			SUN			FLO		
	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC
DUS-VAE w/o WD	99.43	37.82	90.29	99.11	44.95	86.85	98.98	5.69	98.86	98.87	5.20	99.42	99.60	2.94	99.39
DUS-VAE w/o FC	99.01	12.64	97.22	99.13	11.53	97.23	99.15	1.75	99.86	98.76	2.77	99.56	98.79	0.78	99.93
DUS-VAE w/o CLS	98.89	4.95	99.49	99.09	7.19	98.99	98.92	0.23	99.99	99.14	1.45	99.88	99.00	0.69	99.93
DUS-VAE w/o CR	98.98	2.33	99.58	99.24	4.27	99.10	98.92	0.06	99.99	98.91	0.27	99.94	98.50	0.51	99.97
DUS-VAE (full model)	99.10	0.70	99.90	98.91	2.92	99.30	99.30	0.03	99.99	99.30	0.13	99.98	99.30	0.34	99.97
	U	S	H	U	S	H	U	S	H	U	S	H	U	S	H
DUS-VAE w/o WD	37.66	99.43	54.63	33.18	99.11	49.71	50.34	98.98	66.74	54.09	98.21	69.76	60.47	99.60	75.25
DUS-VAE w/o FC	55.58	98.82	71.15	55.27	99.24	71.00	53.42	99.18	69.44	55.83	97.71	71.06	61.95	98.55	76.07
DUS-VAE w/o CR	64.72	98.98	78.26	61.11	99.24	75.64	54.63	98.91	70.39	57.84	98.91	73.00	62.05	98.50	76.14
DUS-VAE (full model)	66.75	98.95	79.72	63.81	98.87	77.56	54.67	99.3	70.52	57.91	99.22	73.10	62.16	99.30	76.46



(a) Effects of latent dimensions (b) Effects of λ_{cls} (c) Effects of λ_w (d) Effects of λ_{cr}

Figure 5. Parameter sensitivity analysis. We use AwA1 and FLO as examples. The results of harmonic mean accuracy are reported.

0.1% and 1.4% in terms of harmonic mean accuracy on AwA1, AwA2, SUN, CUB and FLO, respectively. With a ZSL classifier obtained in f-CLSWGAN [43], our method achieves performance improvements of 7.1% on AwA1, 6.2% on AwA2, 1.2% on SUN and 1.4% on FLO in terms of harmonic mean accuracy. An exciting observation is that our model is able to recognize almost all the seen samples. We achieve 98.9% on AwA1, 98.8% on AwA2, 99.3% on both CUB, SUN and FLO in terms of seen accuracy.

Comparing our results with the very recently published methods besides OOD, we observe that our method boosts the results of more than 30% in terms of harmonic mean accuracy on SUN dataset. We conjecture this phenomenon is caused by the bias problem, i.e., other methods are confused by similar samples between seen and unseen domains. For instance, the accuracy of GEM-ZSL [29] drops from 62.8% in ZSL to 38.1% in GZSL. We tackle this problem with the proposed fictitious class and the two-stage training strategy.

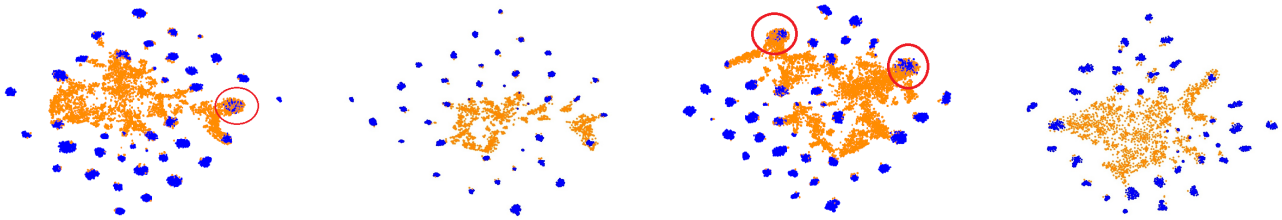
4.4. Results of Domain Classification

We report the results of recognizing seen and unseen domains in Table 2. For a fair comparison, we report our results when the threshold is set to 0.95 and 0.99, respectively. We compare our method with COSMO [5] and OOD [7] in terms of harmonic mean accuracy, False-Positive-Rate and Area-Under-Curve, respectively. From the reported results, we can observe that our method can significantly outperform other methods.

The results of term FPR are 0.7 on AwA1, 0.0 on CUB and 0.1 on SUN, respectively, which indicates that our model can hardly classify unseen samples into seen domains. The results of AUC are 99.9% on all datasets, which also verify that our method is able to distinguish seen and unseen samples perfectly. These results indicate that the proposed fictitious classes are able to establish clear boundaries between the latent representations of seen and unseen samples. Another effect of the proposed fictitious classes is that the embedding latent representations of both modalities more compact. As shown in Table 3, we measure the cosine similarities between latent visual and semantic representations. Comparing with OOD, we can observe that our model achieves higher cosine similarities on all tested datasets. For an intuitive comprehension, we compare the t-SNE [39] visualization results with OOD in Figure 6.

