Learning Aligned Cross-Modal Representation for Generalized Zero-Shot Classification

Zhiyu Fang, Xiaobin Zhu*, Chun Yang, Zheng Han, Jingyan Qin, Xu-Cheng Yin

School of Computer & Communication Engineering, University of Science and Technology Beijing, Beijing, China
{mr.fangzy, han970421, qinjingyanking}@foxmail.com, {zhuxiaobin, chunyang, xuchengyin}@ustb.edu.cn

Abstract

Learning a common latent embedding by aligning the latent spaces of cross-modal autoencoders is an effective strategy for Generalized Zero-Shot Classification (GZSC). However, due to the lack of fine-grained instance-wise annotations, it still easily suffer from the domain shift problem for the discrepancy between the visual representation of diversified images and the semantic representation of fixed attributes. In this paper, we propose an innovative autoencoder network by learning Aligned Cross-Modal Representations (dubbed ACMR) for GZSC. Specifically, we propose a novel Vision-Semantic Alignment (VSA) method to strengthen the alignment of cross-modal latent features on the latent subspaces guided by a learned classifier. In addition, we propose a novel Information Enhancement Module (IEM) to reduce the possibility of latent variables collapse meanwhile encouraging the discriminative ability of latent variables. Extensive experiments on publicly available datasets demonstrate the state-of-the-art performance of our method.

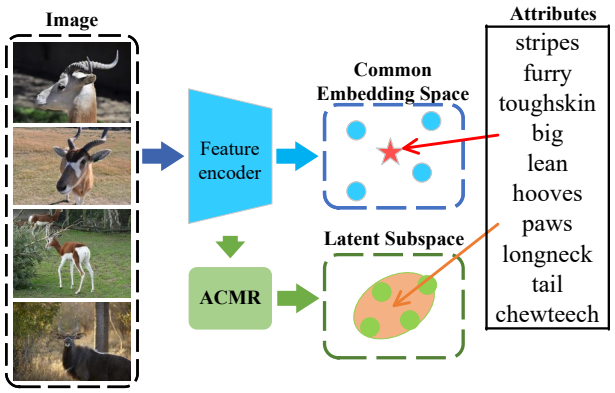


Figure 1: Illustration of the overall concept of our method. Our ACMR strengthens the alignment of cross-modal latent features on latent subspaces guided by a learned classifier.

Introduction

Zero-Shot learning (ZSL) (Larochelle, Erhan, and Bengio 2008; Rahman, Khan, and Porikli 2018; Wang et al. 2018; Wei, Deng, and Yang 2020) aims to recognize entirely new classes without additional data labeling, which is an important issue for learning and recognition in open environments. Conventional ZSL methods try to establish the connection between the visual and semantic space by seen classes, and only the images in unseen classes are available in the testing phase. As an extension, Generalized Zero-Shot Learning (GZSL) removes the constraint in conventional ZSL by allowing images from either seen or unseen classes in the testing phase, which is a more reasonable yet challenging for real-world recognition.

Recently, GZSL has received ever-increasing attention in research and industrial communities. Many GZSL methods (Chen et al. 2018; Jia et al. 2020; Han, Fu, and Yang 2020; Huynh and Elhamifar 2020) try to learn a mapping between visual features of images and their class embeddings for knowledge transfer. Huynh and Elhamifar (Huynh and Elhamifar 2020) adopted an attention mechanism to align attribute embeddings with local visual features. Han *et al.* mapped the class-level semantic features into a redundancy-free feature space constructed from visual features. Some other methods (Yu and Lee 2019; Vyas, Venkateswara, and Panchanathan 2020; Gao et al. 2020; Yue et al. 2021) try to generate synthetic training samples of unseen classes for GZSL. Vyas *et al.* (Vyas, Venkateswara, and Panchanathan 2020) proposed a semantic regularized loss for conditional WGAN to generate visual features for bridging the distribution gap between seen and unseen classes. However, because the distribution of visual features in seen and unseen classes are generally distinguished, unidirectional mappings can easily induce the domain shift problem (Fu et al. 2015; Zhu, Wang, and Saligrama 2019).

To alleviate the domain shift problem, some methods (Xian et al. 2016; Tsai, Huang, and Salakhutdinov 2017; Verma et al. 2018) try to learn common latent spaces by cross-modal alignment restriction. In this paradigm, Variational AutoEncoder (VAE) based architectures are popular for the capability of fitting data distribution in latent space. Schonfeld *et al.* (Schönfeld et al. 2019) employed two VAEs to learn the joint reconstruction between visual features and semantic features in the common latent space. Ma *et al.* (Ma and Hu 2020) incorporated a deep embedding network and modified variational autoencoder to learn a latent space shared by both image features and class embeddings. Li *et*

*Correspondence author
Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

al. (Li et al. 2021) respectively decoupled category-distilling factors and category-dispersing factors from visual and semantic feature for GZSL. However, due to the lack of fine-grained instance-wise annotations, the aligned relationships across different modalities tend to be confused on the common embedding space without structural restriction, which is detrimental for generating effective latent features to train a classifier. A plausible solution is to further restrict cross-modality alignment on subspace, especially for image classification.

From the above-mentioned observations, we propose an innovative autoencoder network by learning Aligned Cross-Modal Representations (dubbed ACMR) for GZSC. Figure 1 illustrates the motivation of our method. Our ACMR is constructed on the popular Variational Auto-Encoders (VAEs) to encode and decode features from different modalities and align them on the latent space. To be specific, a novel Vision-Semantic Alignment (VSA) is proposed to further strengthen the alignment of cross-modal latent features on latent subspace guided by a learned classifier. In addition, a novel Information Enhancement Module (IEM) is proposed to reduce the possibility of latent variables collapse by maximizing joint distribution, meanwhile encouraging the discriminative ability of latent variables. Finally, the aligned cross-modal representations are used to train a softmax classifier for GZSC.

In summary, our main contributions are three-fold:

- We propose an innovative autoencoder network by learning Aligned Cross-Modal Representations for GZSC. Experimental results verify the state-of-the-art performance of our method on four publicly available datasets.
- Proposing an innovative Vision-Semantic Alignment (VSA) method to strengthen the alignment of cross-modal latent features on latent subspace guided by a learned classifier.
- Proposing a novel Information Enhancement Module (IEM) to reduce the possibility of latent variables collapse meanwhile enhancing the discriminative ability of latent variables.

