Noname manuscript No.
(will be inserted by the editor)

Information Bottleneck and Selective Noise Supervision for Zero-Shot Learning

Lei Zhou[1,*] · Yang Liu[1,*] · Xiao Bai[1] · Pengcheng Zhang[1] · Lin Gu[2,3] · Jun Zhou[4] · Yazhou Yao[5] · Jin Zheng[1] · Tatsuya Harada[3,2] · Edwin Hancock[6]

Received: date / Accepted: date

Abstract Zero-shot learning (ZSL) aims to recognize novel classes by transferring semantic knowledge from seen classes to unseen classes. Though many ZSL methods rely on a direct mapping between the visual and the semantic space, the calibration deviation and hubness problem limit the generalization capability to unseen classes. Recently emerged generative ZSL methods generate unseen image features to transform ZSL into a supervised classification problem. However, most generative models still suffer from the seen-unseen bias problem as only seen data is used for training. To address these issues, we propose a novel bidirectional embedding based generative model with a tight visual-semantic coupling constraint. We learn a unified latent space that calibrates the embedded parametric distributions of both visual and semantic spaces. Since the embedding from high-dimensional visual features comprises much non-semantic information, the alignment of visual and semantic in latent space would inevitably be deviated. Therefore, we introduce an information bottleneck (IB) constraint to ZSL for the first time to preserve essential attribute information during the mapping. Specifically, we utilize the uncertainty estimation and the wake-sleep procedure to alleviate the feature noises and improve model abstraction capability. In addition, our method can be easily ex-

* Lei Zhou and Yang Liu contribute equally to this work.
✉ Corresponding author: Xiao Bai
E-mail: baixiao@buaa.edu.cn

[1] School of Computer Science and Engineering, State Key Laboratory of Software Development Environment, Jiangxi Research Institute, Beihang University, Beijing, China
[2] RIKEN AIP, Tokyo, Japan
[3] The University of Tokyo, Tokyo, Japan
[4] School of Information and Communication Technology, Griffith University, Australia
[5] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[6] Department of Computer Science, University of York, York, U.K.

tended to the transductive ZSL setting by generating labels for unseen images. We then introduce a robust self-training loss to solve this label-noise problem. Extensive experimental results show that our method outperforms the state-of-the-art methods in different ZSL settings on most benchmark datasets.

Keywords Zero-shot learning · Information bottleneck · Uncertainty estimation · Label-noise learning · Transductive ZSL

1 Introduction

Thanks to the abundant human annotated data, deep learning has achieved great success in many supervised learning problems, such as image classification and retrieval (Krizhevsky et al., 2017; Zhou et al., 2019, 2020), object detection (Ren et al., 2015; Wang et al., 2018a), semantic segmentation (Long et al., 2015; Chen et al., 2017). However, labeling large-scale training data for each task is time consuming and expensive. Inspired by the human's remarkable ability in recognizing instances of unseen classes solely based on class descriptions without seeing any visual example of such classes, zero-shot learning (ZSL) was proposed as an image classification setting to mimic the human learning process (Lampert et al., 2009). Given the semantic descriptions of both seen and unseen classes but only the seen class training images, ZSL aims to classify test images of unseen classes.

Based on the images and labels that a model can see in the training phase, ZSL includes two settings which are Inductive ZSL(IZSL) and Transductive ZSL(TZSL). IZSL can only utilize the images and labels of seen classes during training. TZSL can use extra images of unseen classes without labels during training. In the test phase, the ZSL problem is further categorized into two settings: conventional ZSL and generalized ZSL. In conventional ZSL, the images to be recognized at test time belong only to unseen classes. In the generalized ZSL (GZSL) setting, the images at test time may belong to both seen or unseen classes. The GSZL setting is practically more useful and challenging, since the assumption that images at test time come only from unseen classes need not hold.

Most early ZSL methods learn a direct or indirect mapping between the visual space and the semantic space (Akata et al., 2013; Romera-Paredes and Torr, 2015; Akata et al., 2015b; Xian et al., 2016; Guo et al., 2016; Kodirov et al., 2017; Xie et al., 2019, 2020; Liu et al., 2021). However, the performance of these methods is often poor on GZSL setting. The reason is that the embedding model is learned only from seen classes, which leads to a bias towards predicting seen classes. To address this issue, more recent approaches (Xian et al., 2018b; Mishra et al., 2018; Wang et al., 2018b; Li et al., 2019a; Schonfeld et al., 2019; Ma and Hu, 2020; Verma et al., 2020; Yu et al., 2020) utilize generative models, e.g., generative adversarial networks (GAN) (Goodfellow et al., 2014) or variational autoencoders (VAE) (Kingma and Welling, 2013), to generate synthetic features of unseen classes. This transfers the ZSL task to

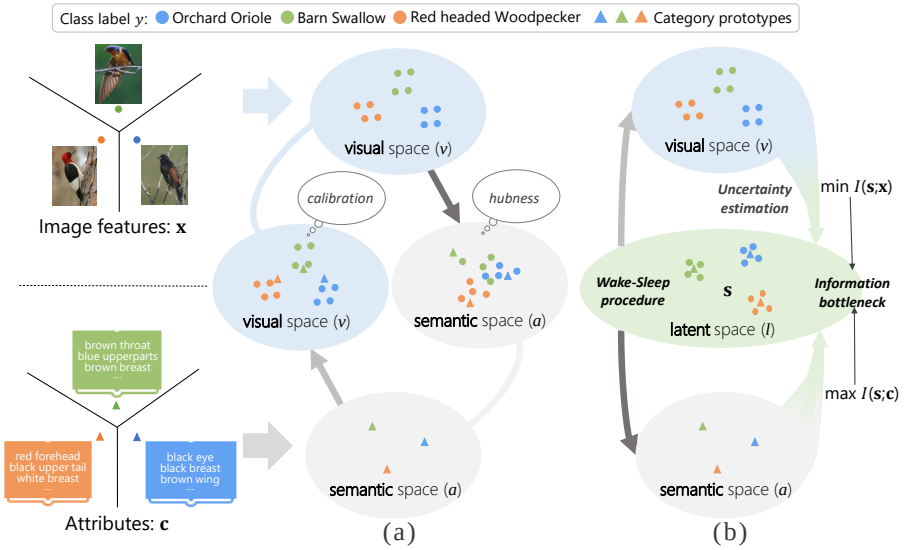


Fig. 1 Comparison of existing direct mapping methods and our latent bidirectional embedding based generative model. (a) Traditional ZSL frameworks are based on direct mapping. The hubness problem and calibration deviation make it difficult to accurately align visual and semantic distributions in respective spaces. (b) Our bidirectional embedding based generative model with a unified latent space. Firstly, an information bottleneck constraint on the latent bidirectional embedding preserves more essential attribute information while eliminating the non-semantic information of visual features. Secondly, uncertainty estimation is utilized to alleviate the visual noises and a bias passing mechanism is designed to solve the unicity of human annotated semantics. Thirdly, a wake-sleep procedure uses both real and generated data for joint training to improve the model representation and abstraction capability.

a supervised classification problem. Since GAN-based loss functions are unstable in training, VAE-based methods (Schonfeld et al., 2019; Ma and Hu, 2020) were developed to tackle this problem and achieved better performance. However, most of these generative models still suffer from the deviation between generated features and unseen classes due to the lack of tight visual-semantic coupling.

Since high-dimensional visual features contain non-semantic information which is redundant for semantic discrimination (Tong et al., 2019; Han et al., 2020b; Shen et al., 2020), it is difficult to well align the semantic categories to the centers of visual sample distributions when mapping the semantic features to the visual space. This causes a calibration deviation problem as illustrated in Figure 1. In addition, when high-dimensional visual features are mapped to a low-dimensional semantic space, the shrink of feature space would aggravate the hubness problem that some instances in the high-dimensional space become the nearest neighbors of a large number of instances (Radovanovic et al., 2010). To address the above challenges, we propose an information bottleneck (IB) (Tishby et al., 2000) constrained bidirectional embedding based generative model which utilizes advantages of both embedding model and

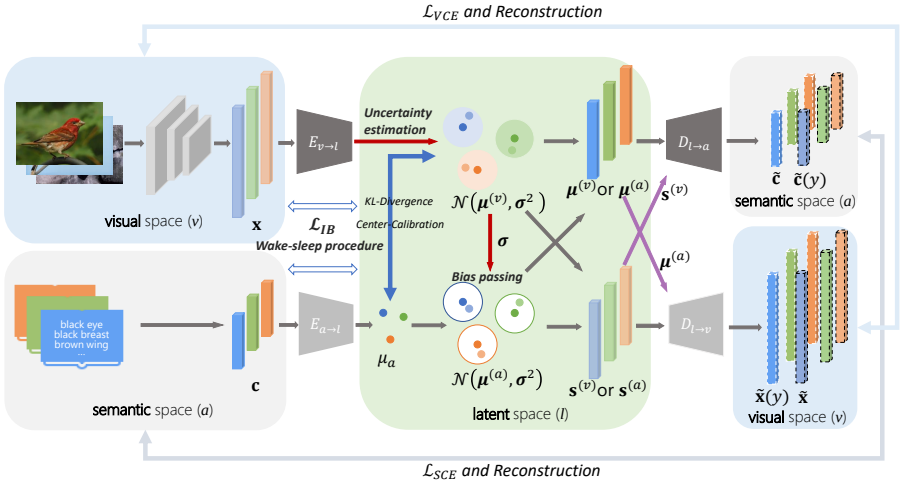


Fig. 2 Illustration of the proposed model. We learn a latent bidirectional embedding based generative model via a modified VAE network. A unified latent space is simultaneously learned to align visual and semantic distributions. A novel Information Bottleneck (IB) loss ($\mathcal{L}_{IB}$) is proposed as the constraint between the latent space and the other two spaces. We exploit the data uncertainty estimation to learn the bias ($\boldsymbol{\sigma}$) of the original visual data and share the bias from visual distribution to the semantic distribution (bias passing). In addition, a wake-sleep procedure is used for joint training of real data and generated data.