4.5. Model Analysis

Ablation Study. We report the results of our model without classification loss, cycle-consistent reconstruction, Wasserstein distance minimization and fictitious class, respectively, in Table 4. Since our model aims to separate seen and unseen samples, we report TPR, FPR, AUC along with the accuracy of seen, unseen, and harmonic mean. The threshold η is set to 0.99. From the results, we can observe that all the reported four parts contribute to the overall performance. Among them, our proposed fictitious class and Wasserstein distance minimization contribute the pri-



(a) OOD on AwA1 (b) Ours on AwA1 (c) OOD on AwA2 (d) Ours on AwA2

Figure 6. Visualization of the latent representations for our method and OOD [7] (best viewed in color). AwA1 and AwA2 are used as examples. Blue dots and orange dots denote seen representations and unseen representations, respectively. Red circles indicate similar seen and unseen representations locate in the same region. Notice that visualization results of OOD are cited from the original paper.

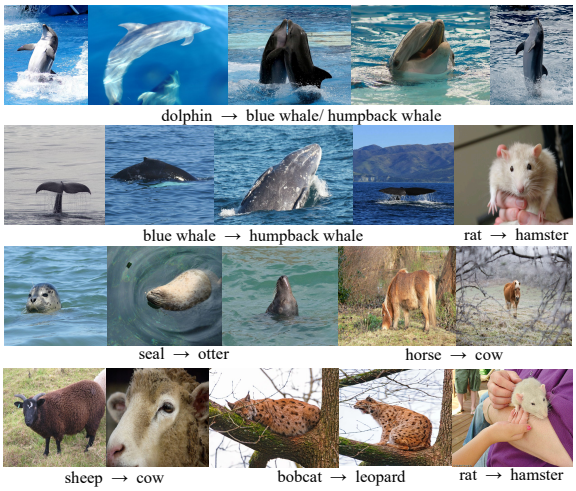


Figure 7. Qualitative results on AWA2. We show several samples which are correctly recognized by our method but failed by the OOD method. $A \rightarrow B$ denotes unseen category A is recognized as seen category B by the OOD method.

mary improvement. This result indicates that the fictitious class is able to distinguish similar latent representations. We can observe that fine-grained datasets, e.g., CUB, SUN and FLO are relatively not vulnerable to missing components comparing with conventional datasets AwA1 and AwA2.

Parameter Sensitivity. We conduct extensive experiments to investigate the effects of $\lambda_{cls}, \lambda_{rc}, \lambda_w$ and latent representation dimension. We report a series of analysis on AwA1 and FLO datasets to study the effects of parameters in Figure 5. From the results in Figure 5(a), we can observe that FLO dataset is not easily affected by latent dimension but AwA1 is more sensitive to latent dimension. The results shown in Figure 5(b) indicate the best effect of our model can be achieved when λ_{cls} is set to 1.0. From the results in Figure 5(c), we can observe that the harmonic mean accuracies slowly grow up as λ_w increases and show a rapid decline when λ_w is greater than 5. From the results in Figure 5(d), we can observe that the performance will decline when λ_{cr} is greater than 1. From all the four figures, we can observe that our model is relatively not sensitive to parameters on fine-grained dataset FLO but can be easily affected by parameters on conventional dataset AwA1.

Visualization. For an intuitive comprehension, we visualize the latent representations of our method and OOD [7] by t-SNE [39] and report the results in Figure 6. Comparing the results with OOD on AwA1 and AwA2 datasets, we can observe that the clusters of our model have tighter boundaries, which indicates that our model is able to better align latent representations. This is coincident with the results of cosine similarities in Table 3. As illustrated in Figure 6(a) and Figure 6(c), we circled clusters that contain both seen and unseen representations in OOD method. We can observe that there are no such clusters in Figure 6(b) and Figure 6(d). These observations all verify that our method performs better on distinguishing seen and unseen domains.

Qualitative Analysis. We report the results of qualitative analysis in Figure 7. We can observe that our model is able to tackle the feature fictitious problem. Although the OOD method can distinguish most of seen and unseen samples, it gets confused with images shown in Figure 7. We can successfully distinguish similar seen and unseen samples, such as humpback whale and dolphin, otter and seal. The results verified that our proposed fictitious class and two-stage training strategy are able to distinguish samples with many similarities.

5. Conclusions

In this paper, we propose fictitious class to separate seen and unseen samples for GZSL. We conduct a two-stage training scheme to leverage seen and unseen semantic attributes, respectively. Experiments on five open benchmarks verified that the proposed method could achieve new state-of-the-art performance. In our further work, we will study how to improve the performance on unseen samples by separating different unseen categories in GZSL.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62176042 and 62073059, in part by Sichuan Science and Technology Program under Grant 2020YFG0080, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2021B1515140013, in part by CCF-Baidu Open Fund (NO.2021PP15002000), and in part by CCF-Tencent Open Fund (NO. RAGR20210107).

