SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

February 2025, Vol. 68, Iss. 2, 122102:1–122102:15
https://doi.org/10.1007/s11432-023-4174-4

Aligning enhanced feature representation for generalized zero-shot learning

Zhiyu FANG, Xiaobin ZHU*, Chun YANG, Hongyang ZHOU, Jingyan QIN & Xu-Cheng YIN

School of Computer & Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

Received 23 April 2023/Revised 14 December 2023/Accepted 28 June 2024/Published online 13 January 2025

Abstract Constructing an effective common latent embedding by aligning the latent spaces of cross-modal variational auto-encoders (VAEs) is a popular strategy for generalized zero-shot learning (GZSL). However, due to the lack of fine-grained instance-wise annotations, existing VAE methods can easily suffer from the posterior collapse problem. In this paper, we propose an innovative asymmetric VAE network by aligning enhanced feature representation (AEFR) for GZSL. Distinguished from general VAE structures, we designed two asymmetric encoders for visual and semantic observations and one decoder for visual reconstruction. Specifically, we propose a simple yet effective gated attention mechanism (GAM) in the visual encoder for enhancing the information interaction between observations and latent variables, alleviating the possible posterior collapse problem effectively. In addition, we propose a novel distributional decoupling-based contrastive learning (D^2-CL) to guide learning classification-relevant information while aligning the representations at the taxonomy level in the latent representation space. Extensive experiments on publicly available datasets demonstrate the state-of-the-art performance of our method. The source code is available at https://github.com/seeyourmind/AEFR.

Keywords generalized zero-shot learning, gated attention mechanism, contrastive learning, multi-modal alignment

Citation Fang Z Y, Zhu X B, Yang C, et al. Aligning enhanced feature representation for generalized zero-shot learning. Sci China Inf Sci, 2025, 68(2): 122102, https://doi.org/10.1007/s11432-023-4174-4

1 Introduction

Zero-shot learning (ZSL) [1–3] aims to recognize new classes without additional data labeling. To mimic the human ability to recognize new things only from specific descriptions, ZSL explores side-information (e.g., attributes [4], word vectors [5], semantic description [6]) to achieve the knowledge transfer from seen to unseen classes. The success of knowledge transfer mainly relies on establishing an effective feature space mapping between visual information and side information. To further generalize to real-world settings, generalized ZSL (GZSL) has received ever-increasing attention in research and industrial communities in which the test instances come from a mixture of seen and unseen classes.

In the paradigm of GZSL, generative models (e.g., generative adversarial networks (GANs) [7] or variational auto-encoder (VAEs) [8]) are popular and achieve promising performance. They often try to learn informative features on a common latent representation space [9–12]. Schönfeld et al. [13] employed two VAEs to learn the joint reconstruction between visual features and semantic features in the common latent space. Xu et al. [14] proposed an attribute prototype network to regress and decorrelate attributes from intermediate features extracted by visual information. Han et al. [11] produced the synthetic samples by a feature generator from attribute vectors into the common embedding space. However, since the lack of fine-grained instance-wise semantic information, the common space learning-based methods can easily suffer from the collapse problem, i.e., the mode collapse [15] prevalent in GAN-based methods and the posterior collapse [12] prevalent in VAE-based methods. A possible strategy is to enhance the discriminative ability of common space feature representations to obtain a robust GZSL model.

In the generative methods, aligning visual and semantic latent variables is a popular and effective strategy. Some studies adopt Wasserstein distance to align the distribution of latent variables [13,16,17]. Some other studies directly maximize the similarity between visual and semantic representations [18,

* Corresponding author (email: zhuxiaobin@ustb.edu.cn)

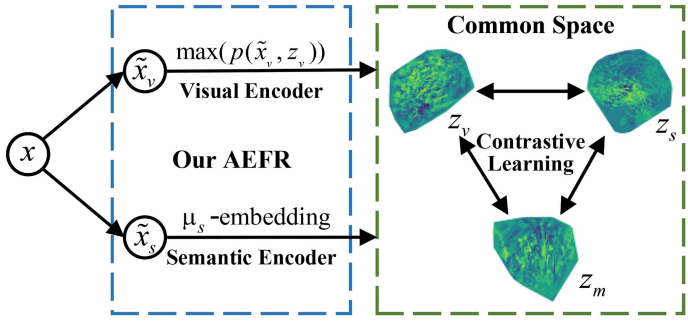

Figure 1 (Color online) Illustration of our motivation. The visual features $\widetilde{x}_v$ will be processed by the visual encoder to obtain enhanced representations via maximizing the joint distribution between $\widetilde{x}_v$ and z_v. The semantic features $\widetilde{x}_s$ will be processed by the semantic encoder to obtain mean embeddings of z_s. Furthermore, our AEFR leverages contrastive learning to optimize z in the latent common space to improve the discriminative ability of feature representations. Z_m is an augmented representation that assists in the multi-modal alignment.

19]. Since the alignment of the latent variables cannot mitigate the strong bias towards seen classes during training, some methods introduced classification constraints [12,20,21] and task distribution alignment [22,23] to improve the discriminability of latent feature representations. Contrastive learning [24] can learn representations with strong generalization ability, and it has demonstrated effectiveness on multi-modal alignment tasks [25]. Although some GZSL methods [11, 26] have adopted contrastive learning, they do not effectively impose constraints due to the multi-modality characteristic of VAEs.

In this paper, we propose an innovative asymmetric VAE network by aligning enhanced feature representations (AEFR) to learn discriminative feature representations for GZSL. Figure 1 illustrates the motivation of our method. Distinguished from existing VAE-based methods that process different modalities independently, we design an asymmetric variational autoencoder consisting of two independent encoders and a shared decoder. To be concrete, we propose a gated attention mechanism (GAM) in the visual encoder to strengthen the information interaction between observations and latent variables, aiming to effectively alleviate the posterior collapse problem. Based on the uniqueness of attributes on each category, we propose a semantic encoder with mean embedding, which cooperates with our proposed variance learning strategy to achieve effective modeling of semantic latent space. In addition, we propose a distributional decoupling-based contrastive learning (D^2-CL) strategy to optimize the alignment between visual and semantic modalities in the shared latent space. Specifically, by combining latent representations from two encoders and the distribution-based augmentation, our model aligns the latent variables at the taxonomy level in the representation space while learning classification-relevant information. Consequently, we train a softmax classifier with the discriminative feature representations encoded by AEFR, achieving superior performance on five benchmarks against the state-of-the-art (SOTA) methods.

In summary, our main contributions are three-fold:

- We propose an innovative asymmetric VAE network for GZSL. Experimental results verify the SOTA performance of our method on five publicly available datasets.
- We propose a simple yet effective GAM for enhancing the information interaction between observations and latent by maximizing their joint distribution, alleviating the possible posterior collapse problem in general VAE structures.
- Proposing a novel D^2-CL to align the representations of different distribution-wise modalities meanwhile supervised learning classification-relevant information.

2 Related work

Generalized zero-shot learning (GZSL), as the realistic extension of ZSL, identifies samples from both seen and unseen classes via model only trained on seen classes. The typical GZSL approaches learn a feature projection function to map the images into the corresponding class semantic prototypes, while classification is then implemented based on various metrics [27–29]. Due to the imbalance between visual and semantic information, those unilateral projection methods suffer from the domain shift problem. To alleviate this issue, Liu et al. [30] proposed a graph-based isometric propagation network to learn the sound relationship between classes with each space via mutual guidance of class dependency in two spaces. Yu et al. [31] employed the strong teacher network to provide the student network with rich

classification priors via knowledge distillation.

On the other hand, most existing methods [32–35] convert the GZSL task to a conventional classification task with unseen samples synthesized by generative models. Among them, common space learning-based methods [36] as the simple yield effective method have received ever-increasing attention in the GZSL community recently. f-VAEGAN-D2 [37] learns the marginal feature distribution of unlabeled images via a conditional generative model and an unconditional discriminator. EPGN [38] strengthens the alignment of latent features in their respective modalities by cross-reconstruction. Yu and Lee [20] proposed a deep generative model based on the VAE with class-specific multi-modal prior and the SGAL training algorithm. Schönfeld et al. [13] employed two independent VAEs to align visual and semantic representations in the shared latent space by one-step adaptation. Chen et al. [21] analyzed the shortcoming of one-step adaption and proposed the hierarchical two-step adaptation to consider the manifold structure relationship between different representations. Fang et al. [12] learned aligned cross-modal representations by an information enhancement module and the guidance of a learned classifier. However, the current methods ignore the collapse problem caused by the lack of fine-grained instance-wise annotations. Furthermore, existing alignment approaches of different latent representations lack multi-modal constraints, which are easily prone to classification shifts on unseen class samples.