Related Work

Generalized Zero-Shot Learning

GZSL is a more realistic than ZSL because it not only learns information which can be adapted to unseen classes but also apply to the testing data from seen classes. The classical GZSL methods learn a mapping function and can be roughly divided into three categories.

One type of methods directly extracts visual features for mapping to semantic space. DAP and IAP (Lampert, Nickisch, and Harmeling 2009) establish the rigid correspondence between original visual features and attribute vectors for GZSL. The popular attention mechanisms also have been adopted to distinguish informative local visual features for GZSL, such as (Xie et al. 2019; Niu et al. 2019; Huynh and Elhamifar 2020; Liu et al. 2021; Ge et al. 2021). Considering the distribution consistency of semantic features in all categories, some methods aims to extract semantic features from textual descriptions for mapping to visual space. DeViSE (Frome et al. 2013) maps the visual features into word2vector embedding space learned by Skip-gram Language Model (Mikolov et al. 2013). Moreover, the generative models (Felix et al. 2018; Wan et al. 2019; Vyas, Venkateswara, and Panchanathan 2020; Feng and Zhao 2021; Yue et al. 2021) try to generate synthetic images of unseen classes for GZSL.

Due to the above methods often conduct mapping in a unidirectional manner, which tend to easily cause the domain shift problem. Therefore, the common-space based methods try to map different features into common spaces for connecting visual features with semantic features (Akata et al. 2015b,a; Sung et al. 2018; Keshari, Singh, and Vatsa 2020). CVAE (Mishra et al. 2018) utilizes conditional VAE to encode the concatenation feature of visual and semantic embedding vectors. OCD (Keshari, Singh, and Vatsa 2020) designs an over-complete distribution of seen and unseen classes to enhance class separability. Although these methods extract the effective features on the common space, they often neglect explicitly aligning different modalities.

Cross-Modal Alignment Models for GZSL

Recent cross-alignment based methods of GZSL generally focus on establishing a shared latent space for visual and semantic features via generative adversarial network (GAN) (Goodfellow et al. 2014) or variational autoencoder (VAE) (Kingma and Welling 2013). The pioneer GAN-based method, SJE (Akata et al. 2015b) learns a convolutional GAN framework to synthesize images from high-capacity text via minimax compatibility between image and text. LiGAN (Li et al. 2019) employs GAN to map features into synthetic sample space. f-VAEGAN-D2 (Xian et al. 2019) learns the marginal feature distribution of unlabeled images via a conditional generative model and an unconditional discriminator. cycle-GAN (Felix et al. 2018) and EPGN (Yu et al. 2020) strengthen the alignment of latent features in their respective modalities by cross-reconstruction. However, GAN-based models are difficult to optimize parameters during the training process.

VAE is an autoencoder for learning the conditional probability distribution from training data, which helps to generate samples having certain desired properties. CADA-VAE (Schönfeld et al. 2019) combines cross-modal and cross-reconstruction, and leverages two independent VAEs to align features via minimizing distance between the latent distributions of visual and semantic features. Based on CADA-VAE, Li *et al.* (Li et al. 2020) learns modality-invariant latent representations by maximizing mutual information and entropy on latent space. DE-VAE (Ma and Hu 2020) adopts a deep embedding model to learn the mapping from the semantic space to the visual feature space. Disentangled-VAE (Li et al. 2021) decouples the latent features by shuffling classification-irrelevant information to obtain discriminative representations. Although the VAE-based methods achieve promising performance, the aligned relationships across different modalities tend to be confused on the common embedding space without structural restriction since the lack of fine-grained instance-wise annotations.

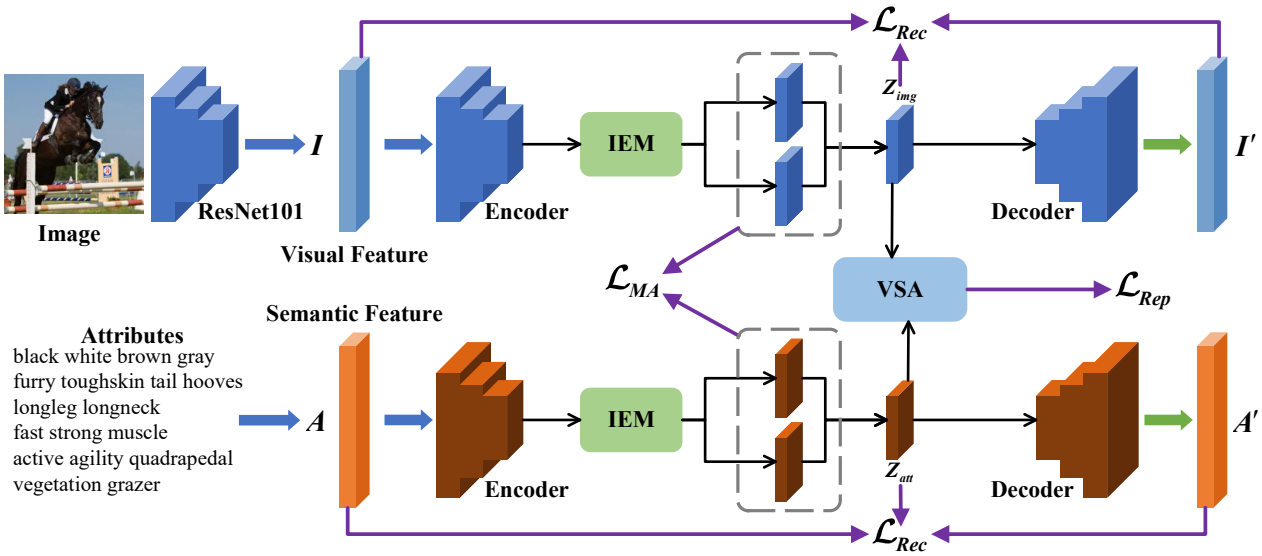


Figure 2: Architecture of our Aligned Cross-Modal Representations (ACMR) network. Two parallel VAEs as backbones extract visual and semantic latent representations, respectively. IEM (detailed in Figure 3) aims to learn more discriminative latent representation. VSA (detailed in Figure 4) restrains the alignment between two modalities by classification optimization.