generative model to align visual and semantic distributions in a unified latent space. As shown in Figure 2, our proposed method first learns a latent bidirectional embedding via a modified VAE network. Visual features $\mathbf{x}$ and attributes semantic $\mathbf{c}$ are embedded to the unified latent space respectively by two encoders. Then the visual-semantic distributions alignment is constrained in the latent space by *KL-Divergence* and *Center-Calibration*. To facilitate the distribution alignment, the redundant non-semantic information in the visual space should be discarded to preserve the attributed related part when it is flowing to the latent space. To achieve this, we design an **IB loss ($\mathcal{L}_{IB}$)** on the latent bidirectional embedding to impose the mutual information relationships between feature spaces. Due to the wide existence of noises such as the labeling noise (Kunran Xu and Gu, 2020), the human annotated semantics are insufficient to fully describe the diversified visual samples (Ding and Liu, 2019). The deviation between visual and semantic distributions will accumulate during the embedding process. Therefore, we learn the bias of the original visual distribution by introducing an **uncertainty estimation** technique (Kendall and Gal, 2017) to alleviate the influence of noises. Since one semantic class may correspond to a variety of visual samples, we also propose a **bias passing** mechanism to share this variety bias to the latent semantic distribution to benefit the distributions alignment. The two decoders are respectively used to generate attributes semantic and visual features for auxiliary training, i.e., the reconstruction losses $\mathcal{L}_{VCE}$ and $\mathcal{L}_{SCE}$. Since VAE does not incorporate the generated samples for learning, the latent features generated by VAE are

largely randomized and uncontrollable (Hu et al., 2017, 2018). Therefore, we introduce a **wake-sleep procedure** (Hinton et al., 1995) that uses both real and generated data for joint training to improve the model representation and abstraction capability. Furthermore, base on the proposed framework, we can easily extend our method to solve the TZSL task. We use the generated semantics as pseudo labels for unseen images and regard the problem as a label noise circumstance. Then a robust loss is introduced to solve this label noise problem. Finally, with the generated latent features, we can solve both the IZSL and TZSL as a supervised classification problem.

The contributions of this paper are as follows.

- We propose a novel ZSL method based on an information bottleneck (IB) constrained generative model with a tight visual-semantic bidirectional embedding. The IB loss minimizes the non-semantic information when embedding the visual domain to latent space. To the best of our knowledge, this is the first work that adopts the IB theory in ZSL.
- We exploit the data uncertainty estimation technique for the first time in ZSL to learn the bias of visual distribution and design a bias passing mechanism, which alleviates the noises and gap between visual features and human annotated semantics.
- We train the model on both real and generated data with a wake-sleep training mechanism to improve the model representation and abstraction capability via a VAE model.
- We further extend our method to adapt the transductive ZSL setting with a robust label noise loss.

Extensive experimental results on four widely used ZSL benchmarks for both generalized ZSL and conventional ZSL show the superiority of our method under both inductive ZSL and transductive ZSL settings compared with state-of-the-art ZSL methods.

The rest of this paper is organized as follows: We first review the related works on two different ZSL settings in Section 2. In Section 3, we describe the problem setting and our proposed method. The extended method for TZSL is introduced at the end of Section 3. Section 4 shows the experimental results on four benchmark datasets under different ZSL settings. Finally, we conclude our method in Section 5.

2 Related Works

In this section, we review related works on IZSL and TZSL. For the IZSL, we divide the existing methods into two categories, embedding models and generative models.

2.1 Embedding Models for IZSL

Embedding models for ZSL focus on learning a direct or indirect mapping between visual and semantic spaces to transfer semantic knowledge from seen classes to unseen classes. There are three typical embedding strategies. The earliest methods learned the mapping function from the visual space to the semantic space, which include, for example DAP and IAP (Lampert et al., 2013), ALE (Akata et al., 2015a), DeViSE (Frome et al., 2013) and ESZSL (Romera-Paredes and Torr, 2015). To alleviate the severe hubness problem caused by embedding from the high-dimensional visual space to the low-dimensional semantic space, reverse mapping from the semantic space to the visual space was proposed for the nearest neighbor classification in the visual space (Changpinyo et al., 2017; Zhang et al., 2017). Some models such as SSE (Zhang and Saligrama, 2015), SYNC (Changpinyo et al., 2016) and BiDiLEL (Wang and Chen, 2017) explore the idea of embedding both visual and semantic features into a common intermediate space. Though these methods perform well in the conventional ZSL setting, their performance deteriorates on the GZSL setting since there are only seen class features for model training.

2.2 Generative Models for IZSL

Recently, abundant generative models (Guo et al., 2017; Chen et al., 2018; Felix et al., 2018; Kumar Verma et al., 2018; Xian et al., 2018b; Zhu et al., 2018; Li et al., 2019a; Schonfeld et al., 2019; Ma and Hu, 2020; Keshari et al., 2020) were proposed to address the training data imbalance problem between seen and unseen classes by synthesizing unseen class features. Among these, both generative adversarial networks (GAN) (Chen et al., 2018; Felix et al., 2018; Zhu et al., 2018; Li et al., 2019a), and variational autoencoders (VAE) (Bucher et al., 2017; Kumar Verma et al., 2018; Schonfeld et al., 2019; Ma and Hu, 2020; Keshari et al., 2020) have been used for ZSL. f-CLSWGAN (Xian et al., 2018b) adapts the Wasserstein GAN (WGAN) (Arjovsky et al., 2017; Gulrajani et al., 2017) by adding a classification loss to enforce the generator to synthesize features that are suited for ZSL. Motivated by the cycle consistency loss (Zhu et al., 2017), cycle-CLSWGAN (Felix et al., 2018) utilizes a multi-modal cycle consistency loss to enforce that the generated visual features map back to their original semantic features, which can generate more robust unseen samples. LisGAN (Li et al., 2019a) exploits conditional WGAN to generate fake unseen classes from random noises and introduces soul samples regularizations to guarantee the generated sample is close to real. Due to the hardness of training GAN based models, CADA-VAE (Schonfeld et al., 2019) adopts a cross-aligned VAE to align the visual and semantic distributions in a latent space. More recently, a new flow-based generative model (Shen et al., 2020) was introduced to ZSL which utilizes an invertible generative flow network to generate distinguishable samples.

Although these generative models have achieved encouraging performance for GZSL, feature generation for unseen classes still needs tight visual-semantic coupling constraints to alleviate the deviation. Our proposed method combines the advantages of both embedding and generative models for an accurate alignment of visual-semantic distributions while generating discriminative image features.

2.3 Transductive ZSL

Different from the IZSL, the TZSL assumes the availability of unlabeled target unseen images (Fu et al., 2015) during training. But the relationship between the unseen image and label is still unavailable. With unseen images, the distribution discrepancy between seen and unseen domains can be exploited to alleviate the domain shift problem. Therefore, the TZSL methods usually achieve better performance than the IZSL setting. QFSL (Song et al., 2018) alleviates the bias problem with a proposed Quasi-Fully Supervised Learning framework. GXE (Li et al., 2019b) utilizes semantic attributes to train a visual feature classifier and calibrates the classifier with unlabeled data. SABR (Paul et al., 2019) learns two different GANs to generate the latent space features for seen and unseen classes, respectively. Recently, SDGN (Wu et al., 2020) integrates a self-supervised learning mechanism into the feature generating model to effectively exploit the unlabeled data and uses the self-supervision lurking in the data structure of different domains to conduct cross-domain mining.

Our method can be easily extended to adapt the TZSL setting. The latent space we learned can effectively eliminate the possible noise in visual and semantic features. We select unseen data with higher confidence and mark these data with pseudo labels. Unlike training the seen classes, the labels of the unseen classes are noisy, hence we introduce a robust loss for the label noise problem during the unseen classes training.

3 Proposed Method

In this section, we first define the problem setting, notations and then present the details of each module of our method. Finally, we extend our method to adapt the TZSL setting with a robust loss for the noisy label.