Contrastive learning has been extensively investigated for object detection [39–41], image classification, and few-shot learning. It aims to build representation learning algorithms that only encode high-level features sufficient enough to distinguish different objects. LeCun et al. [42] pioneered contrastive loss to learn an invariant mapping for dimensionality reduction. Based on the core idea of balancing the relative relationship between positive and negative samples, contrastive learning is widely used in the computer science community [24, 43–45]. Gutmann et al. [46] proposed NCE loss to estimate the unnormalized statistical model by contrastive learning. Schroff et al. [47] proposed triplet loss to optimize a unified embedding for face recognition and clustering. In GZSL, Jiang et al. [26] utilized contrastive learning to exploit the class similarities, aiming to transfer knowledge from source images to similar target classes. Han et al. [11] proposed contrastive embedding to learn discriminative representations via class-wise and instance-wise supervision.

Therefore, we introduce contrastive learning and propose the AEFR model to solve the classification bias problem of existing GZSL methods in representation space. Different from existing methods, instead of using contrastive learning within a single modality, we treat the two modalities of visual and semantics as various feature augmentations of the same instance and achieve cross-modal alignment in a contrastive learning manner. Moreover, we design an asymmetric VAE structure to tackle the imbalance between visual and semantic feature diversity.

3 Proposed approach

3.1 Preliminaries

GZSL aims to learn a classifier $f_{\mathrm{GZSL}} : \mathcal{I} \to \mathcal{C}^s \cup \mathcal{C}^u$, where $\mathcal{I}$, $\mathcal{C}^s$ and $\mathcal{C}^u$ denote the images, the seen classes, and the unseen classes, respectively. And in particular, we assume that $X = \{X^s,\ X^u\}$ denotes the visual feature space of images $\mathcal{I}$, $A = \{A^s,\ A^u\}$ denotes the semantic feature space of categories $\mathcal{C}$, and $Y = \{Y^s,\ Y^u\}$ denotes the corresponding label set. Notably, Y^s and Y^u are disjoint, i.e., $Y^s \cap Y^u = \emptyset$. Consequently, we can individually collect seen domain dataset $D_s{=}\{x_i^s, a_i^s, y_i^s\}_{i=1}^N$ and unseen domain dataset $D_u{=}\{x_i^u, a_i^u, y_i^u\}_{i=1}^M$, where $x_i^s, x_i^u \in X$ is the i-th visual feature, $a_i^s, a_i^u \in A$ is the i-th semantic feature, and $y_i^s, y_i^u \in Y$ is their corresponding labels, of seen/unseen classes, respectively. According to the GZSL setting, we propose a novel auto-encoder trained only on the seen domain dataset D_s to learn discriminative feature representations. Then, we train a classifier $\mathcal{F}$ to recognize images from both seen and unseen categories.

3.2 Asymmetric variational auto-encoder

As illustrated in Figure 2, our AEFR consists of two different encoders (visual encoder E_v and semantic encoder E_s) and one shared decoder. Distinguished from the conventional VAE structures, we design a novel asymmetric structure to focus on the latent representation learning by AEFRs. Specifically, we propose the GAM for the visual encoder to strengthen the visual representation while alleviating the posterior collapse. Considering the uniqueness of attributes, we proposed the attribute encoder with

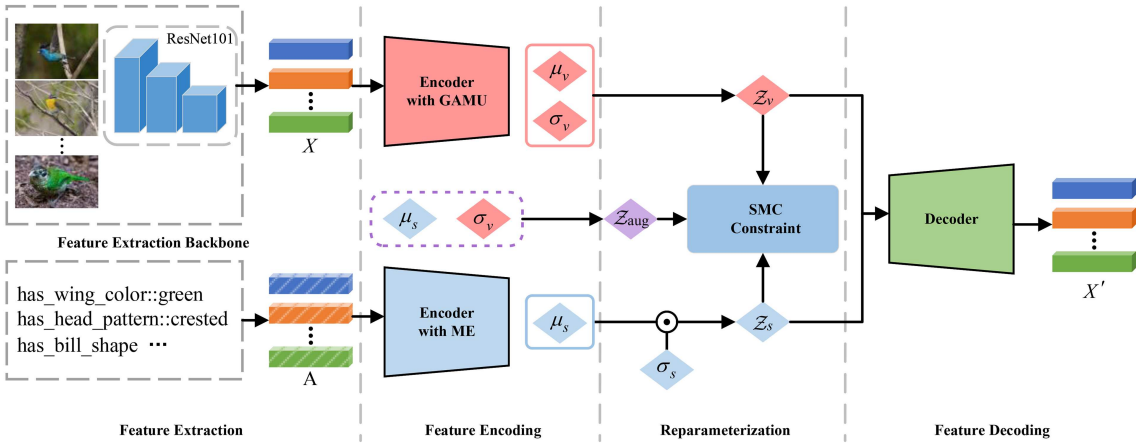

Figure 2 (Color online) Architecture of the proposed asymmetric VAE network (AEFR). GAMU denotes the GAM unit in the visual encoder, which is the core of our GAM. ME denotes the mean embedding in the semantic encoder. D^2-CL denotes the distributional decoupling-based contrastive learning on the representation space. Notations Z_v, Z_s, and Z_{aug} respectively denote latent representations of visual observations, semantic descriptions, and augmented features.

mean embedding and a learnable strategy for variance embedding. And then, we employ a simple single-hidden-layer multi-layer perceptron (MLP) as the decoder to restructure the original visual observations from different latent representations.

GAM. Visual features that come from real images contain abundant information. Hence, they rely on a delicate encoder to extract effective information. Additionally, the inherent posterior collapse problem [12] in VAE will reduce the discriminability of latent representations. Hence, we propose a novel visual encoder with the GAM to learn the enhanced feature representations. Specifically, our GAM directly strengthens the information interaction between the latent variable and the observation via a novel gated attention mechanism unit (GAMU). The detailed structure is shown in Figure 3. Inspired by FLASH [48], our GAM simulates Transformer [49] via integrating gated linear unit (GLU) [50] and self-attention [51] to construct a robust encoder. Our GAM can be mathematically formulated as

$$G(X) = \mathcal{G}\left(\varphi_u(X; W_u) \odot \mathcal{A} \otimes \varphi_v(X; W_v); W_g\right), \tag{1}$$

where φ_u and φ_v denote two learned non-linear functions to obtain the hidden representations from observation X, $\mathcal{A}$ denotes the attention weight computed by the LinearAtten, and $\odot$ and $\otimes$ denote the Hadamard product and the matrix product. We compute the attention weight matrix with fixed keys, which can be formulated as

$$\mathcal{A} = \text{Norm}(XW_a + b_a), \tag{2}$$

where $\mathcal{A}$ denotes the learned attention weight, Norm$(\cdot)$ denotes the normalization function, and W_a and b_a denote the weight and bias of the linear function.

To alleviate the posterior collapse problem, our GAM enhances the information interaction to approximate the joint distribution between observation and latent variable. Then, we can re-formulate Kullback-Leibler (KL) divergence as follows:

$$D_{\text{KL}}(q(z_x, x) \parallel p(z_x)) = q(x) D_{\text{KL}}(q(z_x|x) \parallel p(z_x)) + q(x) \log q(x), \tag{3}$$

where $q(z_x, x)$ is the posterior from GAM, $p(z_x)$ is the prior with standard normal distribution $\mathcal{N}(0, 1)$. Since the second term forces optimization in the direction of entropy increment, it can avoid the posterior collapse problem of the latent variables to standard normal distribution. We have provided detailed explanations and derivations in Appendix A.

Attribute encoder with distribution embeddings. Existing GZSL methods generally adopt attribute vectors as semantic features, each dimension of which represents shared attributes of different categories. Obviously, attribute vectors are highly semantic and category-unique. Therefore, encoding semantic features does not need to be as complex as visual features. We design the semantic encoder with mean embedding to obtain latent representations of semantic modality, which only model the normal means of semantic latent representations by a simple MLP. It can be formulated as

$$E_s(A) = W_2^{\text{T}} \text{ReLu}(W_1^{\text{T}} A) + b, \tag{4}$$

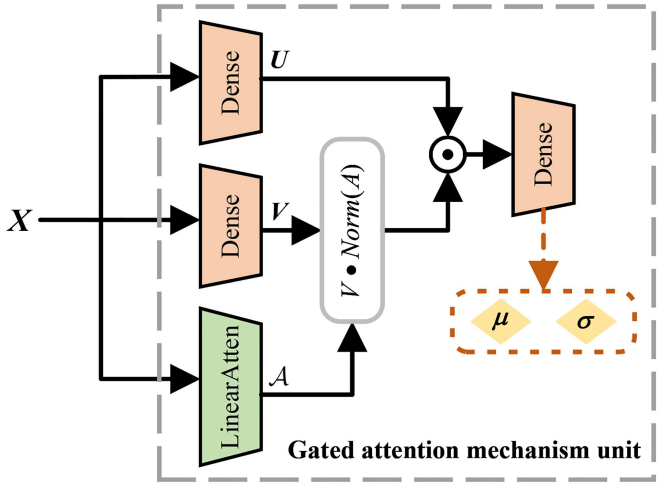

Figure 3 (Color online) Architecture of the GAMU. U, V, and A are φ_u, φ_v, and attention weight $\mathcal{A}$ in (1), respectively. μ and σ denote the mean and variance of latent representation, respectively.

where W_1 and W_2 denote the weights of MLP, ReLu$(\cdot)$ denotes the nonlinear activation function, and b denotes the bias. Furthermore, we define a learnable attribute embedding matrix $\mathcal{M}$ to describe each attribute dimension and obtain the attribute representation for each category by left-multiplying with an attribute probability matrix P_a. Then, we use dimensionality reduction techniques such as PCA to obtain the variance σ_a for the latent variable of attributes that is consistent with the dimensionality of visual latent variables. It can be formulated as

$$\sigma_a = U_a^{\mathrm{T}}(P_a \cdot \mathcal{M} - m_a), \tag{5}$$

where U_a is the projection matrix of PCA, and m_a is the mean vector of the attribute representations.