Methodology

In this section, we first present the definition of GZSL and the analysis of cross-modal alignment in common space. Then we briefly introduce VAE as the basic building block of our model, and elaborate our method. Finally, we summarize the loss function and training process of our method.

Generalized Zero-Shot Learning

Suppose that $X=\{X^S,\ X^U\}$ denotes the visual space of images, $A=\{A^S,\ A^U\}$ denotes the semantic space of attributes, and $Y=\{Y^S,\ Y^U\}$ denotes the corresponding label set. Notably, Y^S and Y^U are disjoint, i.e., $Y \cap Y^U=\emptyset$. Consequently, we can individually collect training dataset $D_s=\{x_i^s, a_i^s, y_i^s\}_{i=1}^N$ and testing 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. The task of ZSL aims to learn a classifier $f_{ZSL} : X \rightarrow Y^U$ for recognizing a testing instance x of unseen classes. For adapting to both seen and unseen classes, GZSL adopts a more realistic setting and learn a classifier $f_{GZSL} : X \rightarrow Y^U \cup Y^S$.

Cross-Modal Alignment

Let Z_x, Z_a respectively denote the feature vectors of visual and semantic modalities, the joint distribution $P(Z_x, Z_a)$ can assess the degree of cross-modal alignment. For example, if two modalities are completely aligned $P(Z_x, Z_a) = 1$, otherwise $0 \leq P(Z_x, Z_a) < 1$. Optimizing cross-modal alignment can be mathematically formulated as:

$$2\log p(z_x, z_a) \geq \log p(z_x|z_a) + \log p(z_a|z_x) + \mathcal{H}(X) + \mathbb{E}_{p(x)}[\log p(z_x, X)] + \mathbb{E}_{p(a)}[\log p(z_a, A)] + \mathcal{H}(A) \quad (1)$$

where $\mathbb{E}(\cdot)$ represents expectation, $\mathcal{H}(\cdot)$ represents entropy.

Aligned Cross-Modal Representation Network

For conducting cross-modal alignment, we can train a deep learning model for mapping image features and attributes into a common space. According to the optimization objective (Equation 1), the common space can be utilized to align the features by the first two items, the features on the common space can contain discriminative information corresponding inputs by the middle two items, and the features can be utilized to reconstruct original modalities by the last two items. Consequently, we summarize Equation 1 into three parts. The first two log-likelihood items $\log p(\cdot)$ can be realized by a mutual alignment constraint C_{MA}, the middle two expectation items $\mathbb{E}$ can be realized by a representation constraint C_{Rep}, and the last two entropy items $\mathcal{H}$ can be realized by a reconstruction constraint C_{Rec}. Hence, the total optimization objective can be simplified as:

$$\mathcal{O} = C_{MA} + C_{Rep} + C_{Rec} \quad (2)$$

Variational Auto-Encoder. Since the excellent ability of fitting data distribution, VAE is widely used in cross modal alignment. We employ two independent VAEs (shown in Figure 2) composed of multilayer perceptron (MLP) to respectively establish visual and semantic modality. In VAE, the encoders are used to generate the latent variables of each modality, and the decoders are used to reconstruct the latent vectors. Therefore, we can estimate the reconstruction constraint C_{Rec} with VAE loss, which can be formulated as:

$$\mathcal{L}_{Rec} = \mathbb{E}_{q_\phi(z_x|x)}[\log p(x|z_x)] - \alpha D_{KL}(q_\phi(z_x|x)\|p_\theta(z_x)) + \mathbb{E}_{q_\phi(z_a|a)}[\log p(a|z_a)] - \alpha D_{KL}(q_\phi(z_a|a)\|p_\theta(z_a)) \quad (3)$$

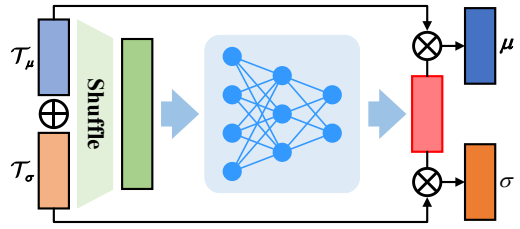


Figure 3: Architecture of our IEM.

where $q_\phi(z_x|x)$ denotes latent feature distributions of visual modality, $q_\phi(z_a|a)$ denotes latent feature distributions of semantic modality, $p(x|z_x)$ and $p(a|z_a)$ denote the observation distributions of corresponding modalities, α weights the KL Divergence. The priors $p_\theta(z_x)$ and $p_\theta(z_a)$ are both standard Gaussian distributions.

Because VAE models the distribution of hidden variables, we can transform the mutual alignment constraint C_{MA} of maximizing log likelihood into minimizing the Wasserstein distance between the latent Gaussian distributions of semantic modality and visual modality. Hence, the mutual alignment constraint can be formulated as:

$$\mathcal{L}_{MA} = \left(\|\mu_x - \mu_a\|_2^2 + \left\| \Sigma_x^{\frac{1}{2}} - \Sigma_a^{\frac{1}{2}} \right\|_{\text{Frobenius}}^2 \right)^{\frac{1}{2}} \tag{4}$$

where μ_x and Σ_x are all predicted by visual encoder, while μ_a and Σ_a are all predicted by semantic encoder.

Information Enhancement Module. Because of the "latent variable collapse" problem (Alemi et al. 2018), the distribution of latent vector $p(Z|I)$ generated by VAE can easily degrade into a standard normal distribution $\mathcal{N}(0,1)$, which is detrimental to learn discriminative representations of latent variables for classification. From this key observation, we propose a novel Information Enhancement Module (IEM, shown in Figure 3) to magnify the joint probability between observation I and the inferred latent variables Z.