3.1 Problem Setting and Notations

The GZSL problem is defined as follows. Let $\mathcal{X}^S$ and $\mathcal{X}^U$ denote the image feature sets of seen classes and unseen classes respectively, $\mathcal{X} = \mathcal{X}^S \cup \mathcal{X}^U$. $\mathcal{S} = \{(\mathbf{x}, y, \mathbf{c}(y))|\mathbf{x} \in \mathcal{X}^S, y \in \mathcal{Y}^S, \mathbf{c}(y) \in \mathcal{C}^S\}$ denotes the training set, where $\mathbf{x} \in \mathbb{R}^D$ are image features extracted by a plain CNN model. y are the seen class labels which are available during training and $\mathbf{c}(y) \in \mathbb{R}^K$ are attribute features. The auxiliary training set is $\mathcal{U} = \{(u, \mathbf{c}(u))|u \in \mathcal{Y}^U, \mathbf{c}(u) \in \mathcal{C}^U\}$, where u

denote unseen class labels. The seen classes and unseen classes are disjoint, i.e., $\mathcal{Y}^S \cap \mathcal{Y}^U = \emptyset$. Here, $\mathcal{C} = \mathcal{C}^S \cup \mathcal{C}^U$ is used to transfer information between seen classes and unseen classes. $\mathcal{C}$ can be human-annotated attributes (Xian et al., 2018b) or articles describing the classes (Zhu et al., 2018). In the conventional ZSL, the task is to learn a classifier $f_{ZSL} : \mathcal{X}^U \rightarrow \mathcal{Y}^U$. However, in more realistic and challenging setup of GZSL, the aim is to learn a classifier $f_{GZSL} : \mathcal{X} \rightarrow \mathcal{Y}^U \cup \mathcal{Y}^S$.

The architecture of the proposed model is shown in Figure 2. It consists of two sets of latent embedding VAEs with $(E_{v\rightarrow l}, D_{l\rightarrow a})$ and $(E_{a\rightarrow l}, D_{l\rightarrow v})$. These two sets of VAEs share the same latent space l. $E_{v\rightarrow l}$ maps visual space v to latent space l, and $D_{l\rightarrow a}$ maps latent space l to semantic space a. $E_{a\rightarrow l}$ maps semantic space a to latent space l, and $D_{l\rightarrow v}$ maps latent space l to visual space v. The visual-semantic distributions are aligned in the latent space l by KL-Divergence and Center-Calibration. Information bottleneck loss and uncertainty estimation are used to facilitate the distribution alignment. A wake-sleep procedure is exploited to improve the quality of generated features by VAE decoder. Following we give the detailed descriptions of each module.

3.2 Latent Bidirectional Embedding with Uncertainty Estimation

The goal of our model is to learn a latent space that can accurately align visual and semantic distributions. We first learn a Visual to Semantic (VS) network $\text{VS} = E_{v\rightarrow l} \circ D_{l\rightarrow a} : \mathbb{R}^D \rightarrow \mathbb{R}^K$ to project the visual features through latent space into semantic space. The latent embedding model is shown in Figure 3. Because there may be inherent noise in the visual features (Chang et al., 2020). To reduce the impact of data uncertainty, we define the latent representation $\mathbf{z}_i^{(v)}$ embedded from each visual sample $\mathbf{x}_i$ as a Gaussian distribution:

$$p(\mathbf{z}_i^{(v)}|\mathbf{x}_i) = \mathcal{N}(\mathbf{z}_i^{(v)}; \boldsymbol{\mu}_i^{(v)}, \boldsymbol{\sigma}_i^2\mathbf{I}), \tag{1}$$

where $\boldsymbol{\mu}_i^{(v)}$ and $\boldsymbol{\sigma}_i^2$ are the mean and variance of the Gaussian distribution learned by the encoder $E_{v\rightarrow l} : \boldsymbol{\mu}_i^{(v)} = E_{v\rightarrow l,\boldsymbol{\phi}_1}(\mathbf{x}_i), \log \boldsymbol{\sigma}_i^2 = E_{v\rightarrow l,\boldsymbol{\phi}_2}(\mathbf{x}_i)$, where $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ refer to the model parameters. The re-parameterization trick (Kingma and Welling, 2013) is used to keep gradients of the model as usual. With this trick, we generate the latent sampling representation $\mathbf{s}_i^{(v)}$ as

$$\mathbf{s}_i^{(v)} = \boldsymbol{\mu}_i^{(v)} + \boldsymbol{\epsilon}\boldsymbol{\sigma}_i, \ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

where $\boldsymbol{\epsilon}$ is a random noise.

Then, $\widetilde{\mathbf{c}}(y_i) = D_{l\rightarrow a}(\mathbf{s}_i^{(v)})$ projects the latent feature $\mathbf{s}_i^{(v)}$ into semantic space, i.e., the mapping of a visual sample $\mathbf{x}_i$ is calculated as $\text{VS}(\mathbf{x}_i)$:

$$\text{VS}(\mathbf{x}_i) = \widetilde{\mathbf{c}}(y_i) = D_{l\rightarrow a}(\mathbf{s}_i^{(v)}) = E_{v\rightarrow l} \circ D_{l\rightarrow a}(\mathbf{x}_i). \tag{3}$$

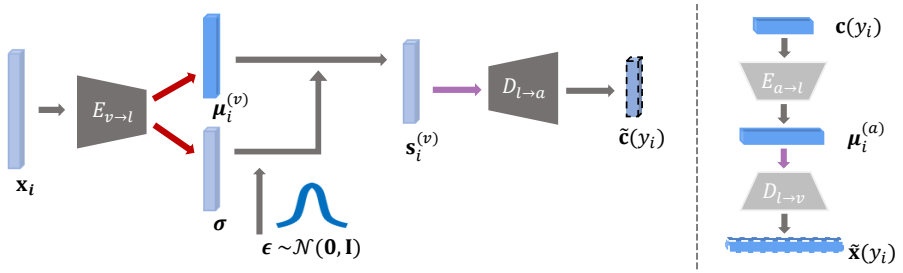


Fig. 3 Illustration of the latent bidirectional embedding and data uncertainty estimation. Left: the visual to semantic network. Here we define the latent representation as a Gaussian distribution to learn the data uncertainty. Right: the semantic to visual network.

The affinity between $\mathrm{VS}(\mathbf{x}_i)$ and the y_i-th attribute feature $\mathbf{c}(y_i)$ could be measured by their inner product $\mathrm{VS}(\mathbf{x}_i)^T\mathbf{c}(y_i)$. Then the probability of $\mathbf{x}_i$ belong to the y_i-th category in semantic space can be calculated as:

$$p^A(y_i|\mathbf{x}_i) = \frac{\exp(\mathrm{VS}(\mathbf{x}_i)^T\mathbf{c}(y_i))}{\sum_{y\in\mathcal{Y}^S}\exp(\mathrm{VS}(\mathbf{x}_i)^T\mathbf{c}(y))}. \tag{4}$$

Then the Semantic Cross-Entropy (SCE) loss can be written as:

$$\mathcal{L}_{SCE} = -\sum_i \log p^A(y_i|\mathbf{x}_i). \tag{5}$$

Similarly, we learn a Semantic to Visual (SV) network $\mathrm{SV} = E_{a\to l}\circ D_{l\to v} : \mathbb{R}^K \to \mathbb{R}^D$, which first projects the semantic feature $\mathbf{c}(y_i)$ to latent space as $\boldsymbol{\mu}_i^{(a)}$, then projects $\boldsymbol{\mu}_i^{(a)}$ to visual space as the generated visual prototype $\widetilde{\mathbf{x}}(y_i)$ for the y_i-th category:

$$\mathrm{SV}(\mathbf{c}(y_i)) = \widetilde{\mathbf{x}}(y_i) = D_{l\to v}(\boldsymbol{\mu}_i^{(a)}) = E_{a\to l}\circ D_{l\to v}(\mathbf{c}(y_i)). \tag{6}$$

The probability of $\mathbf{x}_i$ belong to the y_i-th category in visual space is calculated as:

$$p^V(y_i|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T\mathrm{SV}(\mathbf{c}(y_i)))}{\sum_{y\in\mathcal{Y}^S}\exp(\mathbf{x}_i^T\mathrm{SV}(\mathbf{c}(y)))}. \tag{7}$$

Then the Visual Cross-Entropy (VCE) loss is:

$$\mathcal{L}_{VCE} = -\sum_i \log p^V(y_i|\mathbf{x}_i). \tag{8}$$

The total Cross-Entropy (CE) loss is as follow:

$$\mathcal{L}_{CE} = \mathcal{L}_{SCE} + \mathcal{L}_{VCE}. \tag{9}$$

In order to learn an accurate latent bidirectional embedding, we perform center calibration for each category. Such a structured objective requires the

center embedding of $\mathbf{x}_i$ being closer to the latent embedding of its groundtruth $\mathbf{c}(y_i)$ than other classes, the Center Calibration (CC) is defined as:

$$\mathcal{L}_{CC} = \sum_{i,y}[\Delta + d(E_{v\to l,\phi_1}(\mathbf{x}_i), E_{a\to l}(\mathbf{c}(y))) - d(E_{v\to l,\phi_1}(\mathbf{x}_i), E_{a\to l}(\mathbf{c}(y_i)))]_+, \tag{10}$$

where $d(\cdot,\cdot)$ denotes a certain distance metric. Here, we utilize the Euclidean distance in the experiments. $\Delta > 0$ is a margin to make $\mathcal{L}_{CC}$ more robust.