3.3 Distributional decoupling-based contrastive learning

While the Wasserstein distance is often utilized to constrain the distribution alignment between two modalities in many GZSL methods, the many-to-one relationship between visual features and semantic features makes it challenging to achieve uniform alignment of their distributions. Hence, we propose a novel D^2-CL strategy to align latent representations of different modalities and improve their discriminability. Specifically, we decouple the distributions by aligning the means to enhance the classification consistency of visual features and aligning the variances to enhance the diversity of semantic features. The details are shown in Figure 4.

Distribution-based feature augmentation (DFA). To implement decoupled contrastive learning, we first propose a DFA strategy. With this strategy, we construct an augmented latent variable Z_a that combines both visual and semantic feature distributions. Specifically, we can obtain the means $(\mu_v,\ \mu_s)$ and variances $(\sigma_v,\ \sigma_s)$ of visual and semantic features that follow Gaussian distributions after feature encoding with our two asymmetric encoders. Then, we sample Z_a from a new Gaussian distribution $\mathcal{N}(\mu_a, \sigma_x)$ based on the mean of semantic features and the variance of visual features. Our DFA strategy can be formulated as

$$Z_a = E_s(\mu|a) + E_v(\sigma|x) \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1), \tag{6}$$

where $E_s(\mu|a)$ denotes the mean of semantic representations from E_s, $E_v(\sigma|x)$ denotes the variance of visual representations from E_v, and ϵ denotes a random variable subject to normal distribution. Highlighting the fact that in actual datasets, there are multiple image samples but only one attribute vector for the same category. Our DFA strategy effectively expands the sample space of semantic representations, which provides an abundance of training samples for contrastive learning.

Contrastive learning based feature alignment. Similar to SimCLR [24], we leverage contrastive learning to align the different modal features in the shared latent space. Specifically, we respectively regard the visual, semantic, and augmented representations as different enhancement versions of one instance. Then, we introduce a simple MLP as a feature modulation layer to map representations into the unified

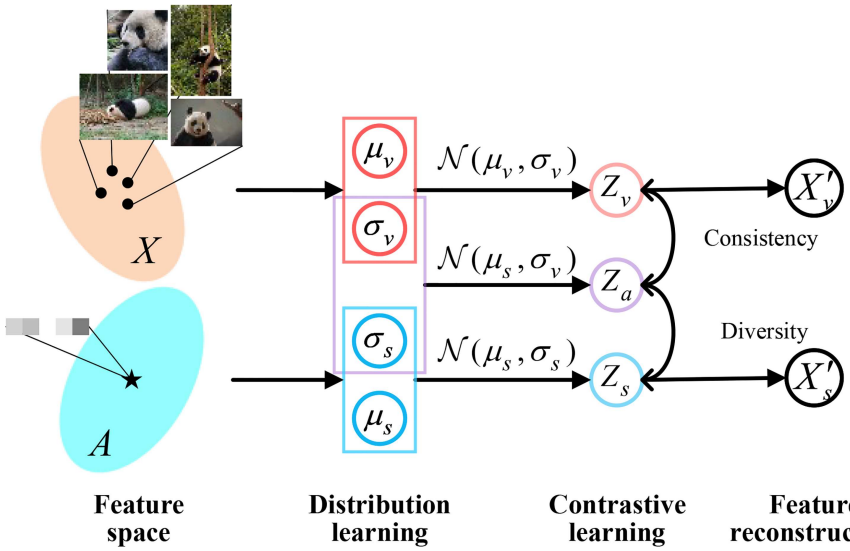

Figure 4 (Color online) Illustration of the D^2-CL. Z_v, Z_s, and Z_a denote visual latent variables that obey $\mathcal{N}(\mu_v, \sigma_v)$, semantic latent variables that obey $\mathcal{N}(\mu_s, \sigma_s)$, and augmented latent variables that obey $\mathcal{N}(\mu_s, \sigma_v)$, respectively.

space. Finally, we define a contrastive loss function with pairs of multi-modal latent representations, which can be formulated as

$$\mathcal{L}_{\text{D2CL}} = -\log \sum_{i=1}^{N} \frac{1}{N_{y_i}} \sum_{j=1}^{N} \mathbf{1}_{y_i=y_j} \left(\frac{\exp(S_{ij}^{(v,a)}/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k\neq i]} \exp(S_{ik}/\tau)} + \frac{\exp(S_{ij}^{(s,a)}/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k\neq i]} \exp(S_{ik}/\tau)} \right), \tag{7}$$

where S_{ik} denotes the similarity of random pair of examples (i, k) composed of arbitrary representations Z. Furthermore, the similarity $S_{ik} = z_i^{\mathrm{T}} \cdot z_k / \|z_i\| \|z_k\|$ is defined by a dot product between L_2 normalized z_i and z_k. Especially, $S_{ij}^{(v,a)}$ denotes the similarity of positive pair of examples (i, j) composed of Z_v and Z_a. $S_{ij}^{(s,a)}$ denotes the similarity of positive pair of examples (i, j) composed of Z_s and Z_a. τ denotes a temperature parameter. To introduce the inductive bias about classification, we divide positive and negative samples based on the labels. i.e., $\mathbf{1}_{y_i=y_j}$ as an indicator function evaluates to 1 iff $y_i = y_j$.

3.4 Generalized zero-shot classifier

Due to training data for unseen classes being unavailable, we follow the mainstream generative model-based paradigm to establish the GZSL classifier trained on the synthesized dataset. To be concrete, we sample the data randomly from the attributes of unseen classes and encode them via AEFR to obtain feature representations. To further improve the robustness of the classifier, we leverage the mix-up strategy [52] to construct new training samples from attributes, which can be formulated as

$$\begin{aligned} \widetilde{Z} &= \gamma Z_s^i + (1-\gamma) Z_s^j, \\ \widetilde{Y} &= \lfloor \gamma Y_i + (1-\gamma) Y_j \rfloor, \end{aligned} \tag{8}$$

where Z_s^i and Z_s^j are semantic features of different classes, Y_i and Y_j are corresponding labels, $\gamma \in [0, 1]$ is sampled from the Beta distribution.

Then we concentrate on unseen semantic representations, mix-up-enhanced representations, and seen visual representations to constitute a training dataset for the GZSL classifier. Finally, a single-layer MLP is trained using the synthetic dataset for GSL classification. The classifier can be formulated as

$$F_c = ZW + b, \{Z \in E_v(x^s) \cup E_s(a^u) \cup \widetilde{Z}\}, \tag{9}$$

where Z is the synthetic dataset, which consists of the visual representations encoded by E_v with images of seen classes x^s, and the semantic representations encoded by E_s with attributes of unseen classes a^u.

3.5 Optimization

Our method consists of two optimization tasks, i.e., optimizing asymmetric contrastive encoder and optimizing GZSL classifier.

Optimizing asymmetric VAE network. The first part of the optimization is the asymmetric VAE losses based on joint probability modification

$$\mathcal{L}_{\text{aVAE}} = \mathbb{E}_{q(z_x|x)}[\log p(x|z_x)] + \mathbb{E}_{q(z_s|s)}[\log p(x|z_s)] - \alpha D_{\text{KL}}(q(z_x, x) \parallel p(z_x)), \tag{10}$$

where $q(z_x|x)$ and $q(z_s|s)$ denote latent feature distributions from E_v and E_s, α is the weight of the KL-divergence between $q(z_x, x)$ and $p(z_x)$. Different from existing VAE-based GZSL methods, our AEFR decodes the latent variables encoded by E_v and E_s via the shared visual decoder.