According to (Chen et al. 2019), we randomly shuffle the features of encoder to extract discriminative features. For obtaining the joint probability, we develop a MLP with one hidden layer to learn the marginal probability of I, and multiply it with the conditional probability of Z. Mathematically, our IEM can be formulated as:

$$\begin{aligned} p(z_x, X) &= p(E(X;\varphi_x) \mid X) p(\Psi(X;\psi_x)) \\ p(z_a, A) &= p(E(A;\varphi_a) \mid A) p(\Psi(A;\psi_a)) \end{aligned} \tag{5}$$

where $p(z_x, X)$, $p(z_a, A)$ denote latent variable distributions with enhanced mutual information of visual modality and semantic modality; $E(X;\varphi_x), E(A;\varphi_a)$ respectively denote visual modality encoder and semantic modality encoder; $\Psi(X;\psi_x), \Psi(A;\psi_a)$ are marginal probability perceptrons of corresponding modalities.

When optimizing the vanilla VAE with joint probability, KL divergence can be re-defined as:

$$\begin{aligned} D_{KL} &= \int q_\phi(z,i) \log \frac{q_\phi(z,i)}{p_\theta(z)} dz \\ &= q_\phi(i) D_{KL}(q_\phi(z \mid i) \parallel p_\theta(z)) + q_\phi(i) \log q_\phi(i) \end{aligned} \tag{6}$$

According to Equation 3, the latent variable will collapse with the decrease of $\mathcal{L}_{Rec}$. Then, $D_{KL}(q_\phi(z|i) \| p_\theta(z))$ will be approximate to 0. But with $q_\phi(i) \log q_\phi(i)$ containing distribution of observation i in re-defined D_{KL}, z will not degrade to $\mathcal{N}(0,1)$ and tend to contain more effective information about true data distribution.

Vision-Semantic Alignment. Vision-Semantic Alignment (VSA, shown in Figure 4) is proposed to align the multi-modal latent variables Z after information enhancement. Since the lack of fine-grained instance-wise annotation, images of each class share a pre-class attribute vector, the alignment between different modalities tend to be confused on the latent space without any structural restriction. To tackle this issue, we further constrain the latent features on latent subspace guided by a learned classifier to strengthen the alignment. We introduce classification information Y into the optimization function and maximize the joint distribution $p(Z, I, Y)$. Then the optimization function can be obtained by replacing $p(z_x, X), p(z_a, A)$ in the middle two items of Equation 1 with $p(z_x, X, Y)$, $p(z_a, A, Y)$. Then the two expectation items of Equation 1 can be respectively formulated as:

$$\begin{aligned} &\mathbb{E}_{p(x)}[\log p(z_x, X)] = \\ &\quad \mathbb{E}_{p(y)}\mathbb{E}_{p(x)}[\log p(Y|z_x, X)] + \mathbb{E}_{p(y)} \log p(z_x, X) \\ &\mathbb{E}_{p(a)}[\log p(z_a, A)] = \\ &\quad \mathbb{E}_{p(y)}\mathbb{E}_{p(a)}[\log p(Y|z_a, A)] + \mathbb{E}_{p(y)} \log p(z_a, A) \end{aligned} \tag{7}$$

where $p(Y|z_x, X), p(Y|z_a, A)$ denote classification probability of corresponding modalities in latent space, the last $\mathbb{E}(\cdot)$ represent joint expectation. This formula highlights that maximizing the optimization objective is equivalent to maximize the conditional probability $p(Y|z_x, X)$ and $p(Y|z_a, A)$ and to maximize the expectation of the joint distribution $\mathbb{E}(\cdot)$. Hence, $\mathbb{E}_{p(y)}\mathbb{E}_{p(x)}[\cdot]$ item can be approximated by a classification loss based on latent variables, $\mathbb{E}(\cdot)$ item can be approximated by a reconstruction loss based on latent variables after IEM. In this way, we can constrain the generation of latent variables by the classifier to improve its discriminability.

By replacing the expectation items in representation constraint C_{Rep} with latent variable representation after information enhancement (shown in Equation 7). The new C_{Rep} are implemented by IEMs and VSA modules, and the entropy items as reconstruction are merged into VAE. Hence we can denote the representation loss as:

$$\begin{aligned} \mathcal{L}_{Rep} = &\frac{1}{N} \sum_{i=1}^{N} L_{cls}(y_i^s, f(z_i^x; W_x)) + \\ &\frac{1}{N} \sum_{i=1}^{N} L_{cls}(y_i^s, f(z_i^a; W_a)) \end{aligned} \tag{8}$$

where z_i^x, z_i^a, y_i^s are visual latent variable, semantic variable, and label in training set, L_{cls} is the classification loss, f is the classifier composed of two independent block.

Loss function. The overall loss of our method can be formulated as:

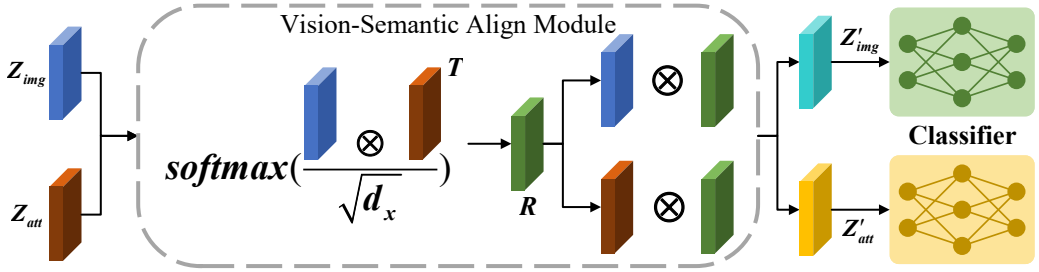


Figure 4: Architecture of Vision-Semantic Alignment Module (VSA). d_x denotes normalization factor, T denotes transpose operator, and R denotes joint weighting coefficient of latent variables.

Dataset	$\mathcal{D}_I$	$\mathcal{D}_A$	$\mathcal{N}_X$	$\mathcal{N}_S/\mathcal{N}_U$
CUB	2048	312	11788	150/50
SUN		102	14340	707/10
AwA1		85	30475	40/10
AwA2		85	37322	40/10

Table 1: Common datasets in GZSL, in terms of dimensionality of visual features $\mathcal{D}_I$, dimensionality of semantic features $\mathcal{D}_A$, number of all instances $\mathcal{N}_X$, number of seen classes $\mathcal{N}_S$ and number of unseen classes $\mathcal{N}_U$.

$$\mathcal{L} = \mathcal{L}_{Rec} + \beta\mathcal{L}_{MA} + \lambda\mathcal{L}_{Rep} \tag{9}$$

where $\mathcal{L}_{Rec}$ is the reconstruction loss of intra-modalities, $\mathcal{L}_{MA}$ is the mutual alignment loss of inter-modalities, and $\mathcal{L}_{Rep}$ is the classification loss of latent representations. β and λ are two hyper-parameters training by an annealing schedule (Bowman et al. 2016).