3.3 Feature Generation with Noise Supervision

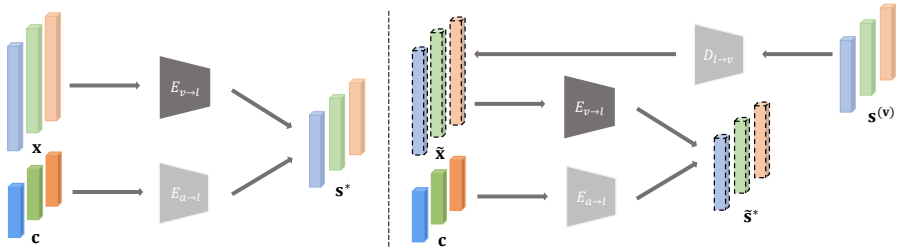
For each category y, there could be many visual samples $\mathbf{x}$, but the semantic description $\mathbf{c}$ of each category is unique. Thus, this unique semantic attribute $\mathbf{c}$ is insufficient to fully describe the variety of visual samples. Therefore, we assume the latent semantic distribution similar to the Gaussian distribution of latent visual features in Equation (1). To adapt to this task, we use two sets of encoder-decoder structures. $E_{v\to l}$ encodes the visual features $\mathbf{x}_i$ to a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_i^{(v)}, \boldsymbol{\sigma}_i^2)$ in the latent space, and $E_{a\to l}$ encodes the semantic feature $\mathbf{c}(y_i)$ to the center $\boldsymbol{\mu}_i^{(a)}$ of category y_i. Since the latent semantic Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_i^{(a)}, \boldsymbol{\sigma}_i^2)$ should be consistent with the latent visual distribution, we design a bias passing mechanism to share the noise (bias) from the visual diversity for the latent semantic distribution. Then we use the decoders $D_{l\to a}$ to decode $\boldsymbol{\mu}_i^{(v)}$ or $\boldsymbol{\mu}_i^{(a)}$ to semantic feature $\widetilde{\mathbf{c}}(y_i)$, and use $D_{l\to v}$ to decode $\mathbf{s}_i^{(v)} \sim \mathcal{N}(\boldsymbol{\mu}_i^{(v)}, \boldsymbol{\sigma}_i^2)$ or $\mathbf{s}_i^{(a)} \sim \mathcal{N}(\boldsymbol{\mu}_i^{(a)}, \boldsymbol{\sigma}_i^2)$ to visual feature $\widetilde{\mathbf{x}}_i$. Finally, the loss with noise supervision for the modified VAE can be written as:

$$\begin{aligned}\mathcal{L}_{VAE} = \mathbb{E}_{q_\phi(\mathbf{s}^{(v)}|\mathbf{x})}[\log p_{\theta_1}(\mathbf{c}|\mathbf{s}^{(v)})] + \mathbb{E}_{q_\phi(\mathbf{s}^{(a)}|\mathbf{c})}[\log p_{\theta_1}(\mathbf{c}|\mathbf{s}^{(a)})] \\ + \mathbb{E}_{q_\phi(\mathbf{s}^{(v)}|\mathbf{x})}[\log p_{\theta_2}(\mathbf{c}|\mathbf{s}^{(v)})] + \mathbb{E}_{q_\phi(\mathbf{s}^{(a)}|\mathbf{c})}[\log p_{\theta_2}(\mathbf{c}|\mathbf{s}^{(a)})] \\ - \beta D_{KL}(q_\phi(\mathbf{s}^{(v)}|\mathbf{x})||\mathcal{N}(\boldsymbol{\mu}^{(a)}, \mathbf{I})),\end{aligned} \tag{11}$$

where, $\boldsymbol{\phi}$ refers to the parameters of $E_{v\to l}$ and $E_{a\to l}$, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ refer to the parameters of $D_{l\to a}$ and $D_{l\to v}$, respectively. The hyperparameter β is set following CADA-VAE (Schonfeld et al., 2019).

3.4 Information Bottleneck Constraint

In our method, information is gradually disentangled from the visual space through the latent space to the semantic space. The semantic feature $\mathbf{c}$ is related disentangled attribute information while the visual feature $\mathbf{x}$ has high-dimensional entangled non-semantic information. Therefore, we hope that the latent feature $\mathbf{s}$ should contain as much semantic information of $\mathbf{c}$ as possible while discarding the redundant non-semantic information of $\mathbf{x}$. In information theory, the dependence between two random variables could be measured by



(a) **wake phase** (b) **sleep phase**

Fig. 4 The wake-sleep procedure. The features represented as dotted lines are generated. In the wake phase, we use real visual data $\mathbf{x}$ to train the model representation capability. In the sleep phase, the generated visual data $\widetilde{\mathbf{x}}$ is used to train the model abstraction capability.

mutual information $I(\cdot;\cdot)$. As illustrated in Figure 1, we maximize the mutual information between the semantic space and the latent space ($I(\mathbf{s};\mathbf{c})$) and minimize the mutual information between the visual space and the latent space ($I(\mathbf{s};\mathbf{x})$). We define the information bottleneck (IB) (Tishby et al., 2000) to constrain the information relationship between spaces:

$$\max I(\mathbf{s};\mathbf{c}) - \eta I(\mathbf{s};\mathbf{x}). \tag{12}$$

Since $\mathbf{s}$ may be sampled from different distributions like $\mathcal{N}(\boldsymbol{\mu}^{(v)}, \boldsymbol{\sigma}^2)$ or $\mathcal{N}(\boldsymbol{\mu}^{(a)}, \boldsymbol{\sigma}^2)$ in our model, we resample $\mathbf{s}^* \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\sigma}^2)$, where $\boldsymbol{\mu}^* = \alpha\boldsymbol{\mu}^{(v)} + (1-\alpha)\boldsymbol{\mu}^{(a)}$ with a uniform distribution $\alpha \sim U(0,1)$.

Since the VAE model does not utilize the generated samples for training, the latent features generated are largely randomized and uncontrollable. Inspired by the work of Hu *et al.* (Hu et al., 2018), we train the modified VAE model in a wake-sleep procedure, using real data and generated data for joint training. The extended wake-sleep procedure is shown in Figure 4. In the wake phase, we use real visual data $\mathbf{x}$ to train the feature representation capability of the model. In the sleep phase, we use generated data $\widetilde{\mathbf{x}}$ to train the abstraction capability of the model. So that the model can generate disentangled latent features. The wake-sleep information bottleneck constraint is as follow:

$$\max[I(\mathbf{s}^*;\mathbf{c}) - \eta I(\mathbf{s}^*;\mathbf{x})] + \lambda[I(\widetilde{\mathbf{s}}^*;\mathbf{c}) - \eta I(\widetilde{\mathbf{s}}^*;\widetilde{\mathbf{x}})], \tag{13}$$

where $\widetilde{\mathbf{s}}^*$ is the latent embedding representation of $\widetilde{\mathbf{x}}$. The weighting factor λ is obtained by grid search on the validation set.

Since the information bottleneck with high dimension is intractable to calculate, we follow the strategy proposed by Alemi *et al.* (Alemi et al., 2016). The Information Bottleneck (IB) loss is shown as follow:

$$\mathcal{L}_{IB} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[-\log q(\mathbf{c}(y_i)|D_{l\to a}(\mathbf{s}_i^*))] + \eta D_{KL}[q_\phi(\mathbf{s}^*|\mathbf{x}_i, \mathbf{c}(y_i))||r(\mathbf{z})], \tag{14}$$

where $r(\mathbf{z})$ is a standard normal distribution in the experiments. η is initialized to 10^{-5} and changed with the IB loss.

Finally, the overall loss of the proposed model is defined as:

$$\mathcal{L} = \mathcal{L}_{VAE} + \gamma\mathcal{L}_{CE} + \delta\mathcal{L}_{CC} + \tau\mathcal{L}_{IB}, \tag{15}$$

where γ, δ, and τ are the weighting factors of the cross entropy loss, center calibration, and information bottleneck loss, respectively. We empirically choose their values to balance the effect of different loss terms in the experiments.

For more intuitive understanding, we summarize our proposed method in Algorithm 1. When the encoder $E_{v\to l}$ is finished training, we utilize this encoder to learn latent representations of both the seen and unseen image features. Then the latent representations are used for classification.

Algorithm 1: The proposed zero-shot learning method.

Input: The image features $\mathbf{x}$, class labels y and attribute features $\mathbf{c}(y)$ of seen classes; The attribute features $\mathbf{c}(u)$ of unseen classes.
Output: The parameters of encoder $E_{v\to l}$.
Steps:
1. Learn the Visual to Semantic (VS) VAE VS $= E_{v\to l} \circ D_{l\to a}$ to project the visual features through latent space into semantic space by the $\mathcal{L}_{SCE}$ Loss (5).
2. Learn the Semantic to Visual (SV) VAE SV $= E_{a\to l} \circ D_{l\to v}$ to project the semantic features through latent space into visual space by the $\mathcal{L}_{VCE}$ Loss (8).
3. A Center-Calibration $\mathcal{L}_{CC}$ Loss (10) is used to align the visual and semantic distributions in the latent space.
4. Generate features with noise supervision by bias passing, the VAE loss with noise supervision is $\mathcal{L}_{VAE}$ Loss (11).
5. An information bottleneck constraint $\mathcal{L}_{IB}$ Loss (14) is utilized to discard the redundant non-semantic information of visual features while preserving more essential information of semantic features.