The second part of optimization is the alignment of latent representations. Assuming $Z_s \sim \mathcal{N}(\mu_a, 1)$, the distribution-based alignment loss will exacerbate the posterior collapse problem on the visual modality. Hence, we propose a prototype-based alignment loss to tackle this issue meanwhile further strengthening the classification ability of aligned latent representations, which can be formulated as

$$\mathcal{L}_{\text{PA}} = \sqrt{\sum_{i=1}^{d}(\mu_x^i - \mu_a^i)^2}, \tag{11}$$

where μ_x denotes the mean of visual latent representations encoded by E_v, and μ_s denotes the mean of semantic latent representations encoded by E_s. Our final loss function can be formulated as

$$\mathcal{L}_{\text{AEFR}} = \mathcal{L}_{\text{aVAE}} + \beta\mathcal{L}_{\text{PA}} + \lambda\mathcal{L}_{\text{D2CL}}, \tag{12}$$

where β and λ is the weighting coefficient. Our method is trained by the Warm-up schedule [13].

Optimizing the GZSL classifier. We employ the synthetic dataset to train the classifier, and minimize the softmax cross-entropy to classify the consistent representations of all categories encoded by AEFR. The details training process is presented in Appendix B. The classification loss is formulated as

$$\mathcal{L}_C = -\mathbb{E}\left[\sum_{i=1}^{C}\mathbf{1}_{[i=y]}\log\frac{\exp F_c(z_i)}{\sum_{j=1}^{C}\exp F_c(z_j)}\right], \tag{13}$$

where C is total number of categories, $\mathbf{1}_{[i=y]}$ is an indicator function evaluated to 1 iff the i-th sample belongs to class y. $F_c(z)$ is the prediction score on the input sample z.

4 Experiments

4.1 Experimental settings

Datasets. We conduct experiments on five benchmark datasets: CUB [53], SUN [54], AwA [4], APY [55], and FLO [56]. Specifically, CUB is a fine-grained categorized dataset collected from professional bird websites, which covers 200 categories, and each category has a 312-dimensional attribute vector. SUN is a fine-grained scene understanding dataset that covers 707 categories, and each category is annotated with a 102-dimensional attribute vector. AwA is an animal image dataset that uses an 85-dimensional attribute to describe 50 categories, which collects 37322 images from public web sources. APY is composed of 20 object categories in PASCAL VOC 2008 and 12 object categories in Yahoo image search, and each category has a 64-dimensional semantic representation. Note that the AwA, SUN, and APY datasets are all based on human-annotated attributes, consisting of common features across all categories. FLO [56] is a fine-grained classification dataset of 102 British flowers created by the University of Oxford. Notably, we adopt the standard seen/unseen split introduced by [4]. The semantic embeddings of FLO and CUB are 1024-dimensional character-based CNN-RNN features [57] extracted from the fine-grained visual descriptions.

Evaluation protocols. The evaluation metrics in the generalized datasets setting include the average classification accuracy (ACA) of the samples from seen and unseen classes, denoted as S and U, respectively. The main performance indicator is the harmonic mean H, which can be formulated as $H = (2 \times U \times S)/(U + S)$.

Table 1 Comparisons of classification accuracy of the proposed model and the SOTA methods on five benchmarks under GZSL setting. U, S and H denote the average top-1 classification accuracy tested on unseen classes, seen classes, and their harmonic mean, respectively. The best results are in bold, and the second-best ones are underlined.

Model	CUB			SUN			AwA			APY			FLO		
	U	*S*	*H*	*U*	*S*	*H*	*U*	*S*	*H*	*U*	*S*	*H*	*U*	*S*	*H*
Feature generation															
cycle-CLSWGAN [32]	59.3	47.9	53.0	33.8	47.2	39.4	–	–	–	–	–	–	59.2	72.5	65.1
f-CLSWGAN [59]	57.7	43.7	49.7	36.6	42.6	39.4	53.8	68.2	60.2	–	–	–	59.0	73.8	65.6
SE [60]	53.3	41.5	46.7	30.5	40.9	34.9	68.1	58.3	62.8	–	–	–	–	–	–
CADA-VAE [13]	51.6	53.5	52.4	47.2	35.7	40.6	55.8	75	63.9	–	–	–	–	–	–
LiGAN [33]	46.5	57.9	51.6	42.9	37.8	40.2	54.3	68.5	60.6	–	–	–	57.7	83.8	68.3
DASCN [61]	59	45.9	51.6	38.5	42.4	40.3	–	–	–	39.7	59.5	47.6	60.5	80.4	69.0
E-PGN [38]	57.2	48.5	52.5	–	–	–	83.6	48	61	–	–	–	71.5	82.2	<u>76.5</u>
FREE [62]	53.1	57.7	55.3	47.4	37.2	<u>41.7</u>	60.4	75.4	67.1	–	–	–	67.4	84.5	75.0
MGA-GAN [35]	46.6	58.3	51.8	45.6	37.3	41.0	59.3	67.7	63.2	–	–	–	58.4	81.9	68.2
CE-GZSL [11]	63.9	66.8	**65.3**	48.8	38.6	**43.1**	63.1	78.6	**70.0**	–	–	–	69.0	78.7	73.5
Our AEFR w/o FT	65.8	63.2	<u>64.4</u>	46.4	37.7	41.6	60.0	82.0	<u>69.3</u>	35.4	65.8	46.0	82.5	88.2	**85.3**
Feature finetuning															
AREN [2]	63.2	69.0	66.0	40.3	32.3	35.9	54.7	79.1	64.7	30.0	47.9	36.9	–	–	–
f-VAEGAN-D2 [37]	63.2	75.6	68.9	50.1	37.8	43.1	57.1	76.1	65.2	–	–	–	63.3	92.4	75.1
TF-VAEGAN [63]	63.8	79.3	70.7	41.8	51.9	46.3	55.5	83.6	66.7	–	–	–	69.5	92.5	79.4
DVBE [18]	64.4	73.2	68.5	44.1	41.6	42.8	62.7	77.5	69.4	37.9	55.9	45.2	–	–	
AGZSL [52]	69.2	76.4	72.6	50.5	43.1	<u>46.5</u>	69.0	86.5	**76.8**	36.2	58.6	44.8	73.7	91.9	81.7
GNDAN [29]	69.2	69.9	69.4	50.0	34.7	41.0	60.2	80.8	69.0	–	–	–	–	–	–
SDGZSL [64]	73.0	77.5	<u>75.1</u>	–	–	–	69.6	78.2	<u>73.7</u>	39.1	60.7	<u>47.5</u>	86.1	89.1	<u>87.8</u>
MSDN [65]	68.7	67.5	68.1	52.2	34.2	41.3	62.0	74.5	67.7	–	–	–	–	–	–
TransZero [66]	69.3	68.3	68.8	52.6	33.4	40.8	61.3	82.3	70.2	–	–	–	–	–	–
ICCE [67]	67.3	65.5	66.4	–	–	–	65.3	82.3	72.8	45.2	46.3	45.7	66.1	86.5	74.9
Ours AEFR	74.6	80.4	**77.4**	51.4	43.8	**47.3**	66.8	80.8	73.1	39.2	63.8	**47.8**	84.4	94.0	**89.0**

Implementation details. We implemented our method with PyTorch and trained on NVIDIA TITAN X GPU. The visual encoder E_v is an MLP-based network containing a GAM unit and two distribution embeddings, which set the dimension of the output z to 64. The semantic encoder E_s and the decoder are single-hidden-layer MLP networks with ReLu activation. Following the setting in other methods [4, 13], we employ the 2048 dimensional visual features extracted by pre-trained ResNet-101 as the inputs of E_v and the attribute vectors as the inputs of E_s in our AEFR. The final softmax classifier includes a fully connected layer and a LogSoftmax activation layer. The classifier is fed with 64-dimensional latent representations and outputs the predictions on all categories. We adopt Adam optimizer [58] to train our AEFR and classifier with 100 epochs. We set the learning rate as 1.5e−04 for training AEFR and 0.5e−03 for training the classifier. For all datasets, the batch size of AEFR is set to 50 and the batch size of the final softmax classifier is set to 32.

4.2 Evaluation results with SOTA

We compare our AEFR against other SOTA methods on five benchmark datasets to demonstrate the effectiveness and advantages of our method. According to whether the visual features are fine-tuned, we divide the selected methods into feature generation-based (FG-based) methods and feature fine-tuning-based (FF-based) methods. The detailed experimental results are listed in Table 1 [2, 11, 13, 18, 29, 32, 33, 35, 37, 38, 52, 59–67]. Our AEFR achieves competitive performance on most benchmark datasets with or without feature finetuning. Specifically, compared with FG-based methods, our AEFR achieves the best performance of 85.3% on FLO, the second-best performance of 46.0% on APY, 69.3% on AwA, and 64.4% on CUB in terms of the harmonic mean (H), respectively. Compared with FF-based methods, our AEFR achieves the SOTA performance of 77.4% on CUB, 47.3% on SUN, 47.8% on APY, and 89.0% on FLO in terms of H, respectively.