Experiments

Datasets and Evaluation Protocol

We conduct experiments on four benchmark datasets: CUB (Wah et al. 2011), SUN (Patterson and Hays 2012), AwA1 (Lampert, Nickisch, and Harmeling 2009), and AwA2 (Xian et al. 2018a). The detailed information of the datasets is summarized in Table 1. 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, which is widely used in objective detection, image classification, and so on. In SUN, images are annotated with 102-dimensional attribute. AwA1 is an animal image dataset that uses 85-dimensional attribute to describe 50 categories. AwA2 is a more sophisticated variant of AwA1, which collects 37,322 images from public web sources. AwA1 and AwA2 have no overlapping images.

For a fair comparison, we adopt the setting as in (Xian et al. 2018a) for training and testing. The evaluation metrics in generalized datasets setting include the average classification accuracy (ACA) of the samples from seen/unseen classes, the harmonic mean of ACA^S and ACA^U, details are listed as followings:

- ACA^U: the average accuracy of per-class on test images from unseen classes, which represents the capacity of classifying unseen classes samples.
- ACA^S: the average accuracy of per-class on test images from seen classes, which is used to represent the capacity of classifying seen classes samples.
- H: the harmonic mean value, which is formulated as

$$H = \frac{2 \times ACA^U \times ACA^S}{ACA^U + ACA^S} \tag{10}$$

Implementation Details

Following the setting in other methods (Xian et al. 2018a; Schönfeld et al. 2019), ResNet-101 pre-trained on ImageNet is used to extract the 2048 dimensional visual features. In IEM, the hidden layer of MLP has 99 units. We respectively utilize 1560, 1680 hidden units for visual encoder and decoder, and 1450, 665 hidden units for semantic encoder and decoder. The size of our aligned cross-modal representation in latent space is 64 for all datasets. The final softmax classifier includes a fully connected layer and a non-linear activation layer. The input size of the fully connected layer is 64, and the output size is equal to the total number of all categories. Our approach is implemented with PyTorch 1.5.0 and trained for 100 epochs by the Adam optimizer (Kingma and Ba 2015).We set learning rate as 1.5e-04 for training VAEs, 3.3e-05 for training IEMs, 7.4e-03 for training VSA, 0.5e-03 for training softmax classifier. For all datasets, the batch size of ACMR is set to 50 and the batch size of final softmax classifier is set to 32.

Comparing with the State-of-the-Arts

Baseline models. We compare our model with 15 state-of-the-art (SOTA) models. Among them, SP-AEN (Chen et al. 2018), AREN (Xie et al. 2019), DAZLE (Huynh and Elhamifar 2020) are deep embedding based methods, cycle-CLSWGAN (Felix et al. 2018), f-CLSWGAN (Xian et al. 2018b), LiGAN (Li et al. 2019), f-VAEGAN-D2 (Xian et al. 2019), DASCN (Ni, Zhang, and Xie 2019), LsrGAN (Vyas, Venkateswara, and Panchanathan 2020), E-PGN (Yu et al. 2020) are GAN based methods, and SE (Verma et al. 2018), CVAE (Mishra et al. 2018), CADA-VAE (Schönfeld et al. 2019), OCD (Keshari, Singh, and Vatsa 2020), DE-VAE (Ma and Hu 2020), Disentangled-VAE (Li et al. 2021) are VAE-based methods.

	Model	CUB			SUN			AwA1			AwA2		
		U	S	H	U	S	H	U	S	H	U	S	H
Embedding	SP-AEN (Chen et al. 2018)	34.7	70.6	46.6	24.9	38.6	30.3	-	-	-	23.3	90.9	37.1
	AREN (Xie et al. 2019)	38.9	78.7	52.1	19.0	38.8	25.5	-	-	-	15.6	92.9	26.7
	DAZLE (Huynh and Elhamifar 2020)	42	65.3	51.1	25.7	82.5	25.8	-	-	-	25.7	82.5	39.2
GAN	cycle-CLSWGAN (Felix et al. 2018)	59.3	47.9	53	33.8	47.2	39.4	63.4	59.6	59.8	-	-	-
	f-CLSWGAN (Xian et al. 2018b)	57.7	43.7	49.7	36.6	42.6	39.4	61.4	57.9	59.6	53.8	68.2	60.2
	LiGAN (Li et al. 2019)	46.5	57.9	51.6	42.9	37.8	40.2	52.6	76.3	62.3	54.3	68.5	60.6
	f-VAEGAN-D2 (Xian et al. 2019)	60.1	48.4	53.6	38	45.1	41.3	70.6	57.6	63.5	-	-	-
	DASCN (Ni, Zhang, and Xie 2019)	59	45.9	51.6	38.5	42.4	40.3	68	59.3	63.4	-	-	-
	LsrGAN (Vyas et al. 2020)	59.1	48.1	53	37.7	44.8	40.9	74.6	54.6	63	74.6	54.6	63
	E-PGN (Yu et al. 2020)	57.2	48.5	52.5	-	-	-	86.3	52.6	65.3	83.6	48	61
VAE	SE (Verma et al. 2018)	53.3	41.5	46.7	30.5	40.9	34.9	67.8	56.3	61.5	68.1	58.3	62.8
	CVAE (Mishra et al. 2018)	–	–	34.5	–	–	26.7	–	–	47.2	–	–	51.2
	CADA-VAE (Schönfeld et al. 2019)	51.6	53.5	52.4	47.2	35.7	40.6	57.3	72.8	64.1	55.8	75	63.9
	OCD (Keshari, Singh, and Vatsa 2020)	59.9	44.8	51.3	42.9	44.8	43.8	-	-	-	73.4	59.5	65.7
	DE-VAE (Ma and Hu 2020)	52.5	56.3	54.3	45.9	36.9	40.9	59.6	76.1	66.9	58.8	78.9	67.4
	Disentangled-VAE (Li et al. 2021)	51.1	58.2	54.4	36.6	47.6	41.4	60.7	72.9	66.2	56.9	80.2	66.6
Ours	ACMR	53.1	57.7	**55.3**	49.1	39.5	**43.8**	59.4	77.6	**67.5**	60.0	80.2	**68.7**

Table 2: Performance of classification on four benchmarks under PS (proposed split). ts=ACA on $U = ACA^U$, $S = ACA^S$, H is the harmonic mean. (The best H is marked bold).