3.5 Transductive ZSL

Our proposed model can be easily extended to the TZSL setting. We adopt the self-training manner to exploit the unseen images. We first generate the pseudo labels $\widetilde{\mathcal{Y}}^U$ for unseen image features $\mathcal{X}^U$ through calculating classification score $\mathbf{O}$ of the learned classifier in the IZSL stage. There will inevitably be noise in $\widetilde{\mathcal{Y}}^U$. We use classification score peakiness based filtering strategy (Li et al., 2019b) to mitigate the influence of noisy labels. The classification score of $\mathbf{x}_i^u \in \mathcal{X}^U$ is $\mathbf{o}^i \in \mathbb{R}^{N^U}$ according to all the N^U classes. The sum of each dimension of $\mathbf{o}^i$ is 1. Let $\mathbf{o}^i$ be the soft label of $\mathbf{x}_i^u$. We assume the maximum and second maximum score of $\mathbf{o}^i$ are $\mathbf{o}^i_{u_m}$ and $\mathbf{o}^i_{u_n}$. The pseudo label of $\mathbf{x}_i^u$ should be u_m. If an unseen sample satisfies $\dfrac{\mathbf{o}^i_{u_m}}{\mathbf{o}^i_{u_n}} > r$, we assign soft label and pseudo label to it. r is the threshold, we set $r = 1.4$ in our experiment. Then we have a high confident training set $\widetilde{\mathcal{U}} = \{(\mathbf{x}^u, \widetilde{u}, \mathbf{c}(\widetilde{u}), \mathbf{o}) | \mathbf{x}^u \in \mathcal{X}^U, \widetilde{u} \in \mathcal{Y}^U, \mathbf{c}(\widetilde{u}) \in \mathcal{C}^U, \mathbf{o} \in \mathbf{O}\}$.

For the seen classes, we still utilize the proposed method for the IZSL to train the model with training set $\mathcal{S}$. The difference is that we use the cross entropy loss to replace center calibration of Equation (10). Simultaneously, we exploit the training set $\widetilde{\mathcal{U}}$ to calibrate the model with unseen classes. Unlike Equation (5) and Equation (8), we use the soft label and soft cross-entropy loss to replace hard label and cross-entropy loss in the visual and semantic space. In the latent space, the probability $\mathbf{x}_i^u$ belong to the $\widetilde{u}_i$-th category is calculated as:

$$p(\widetilde{u}_i|\mathbf{x}_i^u) = \frac{\exp(E_{v\to l,\phi_1}(\mathbf{x}_i^u)^T E_{a\to l}(\mathbf{c}(\widetilde{u}_i)))}{\sum_{u\in\mathcal{Y}^U}\exp(E_{v\to l,\phi_1}(\mathbf{x}_i^u)^T E_{a\to l}(\mathbf{c}(u)))}. \tag{16}$$

Then, we introduce the Generalized Cross-Entropy (GCE) loss (Zhang and Sabuncu, 2018) to alleviate the influence of noisy pseudo labels:

$$\mathcal{L}_{GCE} = \sum_i \frac{1 - p(\widetilde{u}_i|\mathbf{x}_i^u)^q}{q}, \tag{17}$$

where $q \in (0, 1]$ is a hyper-parameter of which a higher value is preferred when the noise level is high. The GCE loss is a generalization of the Categorical Cross-Entropy (CCE) loss and the Mean Absolute Error (MAE). It is equivalent to the CCE loss when q infinitely approaches 0 and turns to MAE loss when $q = 1$. We set $q = 0.4$ in our experiment. The CCE loss is powerful for classification tasks but will be overfitting on the label noise dataset. The MAE loss is worse for label clean classification task but is robust to noisy labels. The hyper-parameter q can be tuned between 0 and 1 to fit different noise levels. In the TZSL experiments, we have also tried the other label-noise learning methods T-Revision (Xia et al., 2019) and SIGUA (Han et al., 2020a) to combat noisy labels. Results show that these different label-noise learning methods achieve similar performance. This may be because that the soft label assignment procedure is more important for the transductive ZSL. Therefore, we choose the compact GCE loss to handle the label-noise problem in our method.

4 Experiments

In this section, we first give the experimental settings. Then the comparison with state-of-the-art methods and ablation studies are conducted to demonstrate the effectiveness of our method. Finally, more analyzing experiments show the superiority of our method.

4.1 Experimental Settings

Datasets. We evaluated our framework on four widely used benchmark datasets including CUB-200-2011 (CUB) (Welinder et al., 2010), SUN attribute (SUN) (Patterson and Hays, 2012), attributes Pascal and Yahoo (aPY) (Farhadi et al.,

Table 1 The details for SUN, CUB, AwA2 and APY. $\mathcal{Y}^S$ and $\mathcal{Y}^U$ are seen class number and unseen class number. Tr, Val, Ts are the number of images at training, validation and test time.

Datasets	Attributes	$\lvert\mathcal{Y}^S\rvert$	$\lvert\mathcal{Y}^U\rvert$	Tr	Val	Ts
SUN	102	645	72	10320	2580	1440
CUB	312	150	50	7057	1764	2967
AwA2	85	40	10	23527	5882	7913
aPY	64	20	12	5932	1483	7924

2009) and Animals with Attributes 2 (AwA2) (Xian et al., 2018a) for the GZSL. We extracted a 2,048-dimensional CNN features for images using ResNet-101 (He et al., 2016) as the visual features. The pre-defined attributes on each dataset were used as the semantic descriptors. Moreover, we adopted the Proposed Split (PS) (Xian et al., 2018a) to divide all classes into seen and unseen classes on each dataset. The dataset details are listed in Table 1.

Implementation details. In our modified VAE model, we utilized multilayer perceptrons to implement the encoders ($E_{v\to l}$ and $E_{a\to l}$) and decoders ($D_{l\to a}$ and $D_{l\to v}$). The encoders $E_{v\to l}$ and $E_{a\to l}$ had 1560 and 1450 hidden units, respectively. The hidden units of $D_{l\to a}$ and $D_{l\to v}$ were 665 and 1660, respectively. The latent embedding dimensions were 64 for AwA2 and aPY and 256 for CUB and SUN. β, λ, γ, δ and τ were set to 0.5, 0.1, 1.0, 0.1 and 1.0. The margin Δ was set to 10^{-3}. Adam optimizer (Kingma and Ba, 2014) was used for training, the epoch size was 120 and the batch size was 64. After the model training, the encoders $E_{v\to l}$ and $E_{a\to l}$ transformed the visual features of seen classes and attribute features of unseen classes into the unified latent space. Finally, we trained a softmax classifier to classify latent features.

Evaluation metrics. The performance of our method is evaluated by per-class Top-1 accuracy. In conventional ZSL, the Top-1 accuracy is evaluated on seen classes, denoted as **T1**. In GZSL, since the test set is composed of seen and unseen images, the Top-1 accuracy is evaluated respectively on seen classes, denoted as **S**, and unseen classes, denoted as **U**. Their harmonic mean (**H**) (Xian et al., 2018a) is used to evaluate the performance of GZSL, which is defined as:

$$\mathbf{H} = \frac{2 \times \mathbf{S} \times \mathbf{U}}{\mathbf{S} + \mathbf{U}}. \tag{18}$$

4.2 Comparison with Baseline Methods

Inductive ZSL. We selected several state-of-the-art GZSL methods for comparison, which include non-feature generation methods ALE (Akata et al., 2013), DeViSE (Frome et al., 2013), ESZSL (Romera-Paredes and Torr, 2015), SJE (Akata et al., 2015b), LATEM (Xian et al., 2016), SYNC (Changpinyo et al., 2016), SAE (Kodirov et al., 2017), SP-AEN (Chen et al., 2018), TCN (Jiang et al., 2019), TripletLoss (Cacheux et al., 2019) and feature generation based

Table 2 Results of the state-of-the-arts generalized zero-shot learning for inductive setting. The best and the second-best results are respectively marked by red and blue.

Methods	CUB			AwA2			SUN			aPY		
	U	**S**	**H**	**U**	**S**	**H**	**U**	**S**	**H**	**U**	**S**	**H**
ALE	23.7	62.8	34.3	14.0	81.8	23.9	21.8	33.1	26.3	4.6	73.3	8.7
DeViSE	23.8	53.0	32.8	17.1	74.7	27.8	16.9	27.4	20.9	4.9	76.9	9.2
ESZSL	12.6	63.8	21.0	5.9	77.8	11.0	11.0	27.9	15.8	2.4	70.1	4.6
SJE	23.5	59.2	33.6	8.0	73.9	14.4	14.7	30.5	19.8	3.7	55.7	6.9
LATEM	15.2	57.3	24.0	11.5	77.3	20.0	14.7	28.8	19.5	0.1	73.0	0.2
SYNC	11.5	70.9	19.8	10.0	90.5	18.0	7.9	43.3	13.4	7.4	66.3	13.3
SAE	7.8	54.0	13.6	1.1	82.2	2.2	8.8	18.0	11.8	0.4	80.9	0.9
SP-AEN	34.7	70.6	46.6	23.3	90.0	37.1	24.9	38.6	30.0	13.7	63.4	22.6
TCN	52.6	52.0	52.3	61.2	65.8	63.4	31.2	37.3	34.0	24.1	64.0	35.1
TripletLoss	55.8	52.3	53.0	48.5	83.2	61.3	47.9	30.4	36.8	-	-	-
SE-GZSL	41.5	53.3	46.7	58.3	68.1	62.8	40.9	30.5	34.9	-	-	-
CVAE-ZSL	-	-	34.5	-	-	51.2	-	-	26.7	-	-	-
f-CLSWGAN	43.7	57.7	49.7	53.8	68.2	60.2	42.6	36.6	39.4	-	-	-
LisGAN	46.5	57.9	51.6	54.3	68.5	60.6	42.9	37.8	40.2	34.3	68.2	45.7
GDAN	39.3	66.7	49.5	32.1	67.5	43.5	38.1	89.9	53.4	30.4	75.0	43.4
CADA-VAE	51.6	53.5	52.4	55.8	75.0	63.9	47.2	35.7	40.6	-	-	-
ABP	47.0	54.8	50.6	55.3	72.6	62.6	45.3	36.8	40.6	-	-	-
OCD-CVAE	44.8	55.9	51.3	59.5	73.4	65.7	44.8	42.9	43.8	-	-	-
LsrGAN	48.1	59.1	53.0	-	-	-	44.8	47.7	40.9	-	-	-
Ours	52.2	56.2	54.1	56.0	80.0	65.9	43.8	37.8	40.6	34.2	69.8	45.9