Comparison of GZSL result with feature generation. FG-based methods leverage the generative model trained on the seen classes to synthesize the data for unseen classes, typically including the GAN-based methods and VAE-based methods. The GAN-based methods generally synthesize the images of

Table 2 Ablation study on the contributions of different modules. The best results of the harmonic mean are highlighted in bold. The datasets used in our experiment are fine-tuned datasets provided based on [63].

GAM	D^2-CL	CUB	SUN	AwA	APY	FLO
×	×	75.3	44.7	65.1	39.2	86.8
√	×	76.8	46.8	71.9	46.9	87.9
×	√	76.0	45.2	69.1	41.7	88.4
√	√	**77.4**	**47.3**	**73.1**	**47.8**	**89.0**

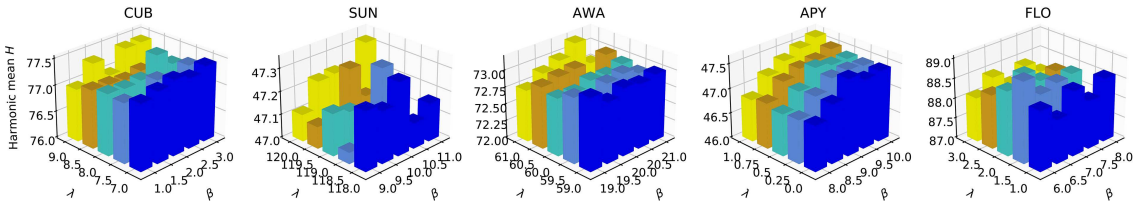

Figure 5 (Color online) Inference about weighting coefficients. We measure the harmonic mean H in GZSL on five datasets.

unseen classes. The VAE-based methods generate the latent embeddings in common space for unseen classes. Compared with GAN-based methods, our AEFR respectively outperforms E-PGN [38] by 11.9% on CUB, by 8.3% on AwA, and outperforms DASCN [61] by 1.3% on SUN in terms of H. Compared with VAE-based methods, our AEFR outperforms FREE [62] by 9.1% on CUB, by 2.2% on AwA, 10.3% on FLO, and underperforms FREE by 0.1% on SUN in terms of H, respectively. Compared with the top result, our method underperforms CE-GZSL [11] by 0.9% on CUB, by 1.5% on SUN, and by 0.7% on AwA in terms of H, respectively. Note that CE-GZSL concatenates visual and semantic vectors with a combination of VAE and GAN backbone. While our method may not have the complexity and forced priors of CE-GZSL, it still achieves comparable performance on CUB, SUN, and AwA. On the contrary, compared with the complex pipeline of CE-GZSL, our lightweight backbone can learn more effective feature representation from the relatively small fine-grained dataset of flowers (a.k.a FLO). Moreover, compared with the typical parallel VAE networks, our method outperforms CADA-VAE [13] by 1.0% on SUN, and by 5.4% on AwA in terms of H, respectively. This means that our proposed asymmetric VAE network is simple yet effective.

Comparison of GZSL result with feature finetuning. Since contrastive learning requires high feature richness, we employ fine-tuned visual features proposed by [63] to further verify the effectiveness of the model. Our method achieves the SOTA performance on four datasets, the value of H can reach 77.4% on CUB, 47.3% on SUN, 73.1% on AwA, 47.8% on APY, and 89.0% on FLO, respectively. Specifically, our method respectively outperforms AGZSL [52] by 4.8% on CUB, by 0.8% on SUN, by 3.0% on APY, and by 7.3% on FLO in terms of H. Besides, our method outperforms SDGZSL [64] by 2.3% on CUB, by 0.3% on APY, and by 1.2% on FLO in terms of H, respectively. Although our method fails to achieve the top result on AwA, we are 3.7% off the best result and only 0.6% off the second-best result.

4.3 Ablation study on AEFR

Contribution of different modules. We propose GAM and D^2-CL to improve the representation ability and discriminability of AEFR, respectively. To evaluate the contribution of different modules, we experiment with different combinations with fine-tuning settings. The comparison results are listed in Table 2. AEFR performs poorer than its full model when the GAM or D^2-CL is only applied for GZSL prediction. Specifically, the H drop by 0.6% and 1.4% on CUB, by 0.5% and 2.1% on SUN, by 1.2% and 4.0% on AwA, by 0.9% and 6.1% on APY, and by 1.1% and 0.6% on FLO, respectively. These results demonstrate that the proposed modules can effectively strengthen the discriminative ability of latent variables. Notably, the results in the first row of Table 2 demonstrate that our proposed framework of asymmetric VAE is effective. In brief, our proposed GAM and D^2-CL can effectively enhance the discriminative ability of features and improve GZSL performance.

Inference about weighting coefficients. We evaluate the influence of weighting coefficients in the loss function (12). We cross-validate weighting coefficients, and plot the harmonic mean (H) with respect to different β and λ, as shown in Figure 5. They reflect the sensitivity of our model to the proposed prototype-based alignment loss and distributional decoupling-based contrastive loss, respectively. With

Table 3 Ablation study on different feature combinations of synthetic data. Z and $\widetilde{Z}$ respectively denote the normal and augmented latent representation of attributes on unseen classes. The best results are in bold.

		CUB			SUN			AwA			APY			FLO		
Z	$\widetilde{Z}$	U	S	H	U	S	H	U	S	H	U	S	H	U	S	H
√	×	70.3	80.3	74.9	**48.4**	44.1	46.1	64.6	**83.1**	72.7	33.5	74.3	46.2	82.4	**94.4**	87.9
√	√	**74.6**	**80.4**	**77.4**	39.2	**63.8**	**47.8**	**66.8**	80.8	**73.1**	39.2	63.8	47.8	**84.4**	94.0	**89.0**

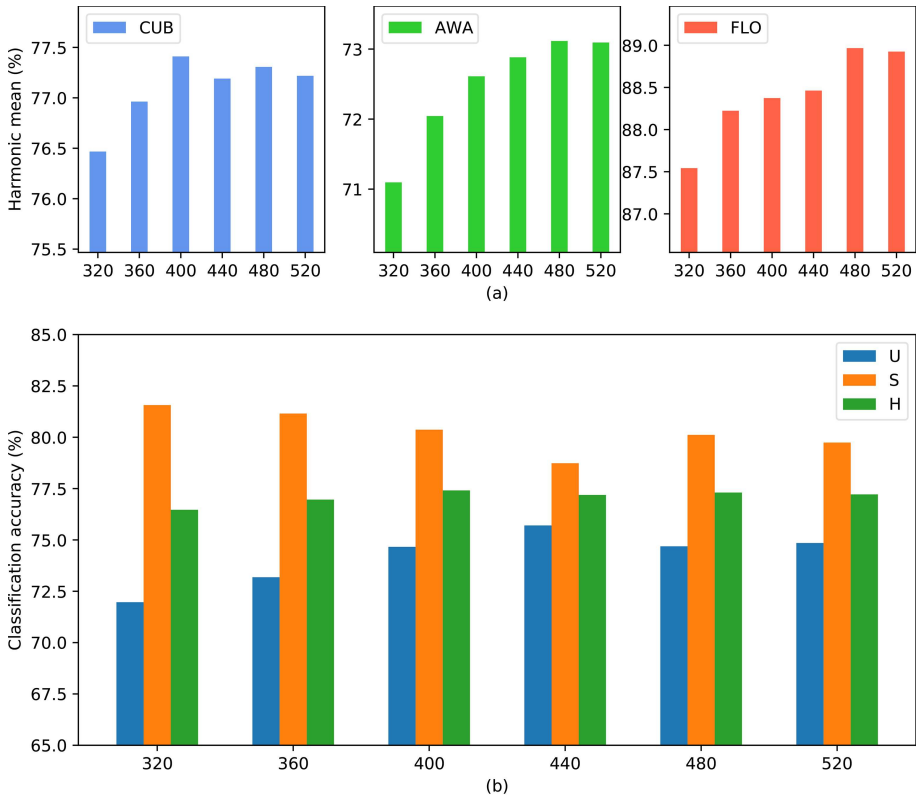

Figure 6 (Color online) Analysis of the impact of the number of synthetic data per unseen class on CUB, AwA, and FLO datasets. (a) The harmonic means on three datasets; (b) the detail results on CUB.

the different values, H results on different datasets change significantly, indicating that our proposed $\mathcal{L}_{\mathrm{PA}}$ and $\mathcal{L}_{\mathrm{D2CL}}$ are necessary and effective.

Evaluation on the components of synthetic data. To alleviate the problem of lacking a training dataset on unseen classes, we synthesize the dataset from the attribute vectors of unseen classes. There are two synthesis strategies in AEFR, one is the standard latent representation Z encoded by the semantic encoder, and another one $\widetilde{Z}$ is the Mix-up-based augmentation of Z. To investigate the impact of the components of synthetic data, we conduct the ablation study on different feature combinations. As listed in Table 3, the details indicate that Mix-up-based augmentation helps improve the generalization of the GZSL classifier.