Model	CUB	SUN	AwA1	AwA2
Dual-VAE	32.7	38.3	52.4	66.1
IE-VAE	36.1	39.0	65.1	66.6
VSA-VAE	54.1	43.5	65.7	67.9
ACMR	**55.2**	**43.8**	**67.3**	**68.7**

Table 3: Ablation study on different module combinations. The results are H (harmonic mean) on four datasets.

The detailed results are listed in Table 2. In our method, the value of H can reach 55.3% on CUB, 43.8% on SUN, 67.5% on AwA1, and 68.7% on AwA2, respectively. Specifically, our method respectively outperforms the original benchmark CADA-VAE by 2.9% on CUB, by 3.2% on SUN, by 3.4% on AwA1, and by 4.8% on AwA2 in terms of H. Compared with deep embedding based methods, our ACMR outperforms DAZLE on three datasets. Especially, our ACMR significantly increases H by 18% and 29.5% on SUN and AwA2, respectively. Moreover, our method outperforms APN (Xu et al. 2020) by 3.1% (H) on AwA2, by 0.1% (H) on SUN, and outperforms RGEN (Xie et al. 2020) by 7% (H) on SUN. Compared with GAN-based methods, our ACMR outperforms LsrGAN by 2.3% on CUB, by 2.9% on SUN, by 4.5% on AwA1, and by 5.7% on AwA2 in terms of H, respectively. We can respectively increase H by 2.8% on CUB, by 2.2% on AwA1, and by 7.7% on AwA2, compared with E-PGN whose H are 52.5%, 65.3% and 61% corresponding datasets. Compared with VAE-based methods, our ACMR outperforms Disentangled-VAE by 0.9% on CUB, by 2.4% on SUN, by 1.3% on AwA1, and by 2.1% on AwA2 in terms of H, respectively. Compared with DE-VAE, we can increase H from 54.3% to 55.3% on CUB, from 40.9% to 43.8% on SUN, from 66.9% to 67.5% on AwA1 and from 67.4% to 68.7% on AwA2. Obviously, our method achieves SOTA performance and outperform other common space based methods with a great margin.

Ablation Study

In this section, we conduct ablation study to evaluate the effectiveness and necessity of our proposed components, including backbone, IEM, VSA, and the hyper-parameters of weighting loss.

About the IEM and VSA. We present the results under different combinations of our proposed modules in ACMR, including Dual-VAE (only two VAEs), IE-VAE (two VAEs and two IEMs), VSA-VAE (two VAEs and VSA module) and our original ACMR. The detailed results are listed in Table 3.

From Table 3, we can see that VSA-VAE outperforms Dual-VAE by 21.4% on CUB, by 5.2% on SUN, by 13.3% on AwA1, and by 1.8% on AwA2 in terms of H. It means that discriminative representations can be learned by aligning cross modal features guided by classifier. In addition, compared with Dual-VAE, IE-VAE increases H from 32.7% to 36.1% on CUB, from 38.3% to 39.0% on SUN, from 52.4% to 65.1% on AwA1, and from 66.1% to 66.6% on AwA2, respectively. Compared with VAS-VAE, ACMR increases H from 54.1% to 55.2% on CUB, from 43.5% to 43.8% on SUN, from 65.7% to 67.3% on AwA1, and from 67.9% to 68.7% on AwA2, respectively. It means that the information enhancement module can effectively strengthen the discriminative ability of latent variables.

Inference about weighting coefficient α, β and λ. These three hyper-parameters are the weighting coefficient of our loss. In this experiment, we vary α from 1 to 5, β from 1 to 12 for all datasets, and λ from 285 to 305 for CUB, SUN, AwA1, from 70 to 90 for AwA2, the detailed results are shown in Figure 5. From α-curve of all datasets, we can see that H generally keeps rising along with the increase of α. From β-curve of all datasets, we can see that H rises

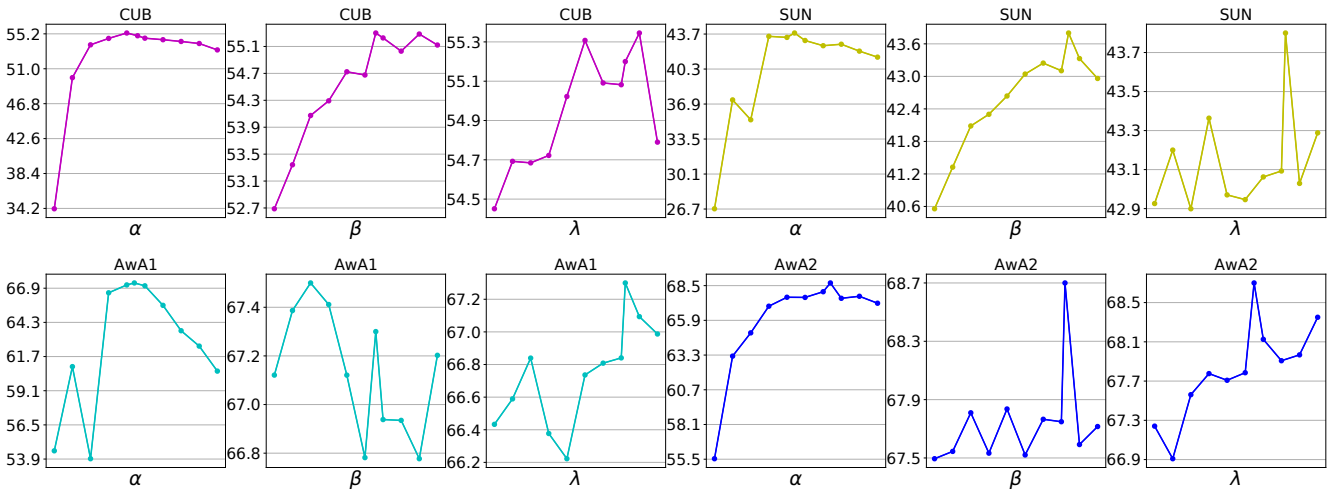


Figure 5: Influence of the weighting coefficients of losses. We measure H (harmonic mean) on 4 datasets. Different colors of curves represent different datasets.