methods SE-GZSL (Kumar Verma et al., 2018), CVAE-ZSL (Mishra et al., 2018), f-CLSWGAN (Xian et al., 2018b), LisGAN (Li et al., 2019a), GDAN (Huang et al., 2019) CADA-VAE (Schonfeld et al., 2019), ABP (Zhu et al., 2019), OCD-CVAE (Keshari et al., 2020), LsrGAN (Vyas et al., 2020). Table 2 shows the results of different methods on four datasets. It can be seen that our proposed method outperforms all the ten compared non-feature generation methods with a large margin for the harmonic mean results. Moreover, our method significantly improves the Top-1 accuracy on unseen classes benefited from the generated unseen class samples. Compared with the feature generation based methods, our method can also achieve the best harmonic mean results on CUB, AwA2, and aPY. Since the IB constrained bidirectional embedding between the visual space and the semantic space can preserve essential attribute information and discard the non-semantic information. To further demonstrate the effectiveness of our method. We also compared our method under the conventional ZSL setting that the test image only belongs to unseen classes. As shown in Table 3, our proposed method achieves the best for three out of the four datasets.

Transductive ZSL. Under the TZSL setting, we mitigated the deviation of the model on seen classes by utilizing the data of unseen classes during training. We compared our method with recent state-of-the-art TZSL methods, which include ALE-trans (Akata et al., 2015b), GFZSL-trans (Verma and Rai, 2017), QFSL (Song et al., 2018), GXE (Li et al., 2019b), GMN (Sariyildiz

Table 3 Results of conventional zero-shot learning for inductive setting. The best and the second-best results are respectively marked by red and blue.

Methods	**CUB**	**AwA2**	**SUN**	**aPY**
ALE	54.9	62.5	58.1	39.7
DeViSE	52.0	59.7	56.5	39.8
ESZSL	53.9	58.6	54.5	38.3
SJE	53.9	61.9	53.7	32.9
LATEM	49.3	55.8	55.3	35.2
SYNC	55.6	46.6	56.3	23.9
SAE	33.3	54.1	40.3	8.3
TCN	59.5	71.2	61.5	38.9
SE-GZSL	59.6	69.2	63.4	-
CVAE-ZSL	52.1	65.8	61.7	-
f-CLSWGAN	57.3	-	60.8	-
LisGAN	58.8	-	61.7	43.1
ABP	58.5	70.4	61.5	-
OCD-CVAE	60.3	71.3	63.5	-
LsrGAN	60.3	-	62.5	-
Ours	62.2	70.1	64.2	43.5

Table 4 Results of the generalized zero-shot learning for transductive setting. The best and the second-best results are respectively marked by red and blue.

Methods	**CUB**			**AwA2**			**SUN**		
	U	**S**	**H**	**U**	**S**	**H**	**U**	**S**	**H**
ALE-trans	23.5	45.1	30.9	12.6	73.0	21.5	19.9	22.6	21.2
GFZSL-trans	24.9	45.8	32.2	31.7	67.2	43.1	0.0	41.6	0.0
QFSL	17.3	39.0	24.0	20.8	74.7	32.6	17.7	25.0	20.7
GXE	57.0	68.7	62.3	80.2	90.0	84.8	45.4	58.1	51.0
GMN	60.2	70.6	65.0	-	-	-	57.1	40.7	47.5
f-VAEGAN	61.4	65.1	63.2	84.8	88.6	86.7	60.6	41.9	49.6
WDVSc	43.3	85.4	57.5	76.4	88.1	81.8	-	-	-
Ours-trans	65.3	66.5	65.9	82.7	89.2	85.8	57.5	44.6	50.2

and Cinbis, 2019), f-VAEGAN (Xian et al., 2019), WDVSc (Wan et al., 2019). Table 4 and Table 5 show the results of different methods on CUB, AwA2, and SUN for GZSL and conventional ZSL, respectively. We can see that our method outperforms all the compared methods for both GZSL and conventional ZSL on the CUB dataset. For AwA2 and SUN datasets, our method achieves the second highest performance. This validates that our method can be readily adapted to the TZSL setting.

4.3 Further Analyses for Inductive Setting

Ablation study. We conducted ablation experiments to verify the effectiveness of the proposed modules. Table 6 shows the influence of different losses.

Table 5 Results of conventional zero-shot learning for transductive setting. The best and the second-best results are respectively marked by red and blue.

Methods	**CUB**	**AwA2**	**SUN**
ALE-trans	54.5	70.7	55.7
GFZSL-trans	49.3	78.6	64.0
QFSL	72.1	79.7	58.3
GXE	61.3	83.2	63.5
GMN	64.6	-	64.3
f-VAEGAN	71.7	89.8	70.1
WDVSc	73.4	87.3	63.4
Ours-trans	73.5	88.1	67.6

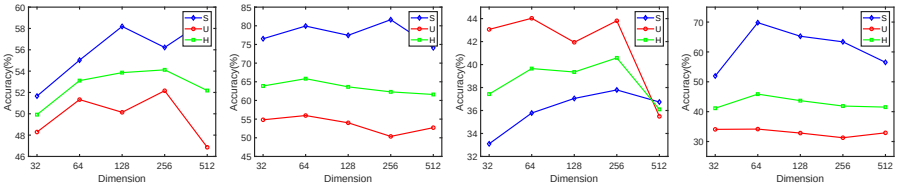
Table 6 Ablation study of the proposed modules.

Loss				**CUB**			**AwA2**			**SUN**			**aPY**		
$\mathcal{L}_{VAE}$	$\mathcal{L}_{CE}$	$\mathcal{L}_{CC}$	$\mathcal{L}_{IB}$	**U**	**S**	**H**	**U**	**S**	**H**	**U**	**S**	**H**	**U**	**S**	**H**
✔	✔			46.6	**56.9**	51.2	52.8	77.9	62.9	40.1	35.7	37.7	32.7	61.8	42.8
✔	✔	✔		50.8	56.6	53.5	54.2	77.0	63.6	**43.8**	36.9	40.1	32.1	66.7	43.3
✔	✔		✔	51.0	56.3	53.5	53.2	**80.7**	64.1	41.8	**37.8**	39.7	**34.7**	65.4	45.3
✔	✔	✔	✔	**52.2**	56.2	**54.1**	**56.0**	80.0	**65.9**	**43.8**	**37.8**	**40.6**	34.2	**69.8**	**45.9**

Table 7 Ablation study of wake-sleep IB constraint on CUB dataset. Wake phase only uses real data for training. Sleep(S/U) phase uses generated seen/unseen classes data for training.

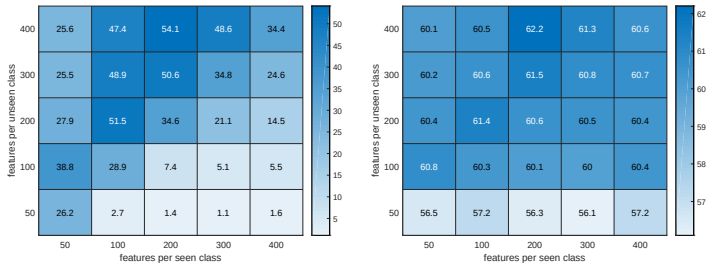
Wake	Sleep(S)	Sleep(U)	**T1**	**U**	**S**	**H**
✔			60.5	50.0	55.4	52.5
✔	✔		61.9	48.7	**58.5**	53.1
✔		✔	60.9	51.6	54.8	53.2
✔	✔	✔	**62.2**	**52.2**	56.2	**54.1**

We can see that our proposed method achieves the best harmonic mean results with all the losses. Specifically, the proposed IB loss can significantly improve the performance. For the proposed wake-sleep IB constraint, we also performed ablation study with different conditions on CUB, as shown in Table 7. It can be seen that the IB loss constrained on the generated seen classes features (Sleep(S)) has significantly improved the classification accuracy of the seen classes and conventional ZSL. Accordingly, the IB loss constrained on the generated unseen classes features (Sleep(U)) also improves the result of unseen classes. Our method achieves the highest harmonic mean result under the wake-sleep IB constraint. In addition, we use $\mathcal{N}(\boldsymbol{\mu}^{(a)}, \mathbf{I})$ to replace the latent semantic distribution $\mathcal{N}(\boldsymbol{\mu}^{(a)}, \boldsymbol{\sigma}^2)$ to verify the effectiveness of the proposed bias passing mechanism. The results on four datasets are CUB($\mathbf{H}$=53.6),



(a) CUB (b) AwA2 (c) SUN (d) aPY

Fig. 5 The influence of the dimension of latent features generated by our model. We measured the Top-1 accuracy on seen classes and unseen classes and the harmonic mean accuracy on CUB, AwA2, SUN, and aPY datasets.