Evaluation on the number of synthetic data. Generative methods transform ZSL into supervised learning by synthesizing pseudo-instances of unseen classes. To observe the influence of the number of synthesized features on the unseen dataset, we fix the synthetic number N_s of per seen class as 200, the number N_m of per mix-up-based features as 100, and only vary the synthetic number N_u of unseen class. From Figure 6(a), we can see that as the number of synthesized features of unseen classes increases, GZSL performance is improved consistently on all datasets until the respective optimal values are reached. From Figure 6(b), we can see that the accuracies on seen and unseen classes show opposite trends as the number increases. Therefore, we need a compromise strategy to set the synthetic number for achieving the optimal harmonic mean on all classes under the GZSL setting.

Table 4 Effectiveness comparison of our proposed modules on CADA-VAE.

Model	CUB			SUN			AwA1			AwA2		
	U	*S*	*H*	*U*	*S*	*H*	*U*	*S*	*H*	*U*	*S*	*H*
Baseline	50.0	53.9	51.9	43.8	37.5	40.4	58.4	71.1	64.4	54.2	78.4	64.1
+GAM	51.1	55.9	$53.4^{\uparrow 1.5}$	47.5	36.1	$41.1^{\uparrow 0.7}$	58.4	72.8	$64.8^{\uparrow 0.4}$	56.8	79.3	$66.2^{\uparrow 2.1}$
+D^2-CL	50.0	56.2	$53.0^{\uparrow 1.1}$	45.1	37.3	$40.8^{\uparrow 0.4}$	59.8	72.9	$65.7^{\uparrow 1.3}$	58.1	78.6	$66.8^{\uparrow 2.7}$

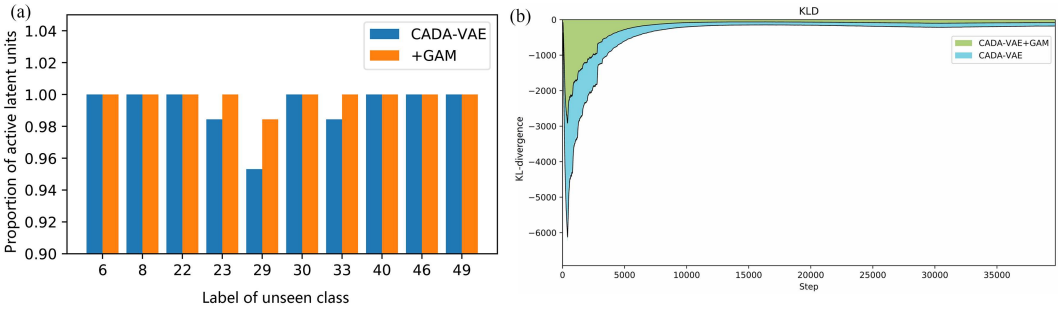

Figure 7 (Color online) Quantitative analysis of posterior collapse on AwA1. (a) AU on unseen classes; (b) the curve of KL-divergence during the training phase.

4.4 Effectiveness analysis on CADA-VAE

We tackle GZSL by proposing a novel cross-modal VAE and extending the classical CADA-VAE. To further demonstrate the effectiveness of our proposed method, we conducted experimental analysis on our GAM and D^2-CL based on the CADA-VAE framework, and the results are shown in Table 4. To ensure the fairness of the experiments, we reproduced the experiments in our own environment based on the open-source code and datasets provided by the authors of CADA-VAE, and tested our methods under the same conditions. Among them, "baseline" represents the original CADA-VAE model. "+GAM" represents adding the GAM module to the encoders of CADA-VAE. "+D2CL" represents adding the feature modulation layer to CADA-VAE network structure, along with our proposed $\mathcal{L}_{\text{D2CL}}$ added to the loss function. From the results in Table 4, it can be seen that the performance of CADA-VAE has consistently improved on all test datasets after adding our proposed modules. Furthermore, the performance of GAM is better than that of D^2-CL on fine-grained classification datasets CUB and SUN, while D2CL performs better than GAM on coarse-grained classification datasets AwA1 and AwA2. This quantitatively demonstrates the effectiveness of our two modules.

Posterior collapse. Since the posterior collapse problem of VAE, the latent variables cannot learn the effective features for classification. Hence, we propose the GAM to alleviate the collapse and help the encoder to generate improved representations. We use joint distribution to estimate mutual information. According to (3), the second term is not equal to zero, then KL divergence has a lower bound, thus ensuring the distribution of latent variables will not collapse to $\mathcal{N}(0,1)$. The detailed explanation is given in Appendix A. To directly illustrate the performance of our GAM, we plot the proportion of active units and the training curves on AwA1, as shown in Figure 7. Specifically, we follow the statistic $\text{AU} = \sum_{d=1}^{D} \text{Cov}_x(\mathbb{E}_{z\sim q(z|x)}[z_d]) > 0.01$ of [68] to measure the activity units of latent variables. It can be seen from Figure 7(a) that the activation unit of the hidden variable will be improved by adding GAM. And the curve in Figure 7(b) shows that GAM contributes to the optimization of KL divergence.

4.5 Visualization of latent representations

To intuitively verify the effectiveness of our proposed AEFR, we visualize the t-SNE distributions of ten unseen classes and the confusion matrix on AwA under the GZSL setting. The details are shown in Figure 8. We show the t-SNE distribution in Figure 8(a), where "Test" represents the AEFR-encoded features of the unseen classes from the testing dataset, "Synthesis" represents the synthetic data generated from the unseen class-based semantic annotation. Moreover, we draw the two features in the same coordinate system for easy observation. Since each class has only one fixed and unique attribute vector, the distribution of synthetic data is striped. However, it can be seen from the Mixture column that the distribution of the synthetic data is roughly consistent with the real testing set, especially "blue+whale".

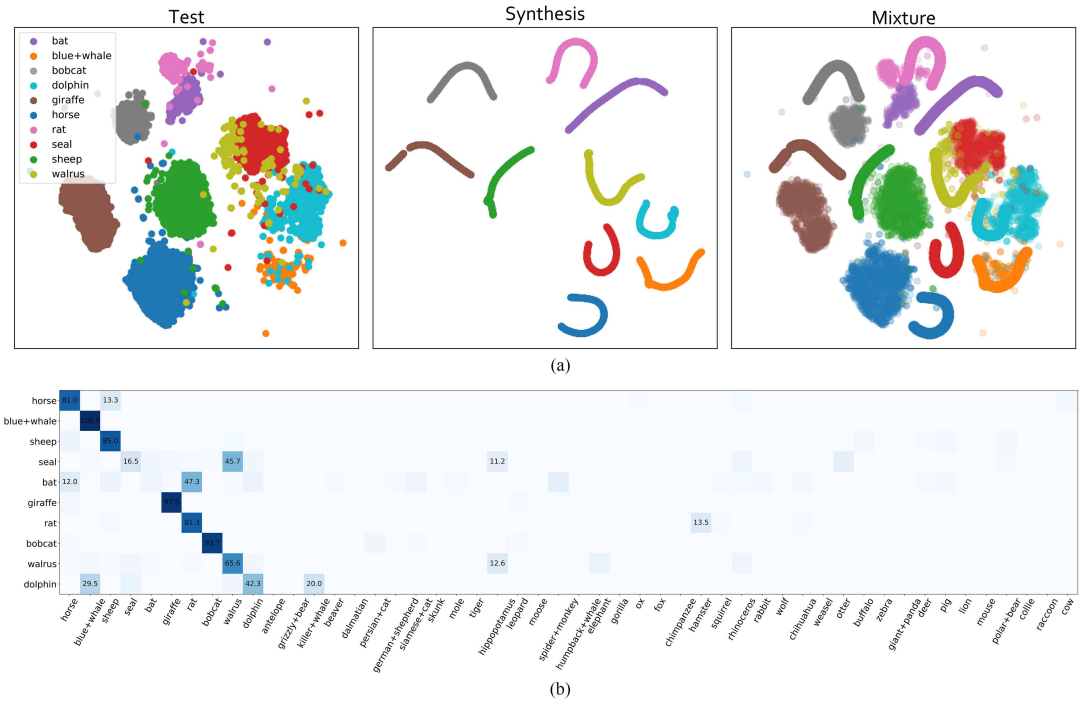

Figure 8 (Color online) Qualitative analysis of our proposed model. (a) t-SNE visualization of latent representations on AwA2; (b) confusion matrix of AEFR on AwA2 under GZSL setting.

Since attribute vectors cannot provide more detailed semantic information, there are also some categories whose distribution deviates from the facts. Furthermore, we leverage the confusion matrix to explore the actual predictions of our proposed AEFR on unseen classes. From Figure 8(b), we can conclude that the classification predictions match the t-SNE distribution of feature representations. Specifically, the classification accuracy of "blue+whale", "giraffe", "bobcat" achieve 100%, 97.5%, and 93.7%, respectively. Due to the visual similarity, (bat, rat) and (seal, walrus) are easily confused, and they have 47.3% and 45.7% misclassification probabilities in the confusion matrix, respectively.