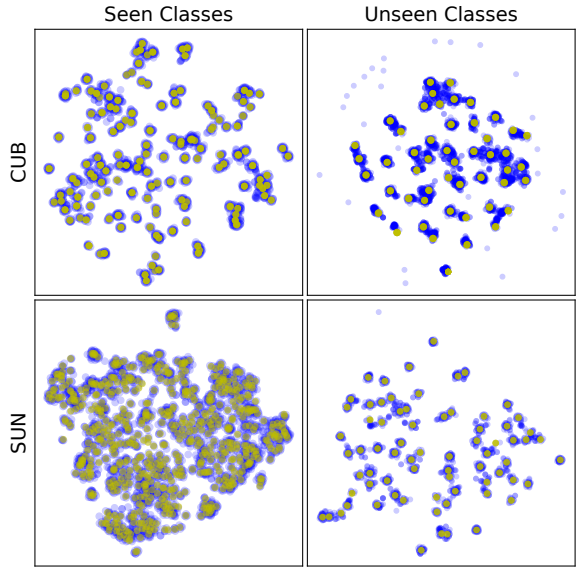


Figure 6: t-SNE visualization of cross-modal features alignment on CUB. The visual features in blue, and the semantic features in yellow. (Left: seen classes. Right: unseen classes)

smoothly on CUB, SUN, and fluctuates greatly on AwA1, AwA2 with the increase of β. From λ-curve of all datasets, we can see that performance fluctuates greatly with the increase of classification weighting coefficient λ, we can conclude that extracting classification subspace features is more helpful to GZSC than extracting common space features.

Visualization. According to the method of qualitatively measuring the extent of two domains (Saito et al. 2019), we leverage t-SNE (Van der Maaten and Hinton 2008) to visualize the features distribution of visual modality and semantic modality in the latent space on CUB and SUN, to visually show that ACMR our proposed can learn well aligned cross-modal representation, as shown in Figure 6. We extract the latent features of seen and unseen classes on testing datasets by the trained ACMR respectively, and use blue dots to represent the features of the visual modality, and yellow dots to represent the features of the semantic modality. From Seen Classes column of CUB and SUN, we can observe that visual features and semantic features overlap almost completely with each other. From Unseen Classes column, although some semantic features have slightly shifted, we can observe that features of two modalities also coincide well. Hence, it verifies the effectiveness of our proposed ACMR.

Conclusion

In this work, we proposed an innovative autoencoder network via learning aligned cross-modal representations (ACMR) for generalized zero-shot classification (GZSC). Our main contributions lie in innovative Vision-Semantic Alignment (VSA) and Information Enhancement Module (IEM). VSA strengthens the alignment of cross-modal latent features on latent subspace by a learned classifier, while IEM reduces the possibility of latent variables collapse and improve the discriminative ability of latent variables in respective modalities. Extensive experiments on four publicly available datasets demonstrate the state-of-the-art performance of our method and verify the effectiveness of our aligned cross-modal representation on latent subspace guided by classification.

Acknowledgements

This research was supported by the National Key Research and Development Program of China (2020AAA09701), National Science Fund for Distinguished Young Scholars (62125601), National Natural Science Foundation of China (62172035, 62006018, 61806017, 61976098).

References

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2015a. Label-embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(7): 1425–1438.

Akata, Z.; Reed, S. E.; Walter, D.; Lee, H.; and Schiele, B. 2015b. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2927–2936.

Alemi, A. A.; Poole, B.; Fischer, I.; Dillon, J. V.; Saurous, R. A.; and Murphy, K. 2018. Fixing a Broken ELBO. In *Proceedings of the International Conference on Machine Learning*, 159–168.

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A. M.; Józefowicz, R.; and Bengio, S. 2016. Generating Sentences from a Continuous Space. In *Proceedings of the Conference on Computational Natural Language Learning*, 10–21.

Chen, L.; Zhang, H.; Xiao, J.; Liu, W.; and Chang, S. 2018. Zero-Shot Visual Recognition Using Semantics-Preserving Adversarial Embedding Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1043–1052.

Chen, Y.; Bai, Y.; Zhang, W.; and Mei, T. 2019. Destruction and Construction Learning for Fine-Grained Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5157–5166.

Felix, R.; Kumar, B. G. V.; Reid, I. D.; and Carneiro, G. 2018. Multi-modal Cycle-Consistent Generalized Zero-Shot Learning. In *Proceedings of the European Conference on Computer Vision*, 21–37.

Feng, L.; and Zhao, C. 2021. Transfer Increment for Generalized Zero-Shot Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6): 2506–2520.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *Advances in Neural Information Processing Systems*, 2121–2129.

Fu, Y.; Hospedales, T. M.; Xiang, T.; and Gong, S. 2015. Transductive Multi-View Zero-Shot Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11): 2332–2345.

Gao, R.; Hou, X.; Qin, J.; Chen, J.; Liu, L.; Zhu, F.; Zhang, Z.; and Shao, L. 2020. Zero-VAE-GAN: Generating Unseen Features for Generalized and Transductive Zero-Shot Learning. *IEEE Transactions Image Processing*, 29: 3665–3680.

Ge, J.; Xie, H.; Min, S.; and Zhang, Y. 2021. Semantic-guided Reinforced Region Embedding for Generalized Zero-Shot Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1406–1414.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, 2672–2680.

Han, Z.; Fu, Z.; and Yang, J. 2020. Learning the Redundancy-Free Features for Generalized Zero-Shot Object Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12862–12871.

Huynh, D.; and Elhamifar, E. 2020. Fine-Grained Generalized Zero-Shot Learning via Dense Attribute-Based Attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4482–4492.