(a) GZSL (H) (b) Conventional ZSL (T1)

Fig. 6 The influence of different numbers of generated features per seen and unseen classes.

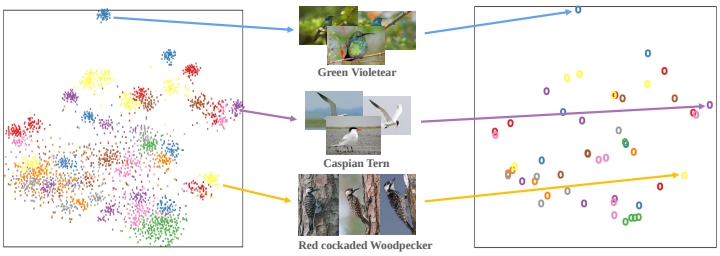
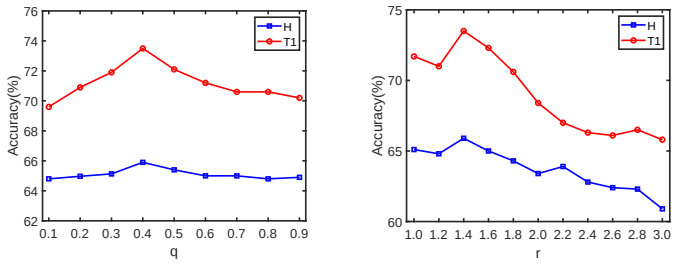


Fig. 7 Visualization of the latent feature distributions. The top is the latent visual embedding features and the bottom is the latent semantic embedding features.

AwA2(**H**=64.3), SUN(**H**=39.5) and aPY(**H**=45.5). It shows that the bias passing mechanism can alleviate the visual and semantic noises problem.

The influence of latent dimensions. We first evaluated our method with different dimensions of latent features, as shown in Figure 5. The harmonic mean results have less fluctuation with different latent feature dimensions on four datasets. Our method achieves the best performance on CUB and SUN with the latent feature dimensions equal to 256. The best results are reached for AwA2 and aPY when the latent feature dimension is 64. We speculate the reason is that on the one hand the CUB and SUN are fine-grained datasets which need more information to distinguish. On the other hand, excessive dimensions lead to redundant information.



(a) The influence of q (b) The influence of r

Fig. 8 The influence of q and r on GZSL (**H**) and conventional ZSL (**T1**) results (%) for transductive setting.

The influence of latent features. We show the influence of different numbers of generated features per seen and unseen classes on CUB. Figure 6 reports the harmonic mean results (H) of GZSL (a) and T1 results of conventional ZSL (b). We can see that the GZSL performance of our method increases with more generated unseen features in most cases and when the number of generated unseen features is twice the generated seen features, our method achieves the best result. The results of conventional ZSL show a similar conclusion with GZSL, which validates the soundness of our method.

Visualization Result. We use the t-SNE (Maaten and Hinton, 2008) to visualize our latent features used for the final GZSL classification. Figure 7 shows the distributions of the latent features of 50 classes on the CUB dataset. The top is the latent visual embedding features and the bottom is the latent semantic embedding features. From the almost consistent distribution, we can see our latent features can well align visual and semantic distributions.

4.4 Further Analyses for Transductive Setting

The influence of q and r. Figure 8(a) shows the results of **H** and **T1** for CUB dataset when varying hyper-parameter q of generalized cross-entropy loss from 0.1 to 0.9. Our method achieves the best performance when $q = 0.4$. GCE loss combines the advantages of CCE loss and MAE loss, which makes our model more robust in the classification task with noisy labels. The influence of r on the results of **H** and **T1** for CUB dataset is shown in Figure 8(b). We varied the value of r from 1.0 to 3.0 in steps of 0.2. At the first, the accuracy of **H** and **T1** increases with the change of r. Our model achieves the best result when $r = 1.4$. Then the performance of the model declines with the change of r. Since the quality of the noisy unseen data used for training becomes higher, the performance of the model is increasing. While the higher the quality of the unseen data, the less noise in the label, and the smaller the amount of training data. Therefore the performance of the model decreases.

Quality analysis of noisy data. We further analyzed the influence of the threshold r on the generated pseudo labels for CUB, AwA2, and SUN.

Table 8 Quality analysis of noisy data. *Acc* is the accuracy of the pseudo label. *Num* is the number of noise data for training. *Max*, *Min* and *Avg* respectively show the maximum, minimum and average number of training samples for different categories.

r		1.0	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6	2.8	3.0
CUB	*Acc*	0.61	0.66	0.70	0.75	0.78	0.81	0.83	0.84	0.85	0.87	0.88
	Num	2967	2420	2065	1786	1573	1404	1276	1162	1077	994	928
	Max	60	59	59	59	59	59	59	58	58	57	56
	Min	50	29	19	15	7	5	3	3	1	0	0
	Avg	59	48	41	35	31	28	25	23	21	19	18
AwA2	*Acc*	0.61	0.66	0.7	0.72	0.74	0.75	0.77	0.78	0.79	0.8	0.82
	Num	7913	6080	4741	3736	2976	2341	1844	1481	1193	979	832
	Max	1645	1354	1106	893	699	515	427	389	345	309	281
	Min	174	151	100	71	49	30	19	12	10	9	9
	Avg	791	608	474	373	297	234	184	148	119	97	83
SUN	*Acc*	0.63	0.67	0.70	0.73	0.75	0.76	0.78	0.79	0.81	0.82	0.83
	Num	1440	1267	1136	1040	964	900	837	792	752	720	688
	Max	20	20	20	20	20	20	20	20	20	19	19
	Min	20	12	9	6	5	4	4	4	2	2	1
	Avg	20	17	15	14	13	12	11	11	10	10	9

As shown in Table 8, *Acc* is the accuracy of the generated pseudo label with different thresholds r. It shows that a larger threshold r can make the confidence of the pseudo label higher. *Num* illustrates the relationship between the number of noisy data for training and threshold r. Although the confidence of pseudo labels has become higher, the number of samples used for training is decreasing. *Max*, *Min* and *Avg* respectively show the maximum, minimum and average number of training samples for different categories with different r. We can see that when r is large, the sample numbers for different categories will be unbalanced, especially the number of training samples in some categories will even become 0. Therefore, we set $r = 1.4$ to achieve a good tradeoff in sample quality, sample training number, and sample balance in the experiment.

5 Conclusion

In this paper, we introduced a novel bidirectional embedding based generative model for zero-shot learning. This method learns a unified latent space to align the feature distributions of both visual domain and semantic domain. A novel information bottleneck (IB) constrained latent bidirectional embedding allows the latent features to contain more essential attributes related information while discarding non-semantic information flowed from the visual features. In addition, data uncertainty estimation and wake-sleep procedure are introduced to facilitate latent distributions alignment. The proposed method has outperformed several state-of-the-art methods in different ZSL settings in experimental comparison, showing the advantages of our approach. Furthermore,

our method can be lightly extended to adapt the transductive ZSL task and also achieves competitive performance.

For the future works, further investigation on information bottleneck and uncertainty estimation theories for the cross-domain alignment problem is significant. For example, the other feature generation based ZSL models, such as GANs and generative flows, can also utilize information bottleneck and uncertainty estimation methods to generate more reliable unseen samples.

Declarations

Funding (This work was supported by the National Natural Science Foundation of China project no. 61772057, Beijing Natural Science Foundation (4202039), the support funding Jiangxi Research Institute of Beihang University. And supported by the Academic Excellence Foundation of BUAA for PhD Students.)

Conflicts of interests (The authors declare that they have no conflict of interest.)

Ethics approval (Not Applicable)

Consent to participate (Not Applicable)

Consent for publication (Not Applicable)

Availability of data and material (The data used in this work is all public.)

Code availability (The codes of the proposed method will be released after publishing.)

Authors' contributions (All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Lei Zhou, Yang Liu, Pengcheng Zhang, Xiao Bai, Lin Gu, Jun Zhou, Yazhou Yao, Tatsuya Harada, Jin Zheng and Edwin Hancock. The first draft of the manuscript was written by Lei Zhou and Yang Liu. And all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.)