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 819–826, 2013. 1

[2] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438, 2015. 1, 2, 5, 6

[3] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2927–2936, 2015. 1, 2, 6

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. 2, 4

[5] Yuval Atzmon and Gal Chechik. Adaptive confidence smoothing for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11671–11680, 2019. 1, 2, 5, 6, 7

[6] 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. 6

[7] Xingyu Chen, Xuguang Lan, Fuchun Sun, and Nanning Zheng. A boundary based out-of-distribution classifier for generalized zero-shot learning. In *European Conference on Computer Vision*, pages 572–588. Springer, 2020. 1, 2, 3, 5, 6, 7, 8

[8] Zhi Chen, Jingjing Li, Yadan Luo, Zi Huang, and Yang Yang. Canzsl: Cycle-consistent adversarial networks for zero-shot learning from natural language. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 874–883, 2020. 1

[9] Zhi Chen, Yadan Luo, Ruihong Qiu, Sen Wang, Zi Huang, Jingjing Li, and Zheng Zhang. Semantics disentangling for generalized zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.*, 2021. 2, 6

[10] Zhi Chen, Yadan Luo, Sen Wang, Ruihong Qiu, Jingjing Li, and Zi Huang. Mitigating generation shifts for generalized zero-shot learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2021. 1

[11] Zhi Chen, Sen Wang, Jingjing Li, and Zi Huang. Rethinking generative zero-shot learning: An ensemble learning perspective for recognising visual patches. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3413–3421, 2020. 1

[12] Yu-Ying Chou, Hsuan-Tien Lin, and Tyng-Luh Liu. Adaptive and generative zero-shot learning. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 3, 6

[13] Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical variational auto-encoders. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pages 856–865. Association For Uncertainty in Artificial Intelligence (AUAI), 2018. 2, 3

[14] Zhengming Ding, Ming Shao, and Yun Fu. Generative zero-shot learning via low-rank embedded semantic dictionary. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2861–2874, 2018. 1, 2

[15] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2014. 2

[16] Rafael Felix, Ian Reid, Gustavo Carneiro, et al. Multi-modal cycle-consistent generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37, 2018. 6

[17] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. 2013. 1, 6

[18] 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

[19] 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. 6

[20] 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

[21] Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8776–8786, 2020. 3

[22] Zhong Ji, Yanwei Fu, Jichang Guo, Yanwei Pang, Zhongfei Mark Zhang, et al. Stacked semantics-guided attention model for fine-grained zero-shot learning. In *Advances in Neural Information Processing Systems*, pages 5995–6004, 2018. 3

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 6

[24] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958. IEEE, 2009. 5

[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. 1

[26] Jingjing Li, Mengmeng Jing, Ke Lu, Zhengming Ding, Lei Zhu, and Zi Huang. Leveraging the invariant side of generative zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7402–7411, 2019. 1, 2, 6

[27] Jingjing Li, Mengmeng Jing, Ke Lu, Lei Zhu, Yang Yang, and Zi Huang. Alleviating feature confusion for generative

zero-shot learning. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1587–1595, 2019. 1, 2

[28] Jingjing Li, Mengmeng Jing, Lei Zhu, Zhengming Ding, Ke Lu, and Yang Yang. Learning modality-invariant latent representations for generalized zero-shot learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1348–1356, 2020. 1, 6

[29] 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, 7

[30] Yang Long, Li Liu, Ling Shao, Fumin Shen, Guiguang Ding, and Jungong Han. From zero-shot learning to conventional supervised classification: Unseen visual data synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1627–1636, 2017. 2

[31] 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. 1, 2

[32] Ashish Mishra, Shiva Krishna Reddy, Anurag Mittal, and Hema A Murthy. A generative model for zero shot learning using conditional variational autoencoders. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2188–2196, 2018. 1

[33] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. 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*, pages 479–495. Springer, 2020. 1, 2, 6

[34] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 5

[35] 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. 5

[36] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pages 2152–2161. PMLR, 2015. 1, 2, 6

[37] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8247–8255, 2019. 1, 2, 3, 6

[38] Ivan Skorokhodov and Mohamed Elhoseiny. Class normalization for (continual)? generalized zero-shot learning. In *International Conference on Learning Representations*, 2020. 1, 2, 6

[39] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7, 8

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[41] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4281–4289, 2018. 1

[42] 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. 5

[43] 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, 6, 7

[44] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *34th Conference on Neural Information Processing Systems*. Curran Associates, Inc., 2020. 6

[45] Yunlong Yu, Zhong Ji, Jungong Han, and Zhongfei Zhang. Episode-based prototype generating network for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14035–14044, 2020. 2

[46] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15404–15414, 2021. 6

[47] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2021–2030, 2017. 2

[48] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 4166–4174, 2015. 2

[49] Yizhe Zhu, Jianwen Xie, Zhiqiang Tang, Xi Peng, and Ahmed Elgammal. Semantic-guided multi-attention localization for zero-shot learning. *Advances in Neural Information Processing Systems*, 32, 2019. 3