Jia, Z.; Zhang, Z.; Wang, L.; Shan, C.; and Tan, T. 2020. Deep Unbiased Embedding Transfer for Zero-Shot Learning. *IEEE Transactions on Image Processing*, 29: 1958–1971.

Keshari, R.; Singh, R.; and Vatsa, M. 2020. Generalized Zero-Shot Learning via Over-Complete Distribution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 13297–13305.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *Proceedings of the International Conference on Learning Representations (Poster)*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 951–958.

Larochelle, H.; Erhan, D.; and Bengio, Y. 2008. Zero-data Learning of New Tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 646–651.

Li, J.; Jing, M.; Lu, K.; Ding, Z.; Zhu, L.; and Huang, Z. 2019. Leveraging the Invariant Side of Generative Zero-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7402–7411.

Li, J.; Jing, M.; Zhu, L.; Ding, Z.; Lu, K.; and Yang, Y. 2020. Learning Modality-Invariant Latent Representations for Generalized Zero-shot Learning. In *Proceedings of the ACM International Conference on Multimedia*, 1348–1356.

Li, X.; Xu, Z.; Wei, K.; and Deng, C. 2021. Generalized Zero-Shot Learning via Disentangled Representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1966–1974.

Liu, Y.; Zhou, L.; Bai, X.; Huang, Y.; Gu, L.; Zhou, J.; and Harada, T. 2021. Goal-Oriented Gaze Estimation for Zero-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3794–3803.

Ma, P.; and Hu, X. 2020. A Variational Autoencoder with Deep Embedding Model for Generalized Zero-Shot Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11733–11740.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In Bengio, Y.; and LeCun, Y., eds., *International Conference on Learning Representations, Workshop Track Proceedings*.

Mishra, A.; Reddy, M. S. K.; Mittal, A.; and Murthy, H. A. 2018. A Generative Model for Zero Shot Learning Using Conditional Variational Autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2188–2196.

Ni, J.; Zhang, S.; and Xie, H. 2019. Dual Adversarial Semantics-Consistent Network for Generalized Zero-Shot

Learning. In *Advances in Neural Information Processing Systems*, 6143–6154.

Niu, L.; Cai, J.; Veeraraghavan, A.; and Zhang, L. 2019. Zero-Shot Learning via Category-Specific Visual-Semantic Mapping and Label Refinement. *IEEE Transactions on Image Processing*, 28(2): 965–979.

Patterson, G.; and Hays, J. 2012. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2751–2758.

Rahman, S.; Khan, S.; and Porikli, F. 2018. A Unified Approach for Conventional Zero-Shot, Generalized Zero-Shot, and Few-Shot Learning. *IEEE Transactions on Image Processing*, 27(11): 5652–5667.

Saito, K.; Kim, D.; Sclaroff, S.; Darrell, T.; and Saenko, K. 2019. Semi-Supervised Domain Adaptation via Minimax Entropy. In *Proceedings of the IEEE International Conference on Computer Vision*, 8049–8057.

Schönfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; and Akata, Z. 2019. Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8247–8255.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H. S.; and Hospedales, T. M. 2018. Learning to Compare: Relation Network for Few-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1199–1208.

Tsai, Y. H.; Huang, L.; and Salakhutdinov, R. 2017. Learning Robust Visual-Semantic Embeddings. In *Proceedings of IEEE International Conference on Computer Vision*, 3591–3600.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Verma, V. K.; Arora, G.; Mishra, A.; and Rai, P. 2018. Generalized Zero-Shot Learning via Synthesized Examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4281–4289.

Vyas, M. R.; Venkateswara, H.; and Panchanathan, S. 2020. Leveraging Seen and Unseen Semantic Relationships for Generative Zero-Shot Learning. In *Proceedings of the European Conference on Computer Vision*, 70–86.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wan, Z.; Chen, D.; Li, Y.; Yan, X.; Zhang, J.; Yu, Y.; and Liao, J. 2019. Transductive Zero-Shot Learning with Visual Structure Constraint. In *Advances in Neural Information Processing Systems*, 9972–9982.

Wang, W.; Pu, Y.; Verma, V. K.; Fan, K.; Zhang, Y.; Chen, C.; Rai, P.; and Carin, L. 2018. Zero-Shot Learning via Class-Conditioned Deep Generative Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4211–4218.

Wei, K.; Deng, C.; and Yang, X. 2020. Lifelong Zero-Shot Learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 551–557.

Xian, Y.; Akata, Z.; Sharma, G.; Nguyen, Q. N.; Hein, M.; and Schiele, B. 2016. Latent Embeddings for Zero-Shot Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 69–77.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2018a. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9): 2251–2265.

Xian, Y.; Lorenz, T.; Schiele, B.; and Akata, Z. 2018b. Feature Generating Networks for Zero-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5542–5551.

Xian, Y.; Sharma, S.; Schiele, B.; and Akata, Z. 2019. F-VAEGAN-D2: A Feature Generating Framework for Any-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10275–10284.

Xie, G.; Liu, L.; Jin, X.; Zhu, F.; Zhang, Z.; Qin, J.; Yao, Y.; and Shao, L. 2019. Attentive Region Embedding Network for Zero-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9384–9393.

Xie, G.; Liu, L.; Zhu, F.; Zhao, F.; Zhang, Z.; Yao, Y.; Qin, J.; and Shao, L. 2020. Region Graph Embedding Network for Zero-Shot Learning. In *Proceedings of the European Conference on Computer Vision*, 562–580.

Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; and Akata, Z. 2020. Attribute Prototype Network for Zero-Shot Learning. In *Advances in Neural Information Processing Systems*.

Yu, H.; and Lee, B. 2019. Zero-shot Learning via Simultaneous Generating and Learning. In *Advances in Neural Information Processing Systems*, 46–56.

Yu, Y.; Ji, Z.; Han, J.; and Zhang, Z. 2020. Episode-Based Prototype Generating Network for Zero-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 14032–14041.

Yue, Z.; Wang, T.; Sun, Q.; Hua, X.; and Zhang, H. 2021. Counterfactual Zero-Shot and Open-Set Visual Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 15404–15414.

Zhu, P.; Wang, H.; and Saligrama, V. 2019. Generalized Zero-Shot Recognition Based on Visually Semantic Embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2995–3003.