4.6 Effectiveness analysis under few-shot setting

We conduct experiments on the CUB dataset with 1-shot, 5-shot, and 10-shot configurations to investigate the performance of our proposed AEFR in few-shot settings. The optimization of AEFR involves optimizing the asymmetric VAE network and the GZSL classifier. To explore the impact of few-shot settings on the model performance, we devised three training strategies: few-shot training on the asymmetric VAE, few-shot training on the classifier, and few-shot training on the entire model. Specifically, for few-shot training on the asymmetric VAE, we randomly select n-shot visual samples (where n equals 1, 5, or 10) from each unseen class and mix them with seen class training data to construct a new dataset for training our asymmetric VAE. For few-shot training on the classifier, we utilize the asymmetric VAE, pre-trained on seen classes, to generate few-shot training data for the unseen classes. For few-shot training on the entire model, we combine the aforementioned approaches to train our AEFR. The results from Table 5 indicate that the model's performance improves upon acquiring a small amount of data from unseen classes, and this enhancement is directly proportional to the volume of unseen class data obtained. Specifically, under the 1-shot setting, our model exhibited an increase in the H metric from 77.4 to 79.5, a 4.4% improvement was observed under the 5-shot setting, and a 5.7% improvement was observed under the 10-shot setting.

5 Conclusion

In this work, we proposed an innovative asymmetric VAE network by AEFR for GZSL to earn efficient discriminative feature representation. Our main contributions lie in innovative asymmetric VAE structure with GAM and D^2-CL. To improve the discriminability of representations, our AEFR leverages the

Table 5 Comparisons of classification accuracy of the different training strategies on CUB under a few-shot learning setting. The best H are in bold.

	Asymmetric VAE			GZSL Classifier			AEFR		
	U	S	H	U	S	H	U	S	H
1-shot	77.9	78.6	78.2	78.5	79.2	78.8	77.8	81.3	**79.5**
5-shot	80.9	80.8	80.9	81.8	79.0	80.4	83.2	80.5	**81.8**
10-shot	84.7	80.1	**82.3**	82.3	79.0	**80.6**	86.1	80.3	**83.1**

GAM to enhance the information interaction between observations and latent variables while mitigating the posterior collapse problem via maximizing the joint probability. Moreover, AEFR aligns enhanced representations distribution-wise, including the classification-relevant information guided by D^2-CL. Extensive experiments on five publicly available datasets demonstrate the SOTA performance of our method and verify the effectiveness of our proposed modules.

Acknowledgements This work was supported by National Science and Technology Major Project (Grant No. 2020AAA0109701), National Science Fund for Distinguished Young Scholars (Grant No. 62125601), and National Natural Science Foundation of China (Grant No. 62076024).

References

1 Li Y, Wang D, Hu H, et al. Zero-shot recognition using dual visual-semantic mapping paths. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2017. 5207–5215
2 Xie G S, Liu L, Jin X, et al. Attentive region embedding network for zero-shot learning. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2019. 9384–9393
3 Liu Y, Zhou L, Bai X, et al. Goal-oriented gaze estimation for zero-shot learning. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2021. 3794–3803
4 Xian Y, Lampert C H, Schiele B, et al. Zero-shot learning — a comprehensive evaluation of the good, the bad and the ugly. IEEE Trans Pattern Anal Mach Intell, 2018, 41: 2251–2265
5 Frome A, Corrado G S, Shlens J, et al. DeViSE: a deep visual-semantic embedding model. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013. 2121–2129
6 Miller G A. WordNet: a lexical database for English. Commun ACM, 1995, 38: 39–41
7 Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of International Conference on Neural Information Processing Systems, 2014. 2672–2680
8 Kingma D P, Welling M. Auto-encoding variational bayes. In: Proceedings of International Conference on Learning Representations, 2014. 1050–1051
9 Chen L, Zhang H, Xiao J, et al. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2018. 1043–1052
10 Jia Z, Zhang Z, Wang L, et al. Deep unbiased embedding transfer for zero-shot learning. IEEE Trans Image Process, 2020, 29: 1958–1971
11 Han Z, Fu Z, Chen S, et al. Contrastive embedding for generalized zero-shot learning. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2021. 2371–2381
12 Fang Z, Zhu X, Yang C, et al. Learning aligned cross-modal representation for generalized zero-shot classification. In: Proceedings of AAAI Conference on Artificial Intelligence, 2022. 6605–6613
13 Schönfeld E, Ebrahimi S, Sinha S, et al. Generalized zero- and few-shot learning via aligned variational autoencoders. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2019. 8247–8255
14 Xu W, Xian Y, Wang J, et al. Attribute prototype network for zero-shot learning. In: Proceedings of International Conference on Neural Information Processing Systems, 2020. 21969–21980
15 Lin Z, Khetan A, Fanti G, et al. PacGAN: the power of two samples in generative adversarial networks. In: Proceedings of the 32nd Conference on Neural Information Processing Systems, 2018. 1505–1514
16 Li X, Xu Z, Wei K, et al. Generalized zero-shot learning via disentangled representation. In: Proceedings of AAAI Conference on Artificial Intelligence, 2021. 1966–1974
17 Su H, Li J, Chen Z, et al. Distinguishing unseen from seen for generalized zero-shot learning. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2022. 7885–7894
18 Min S, Yao H, Xie H, et al. Domain-aware visual bias eliminating for generalized zero-shot learning. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2020. 12661–12670
19 Wang C, Min S, Chen X, et al. Dual progressive prototype network for generalized zero-shot learning. In: Proceedings of International Conference on Neural Information Processing Systems, 2021. 2936–2948
20 Yu H, Lee B. Zero-shot learning via simultaneous generating and learning. In: Proceedings of International Conference on Neural Information Processing Systems, 2019. 46–56
21 Chen S, Xie G, Liu Y, et al. HSVA: hierarchical semantic-visual adaptation for zero-shot learning. In: Proceedings of International Conference on Neural Information Processing Systems, 2021. 16622–16634
22 Verma V K, Brahma D, Rai P. Meta-learning for generalized zero-shot learning. In: Proceedings of AAAI Conference on Artificial Intelligence, 2020. 6062–6069
23 Liu Z, Li Y, Yao L, et al. Task aligned generative meta-learning for zero-shot learning. In: Proceedings of AAAI Conference on Artificial Intelligence, 2021. 8723–8731
24 Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations. In: Proceedings of International Conference on Machine Learning, 2020. 1597–1607
25 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning, 2021. 8748–8763
26 Jiang H, Wang R, Shan S, et al. Transferable contrastive network for generalized zero-shot learning. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2019. 9764–9773
27 Liu S, Long M, Wang J, et al. Generalized zero-shot learning with deep calibration network. In: Proceedings of International Conference on Neural Information Processing Systems, 2018. 2009–2019
28 Zhang H, Tian L, Wang Z, et al. Multiscale visual-attribute co-attention for zero-shot image recognition. IEEE Trans Neural Netw Learn Syst, 2023, 34: 6003–6014