References

Akata Z, Perronnin F, Harchaoui Z, Schmid C (2013) Label-embedding for attribute-based classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 819–826

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

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

Alemi AA, Fischer I, Dillon JV, Murphy K (2016) Deep variational information bottleneck. arXiv preprint arXiv:161200410

Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. In: International Conference on Machine Learning, pp 214–223

Bucher M, Herbin S, Jurie F (2017) Generating visual representations for zero-shot classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2666–2673

Cacheux YL, Borgne HL, Crucianu M (2019) Modeling inter and intra-class relations in the triplet loss for zero-shot learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp 10333–10342

Chang J, Lan Z, Cheng C, Wei Y (2020) Data uncertainty learning in face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5710–5719

Changpinyo S, Chao WL, Gong B, Sha F (2016) Synthesized classifiers for zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5327–5336

Changpinyo S, Chao WL, Sha F (2017) Predicting visual exemplars of unseen classes for zero-shot learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3476–3485

Chen L, Zhang H, Xiao J, Liu W, Chang SF (2018) Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1043–1052

Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence 40(4):834–848

Ding Z, Liu H (2019) Marginalized latent semantic encoder for zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6191–6199

Farhadi A, Endres I, Hoiem D, Forsyth D (2009) Describing objects by their attributes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1778–1785

Felix R, Kumar VB, Reid I, Carneiro G (2018) Multi-modal cycle-consistent generalized zero-shot learning. In: European Conference on Computer Vision, pp 21–37

Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Ranzato M, Mikolov T (2013) Devise: A deep visual-semantic embedding model. In: Advances in neural information processing systems, pp 2121–2129

Fu Y, Hospedales T, XIANG T, Gong S (2015) Transductive multi-view zero-shot learning. IEEE Transactions on Pattern Analysis and Machine Intelligence

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial networks. In: Advances in neural information processing systems, pp 2672–2680

Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC (2017) Improved training of wasserstein gans. In: Advances in neural information processing systems, pp 5767–5777

Guo Y, Ding G, Jin X, Wang J (2016) Transductive zero-shot recognition via shared model space learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 3494–3500

Guo Y, Ding G, Han J, Gao Y (2017) Synthesizing samples for zero-shot learning. In: International Joint Conference on Artificial Intelligence, pp 1774–1780

Han B, Niu G, Yu X, Yao Q, Xu M, Tsang I, Sugiyama M (2020a) Sigua: Forgetting may make learning with noisy labels more robust. In: International Conference on Machine Learning, PMLR, pp 4006–4016

Han Z, Fu Z, Yang J (2020b) Learning the redundancy-free features for generalized zero-shot object recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 12865–12874

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

Hinton GE, Dayan P, Frey BJ, Neal RM (1995) The" wake-sleep" algorithm for unsupervised neural networks. Science 268(5214):1158–1161

Hu Z, Yang Z, Liang X, Salakhutdinov R, Xing EP (2017) Toward controlled generation of text. In: International Conference on Machine Learning, pp 1587–1596

Hu Z, Yang Z, Salakhutdinov R, Xing EP (2018) On unifying deep generative models. In: International Conference on Learning Representations

Huang H, Wang C, Yu PS, Wang CD (2019) Generative dual adversarial network for generalized zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 801–810

Jiang H, Wang R, Shan S, Chen X (2019) Transferable contrastive network for generalized zero-shot learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp 9765–9774

Kendall A, Gal Y (2017) What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in neural information processing systems, pp 5574–5584

Keshari R, Singh R, Vatsa M (2020) Generalized zero-shot learning via overcomplete distribution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 13300–13308

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980

Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint arXiv:13126114

Kodirov E, Xiang T, Gong S (2017) Semantic autoencoder for zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3174–3183

Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Communications of the ACM 60(6):84–90

Kumar Verma V, Arora G, Mishra A, Rai P (2018) Generalized zero-shot learning via synthesized examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4281–4289

Kunran Xu YL Lai Rui, Gu L (2020) Feature normalized knowledge distillation for image classification. In: European Conference on Computer Vision

Lampert CH, Nickisch H, Harmeling S (2009) Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 951–958

Lampert CH, Nickisch H, Harmeling S (2013) Attribute-based classification for zero-shot visual object categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(3):453–465

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

Li K, Min MR, Fu Y (2019b) Rethinking zero-shot learning: A conditional visual classification perspective. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3583–3592

Liu Y, Zhou L, Bai X, Huang Y, Gu L, Zhou J, Harada T (2021) Goal-oriented gaze estimation for zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3794–3803

Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440

Ma P, Hu X (2020) A variational autoencoder with deep embedding model for generalized zero-shot learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 11733–11740

Maaten Lvd, Hinton G (2008) Visualizing data using t-sne. Journal of Machine Learning Research 9(Nov):2579–2605

Mishra A, Krishna Reddy S, Mittal A, Murthy HA (2018) A generative model for zero shot learning using conditional variational autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop, pp 2188–2196

Patterson G, Hays J (2012) Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2751–2758

Paul A, Krishnan NC, Munjal P (2019) Semantically aligned bias reducing zero shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7056–7065

Radovanovic M, Nanopoulos A, Ivanovic M (2010) Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of Machine Learning Research 11(sept):2487–2531

Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99

Romera-Paredes B, Torr P (2015) An embarrassingly simple approach to zero-shot learning. In: International Conference on Machine Learning, pp 2152–2161

Sariyildiz MB, Cinbis RG (2019) Gradient matching generative networks for zero-shot learning. In: Proceedings of the IEEE Conference on Computer

Vision and Pattern Recognition, pp 2168–2178
Schonfeld E, Ebrahimi S, Sinha S, Darrell T, Akata Z (2019) Generalized zero- and few-shot learning via aligned variational autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8247–8255
Shen Y, Qin J, Huang L, Liu L, Zhu F, Shao L (2020) Invertible zero-shot recognition flows. In: European Conference on Computer Vision, Springer, pp 614–631
Song J, Shen C, Yang Y, Liu Y, Song M (2018) Transductive unbiased embedding for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1024–1033
Tishby N, Pereira FC, Bialek W (2000) The information bottleneck method. arXiv preprint physics/0004057
Tong B, Wang C, Klinkigt M, Kobayashi Y, Nonaka Y (2019) Hierarchical disentanglement of discriminative latent features for zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 11467–11476
Verma VK, Rai P (2017) A simple exponential family framework for zero-shot learning. In: Joint European conference on machine learning and knowledge discovery in databases, Springer, pp 792–808
Verma VK, Brahma D, Rai P (2020) Meta-learning for generalized zero-shot learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 6062–6069
Vyas MR, Venkateswara H, Panchanathan S (2020) Leveraging seen and unseen semantic relationships for generative zero-shot learning. In: European Conference on Computer Vision, Springer, pp 70–86
Wan Z, Chen D, Li Y, Yan X, Zhang J, Yu Y, Liao J (2019) Transductive zero-shot learning with visual structure constraint. In: Advances in Neural Information Processing Systems, pp 9972–9982
Wang C, Bai X, Wang S, Zhou J, Ren P (2018a) Multiscale visual attention networks for object detection in vhr remote sensing images. IEEE Geoscience and Remote Sensing Letters 16(2):310–314
Wang Q, Chen K (2017) Zero-shot visual recognition via bidirectional latent embedding. International Journal of Computer Vision 124(3):356–383
Wang W, Pu Y, Verma VK, Fan K, Zhang Y, Chen C, Rai P, Carin L (2018b) Zero-shot learning via class-conditioned deep generative models. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 4211–4218
Welinder P, Branson S, Mita T, Wah C, Schroff F, Belongie S, Perona P (2010) Caltech-ucsd birds 200
Wu J, Zhang T, Zha ZJ, Luo J, Zhang Y, Wu F (2020) Self-supervised domain-aware generative network for generalized zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 12767–12776
Xia X, Liu T, Wang N, Han B, Gong C, Niu G, Sugiyama M (2019) Are anchor points really indispensable in label-noise learning? Advances in Neural Information Processing Systems 32:6838–6849

Xian Y, Akata Z, Sharma G, Nguyen Q, Hein M, Schiele B (2016) Latent embeddings for zero-shot classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 69–77

Xian Y, Lampert CH, Schiele B, 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, Akata Z (2018b) Feature generating networks for zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5542–5551

Xian Y, Sharma S, Schiele B, 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, pp 10275–10284

Xie GS, Liu L, Jin X, Zhu F, Zhang Z, Qin J, Yao Y, Shao L (2019) Attentive region embedding network for zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9384–9393

Xie GS, Liu L, Zhu F, Zhao F, Zhang Z, Yao Y, Qin J, Shao L (2020) Region graph embedding network for zero-shot learning. In: European Conference on Computer Vision, pp 562–580

Yu Y, Ji Z, Han J, Zhang Z (2020) Episode-based prototype generating network for zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14035–14044

Zhang L, Xiang T, Gong S (2017) Learning a deep embedding model for zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2021–2030

Zhang Z, Sabuncu M (2018) Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems 31:8778–8788

Zhang Z, Saligrama V (2015) Zero-shot learning via semantic similarity embedding. In: Proceedings of the IEEE International Conference on Computer Vision, pp 4166–4174

Zhou L, Xiao B, Liu X, Zhou J, Hancock ER, et al. (2019) Latent distribution preserving deep subspace clustering. In: 28th International Joint Conference on Artificial Intelligence, pp 4440–4446

Zhou L, Bai X, Liu X, Zhou J, Hancock ER (2020) Learning binary code for fast nearest subspace search. Pattern Recognition 98:107040

Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232

Zhu Y, Elhoseiny M, Liu B, Peng X, Elgammal A (2018) A generative adversarial approach for zero-shot learning from noisy texts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1004–1013

Zhu Y, Xie J, Liu B, Elgammal A (2019) Learning feature-to-feature translator by alternating back-propagation for generative zero-shot learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp

9844–9854