29 Chen S, Hong Z, Xie G, et al. GNDAN: graph navigated dual attention network for zero-shot learning. IEEE Trans Neural Netw Learn Syst, 2024, 35: 4516–4529
30 Liu L, Zhou T, Long G, et al. Isometric propagation network for generalized zero-shot learning. In: Proceedings of International Conference on Learning Representations, 2021
31 Yu Y, Li B, Ji Z, et al. Knowledge distillation classifier generation network for zero-shot learning. IEEE Trans Neural Netw Learn Syst, 2023, 34: 3183–3194
32 Felix R, Kumar B G V, Reid I D, et al. Multi-modal cycle-consistent generalized zero-shot learning. In: Proceedings of European Conference on Computer Vision, 2018. 21–37
33 Li J, Jing M, Lu K, et al. Leveraging the invariant side of generative zero-shot learning. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2019. 7402–7411
34 Xu B, Zeng Z, Lian C, et al. Generative mixup networks for zero-shot learning. IEEE Trans Neural Netw Learn Syst, 2022. doi: 10.1109/TNNLS.2022.3142181
35 Xie G S, Zhang Z, Liu G, et al. Generalized zero-shot learning with multiple graph adaptive generative networks. IEEE Trans Neural Netw Learn Syst, 2022, 33: 2903–2915
36 Wang S, Li C, Li Y, et al. Self-supervised information bottleneck for deep multi-view subspace clustering. IEEE Trans Image Process, 2023, 32: 1555–1567
37 Xian Y, Sharma S, Schiele B, et al. F-VAEGAN-D2: a feature generating framework for any-shot learning. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2019. 10275–10284
38 Yu Y, Ji Z, Han J, et al. Episode-based prototype generating network for zero-shot learning. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2020. 14032–14041
39 Zhang S X, Zhu X, Chen L, et al. Arbitrary shape text detection via segmentation with probability maps. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 2736–2750
40 Zhang S X, Yang C, Zhu X, et al. Inverse-like antagonistic scene text spotting via reading-order estimation and dynamic sampling. IEEE Trans Image Process, 2024, 33: 825–839
41 Li Y, Wang S, Li C, et al. Towards very deep representation learning for subspace clustering. IEEE Trans Knowl Data Eng, 2024, 36: 3568–3579
42 Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2006. 1735–1742
43 Cheng P, Hao W, Dai S, et al. CLUB: a contrastive log-ratio upper bound of mutual information. In: Proceedings of International Conference on Machine Learning, 2020. 1779–1788
44 Fang Z, Lei S L, Zhu X, et al. Transformer-based reasoning for learning evolutionary chain of events on temporal knowledge graph. 2024. ArXiv:2405.00352
45 Zhou H, Zhu X, Zhu J, et al. Learning correction filter via degradation-adaptive regression for blind single image super-resolution. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2023. 12365–12375
46 Gutmann M, Hyvärinen A. Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In: Proceedings of International Conference on Artificial Intelligence and Statistics, 2010. 297–304
47 Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2015. 815–823
48 Hua W, Dai Z, Liu H, et al. Transformer quality in linear time. In: Proceedings of International Conference on Machine Learning, 2022. 9099–9117
49 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of International Conference on Neural Information Processing Systems, 2017. 5998–6008
50 Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning. In: Proceedings of International Conference on Machine Learning, 2017. 1243–1252
51 Tay Y, Bahri D, Metzler D, et al. Synthesizer: rethinking self-attention for transformer models. In: Proceedings of International Conference on Machine Learning, 2021. 10183–10192
52 Chou Y Y, Lin H T, Liu T L. Adaptive and generative zero-shot learning. In: Proceedings of International Conference on Learning Representations, 2021
53 Wah C, Branson S, Welinder P, et al. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, 2011
54 Patterson G, Hays J. Sun attribute database: discovering, annotating, and recognizing scene attributes. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2012. 2751–2758
55 Farhadi A, Endres I, Hoiem D, et al. Describing objects by their attributes. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2009. 1778–1785
56 Nilsback M E, Zisserman A. Automated flower classification over a large number of classes. In: Proceedings of Indian Conference on Computer Vision, Graphics & Image Processing, 2008. 722–729
57 Reed S E, Akata Z, Lee H, et al. Learning deep representations of fine-grained visual descriptions. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2016. 49–58
58 Kingma D P, Ba J. Adam: a method for stochastic optimization. In: Proceedings of International Conference on Learning Representations, 2015
59 Xian Y, Lorenz T, Schiele B, et al. Feature generating networks for zero-shot learning. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2018. 5542–5551
60 Verma V K, Arora G, Mishra A, et al. Generalized zero-shot learning via synthesized examples. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2018. 4281–4289
61 Ni J, Zhang S, Xie H. Dual adversarial semantics-consistent network for generalized zero-shot learning. In: Proceedings of International Conference on Neural Information Processing Systems, 2019. 6143–6154
62 Chen S, Wang W, Xia B, et al. Free: feature refinement for generalized zero-shot learning. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2021. 122–131
63 Narayan S, Gupta A, Khan F S, et al. Latent embedding feedback and discriminative features for zero-shot classification. In: Proceedings of European Conference on Computer Vision, 2020. 479–495
64 Chen Z, Luo Y, Qiu R, et al. Semantics disentangling for generalized zero-shot learning. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2021. 8692–8700
65 Chen S, Hong Z, Xie G S, et al. MSDN: mutually semantic distillation network for zero-shot learning. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2022. 7612–7621
66 Chen S, Hong Z, Liu Y, et al. TransZero: attribute-guided transformer for zero-shot learning. In: Proceedings of AAAI Conference on Artificial Intelligence, 2022. 330–338
67 Kong X, Gao Z, Li X, et al. En-Compactness: self-distillation embedding & contrastive generation for generalized zero-shot learning. In: Proceedings of IEEE Computer Vision and Pattern Recognition, 2022. 9306–9315
68 Burda Y, Grosse R B, Salakhutdinov R. Importance weighted autoencoders. In: Proceedings of International Conference on Learning Representations, 2016

Appendix A Equation (3): re-defined KL divergence based on joint distribution

In vanilla VAE [8], if the KL divergence dominates the optimization of model, the variational posterior will collapse to the prior when the KL divergence is 0, which is known as latent variable collapse[1)] or posterior collapse[2)]. A common practice is to maximize the mutual information between the observation and the latent variable. In our method, we adopt an approximate method, directly maximizing the joint distribution between the observation i and the latent variable z. The re-defined KL divergence can be formulated as

$$\begin{aligned} D_{\mathrm{KL}} &= \int q_\phi(z,i)\log\frac{q_\phi(z,\imath)}{p_\theta(z)}\mathrm{d}z \\ &= \int q_\phi(z\mid i)q_\phi(i)\log\frac{q_\phi(z\mid i)q_\phi(i)}{p_\theta(z)}\mathrm{d}z \\ &= \int q_\phi(z\mid i)q_\phi(i)\log\frac{q_\phi(z\mid i)}{p_\theta(z)}\mathrm{d}z + \int q_\phi(z\mid i)q_\phi(i)\log q_\phi(i)\mathrm{d}z \\ &= q_\phi(i)D_{\mathrm{KL}}\left(q_\phi(z\mid i)\|p_\theta(z)\right) + q_\phi(i)\log q_\phi(i), \end{aligned} \tag{A1}$$

where $D_{\mathrm{KL}}\left(q_\phi(z\mid i)\|p_\theta(z)\right)$ is the KL divergence of vanilla VAE, $q_\phi(i)\log q_\phi(i)$ corresponding the observation distribution is the variational lower bound of our KL divergence.

Appendix B Training processing of GZSL classifier

Algorithm B1 Training processing of GZSL classifier.

Input:
$D_{\mathrm{tr}} = \{(X_{\mathrm{tr}}, Y_{\mathrm{tr}}) | D_{\mathrm{tr}} \in D_s\}$: training dataset of seen classes;
$D_{\mathrm{te}} = \{(X_{\mathrm{te}}, Y_{\mathrm{te}}) | D_{\mathrm{te}} \in D_s \cup D_u\}$: testing dataset of GZSL task;
A_u: attribute vectors of unseen classes;
C_u: the corresponding labels of unseen classes;
C: the numbers of all classes;
Encoder: the two encoders of our AEFR model;
(N_s, N_u, N_m): sampling numbers.
Construct testing dataset for GZSL classifier
$\overline{X}_{\mathrm{te}} = \mathrm{Encoder}(X^{\mathrm{te}})$;
Get the new testing dataset $\overline{D}_{\mathrm{te}} = (\overline{X}_{\mathrm{te}}, Y_{\mathrm{te}})$;
Construct training dataset for GZSL classifier
$\mathrm{feat}_s, \mathrm{label}_s = \mathrm{sample\_train\_data}(X_{\mathrm{tr}}, Y_{\mathrm{tr}}, N_s)$;
$\mathrm{feat}_u, \mathrm{label}_u = \mathrm{sample\_train\_data}(A_u, C_u, N_u)$;
$\mathrm{feat}_m, \mathrm{label}_m = \mathrm{sample\_train\_data}(A_u, C_u, N_m)$;
$(Z_x^s, Z_a^u, Z_m^u) = \mathrm{Encoder}([\mathrm{feat}_s, \mathrm{feat}_u, \mathrm{feat}_m])$;
$(Z_m^u, \mathrm{label}_m) = \mathrm{MIXUP}(z_m^u, \mathrm{label}_m)$;
$\overline{X}_{\mathrm{tr}} = \mathrm{concat}([Z_x^s, Z_a^u, Z_m^u])$;
$\overline{Y}_{\mathrm{tr}} = \mathrm{concat}([\mathrm{label}_s, \mathrm{label}_u, \mathrm{label}_m])$;
Get the new training dataset $\overline{D}_{\mathrm{tr}} = (\overline{X}_{\mathrm{tr}}, \overline{Y}_{\mathrm{tr}})$;
# Train and test the GZSL classifier with $(\overline{D}_{\mathrm{tr}}, \overline{D}_{\mathrm{te}})$
classifier = LINEAR_LOGSOFTMAX(C).init();
accuracy = classifier.fit($\overline{D}_{\mathrm{tr}}, \overline{D}_{\mathrm{te}}$);
return classifier.

1) Alemi A A, Poole B, Fischer I, et al. Fixing a broken ELBO. In: Proceedings of International Conference on Machine Learning, 2018. 159–168.

2) Razavi A, van den Oord A, Poole B, et al. Preventing posterior collapse with delta-VAEs. In: Proceedings of International Conference on Learning Representations, 2019.