Hindawi
Complexity
Volume 2021, Article ID 6656797, 11 pages
https://doi.org/10.1155/2021/6656797

Research Article

Dual Generative Network with Discriminative Information for Generalized Zero-Shot Learning

Tingting Xu, Ye Zhao, and Xueliang Liu

School of Computer and Information, Hefei University of Technology, Hefei 230000, China

Correspondence should be addressed to Ye Zhao; zhaoye@hfut.edu.cn

Received 17 December 2020; Revised 15 January 2021; Accepted 20 February 2021; Published 28 February 2021

Academic Editor: Chenquan Gan

Zero-shot learning is dedicated to solving the classification problem of unseen categories, while generalized zero-shot learning aims to classify the samples selected from both seen classes and unseen classes, in which "seen" and "unseen" classes indicate whether they can be used in the training process, and if so, they indicate seen classes, and vice versa. Nowadays, with the promotion of deep learning technology, the performance of zero-shot learning has been greatly improved. Generalized zero-shot learning is a challenging topic that has promising prospects in many realistic scenarios. Although the zero-shot learning task has made gratifying progress, there is still a strong deviation between seen classes and unseen classes in the existing methods. Recent methods focus on learning a unified semantic-aligned visual representation to transfer knowledge between two domains, while ignoring the intrinsic characteristics of visual features which are discriminative enough to be classified by itself. To solve the above problems, we propose a novel model that uses the discriminative information of visual features to optimize the generative module, in which the generative module is a dual generation network framework composed of conditional VAE and improved WGAN. Specifically, the model uses the discrimination information of visual features, according to the relevant semantic embedding, synthesizes the visual features of unseen categories by using the learned generator, and then trains the final softmax classifier by using the generated visual features, thus realizing the recognition of unseen categories. In addition, this paper also analyzes the effect of the additional classifiers with different structures on the transmission of discriminative information. We have conducted a lot of experiments on six commonly used benchmark datasets (AWA1, AWA2, APY, FLO, SUN, and CUB). The experimental results show that our model outperforms several state-of-the-art methods for both traditional as well as generalized zero-shot learning.

1. Introduction

In recent years, deep learning [1–4] has achieved great success in a wide range of computer vision and machine learning tasks [5], including face recognition, emotion classification, and visual question answering. In most cases, these deep learning models are more effective than human beings in many aspects, because they can observe potential information that may be ignored by human eyes in pictures. However, as the inventor of neural network, human beings are better at identifying objects they have never seen before through some prior semantic knowledge about these novel objects. In this respect, the effect of deep learning is not as good as that of humans. It precisely is because deep learning tasks for image recognition rely heavily on fully-supervised training, so they need a very large amount of labeled data. However, some object classes are difficult to obtain, such as the image data of endangered species and newly produced commodities. Moreover, even if they get the labeled data of related classes, they will still face the problem of unbalanced data. It is very difficult to obtain images of these objects, let alone a large number of labeled samples. Therefore, training models with a large number of labeled data are unrealistic. In this background, the concept of zero-shot learning has been put forward, which has attracted wide attention in the field of computer vision and has been greatly developed.

As there are too many classes in the real world, it is impossible to collect enough labeled data for each class. In this case, the task of zero-shot learning is desirable, but it is

challenging. In the literature [6–10], zero-shot learning is usually realized by using the marked samples of seen categories and category-related semantic embedding which is regarded as auxiliary information. The semantic embedding, which encodes the interclass relationships, is usually attribute, word vector, or sentence embedding. Therefore, seen classes and unseen classes are shared in semantic embedding space. In traditional zero-shot learning settings [11, 12], the goal is to train an image classifier on the seen classes and then test the trained classifier on unseen classes, where the seen classes and unseen classes are disjoint. However, the traditional zero-shot learning setting is not realistic, and it is not always applicable in the real world, because in reality, the test images can come from the seen classes. Therefore, there is such a trend that we hope the trained classifier can not only identify unseen classes but also seen classes, which is called generalized zero-shot learning [13, 14]. In the following articles, we uniformly express the traditional zero-shot learning as ZSL and the generalized zero-shot learning as GZSL. The main difference between ZSL and GZSL is whether the label space contains seen classes during the test period. In this work, we have conducted comparative experiments to study both ZSL and GZSL by synthesizing visual features of unseen classes with using the potential and valuable discriminative information.

In this paper, we point out the existing problems of ZSL and GZSL works reported recently, and we analyze the effectiveness of the dual generative network proposed in this paper as well as the discriminative information of visual feature representation. In the early days, as is illustrated in Figure 1(a), most methods [7, 11, 15–18] mapped image visual features to the semantic space to solve ZSL tasks based on class attribute embeddings or other side knowledge. However, using semantic space as the mapping space will suffer from the hubness problem pointed out in [19–21]. It is because projecting high-dimensional visual features to low-dimensional semantic space will greatly reduce the diversity of features that some points from different classes may become more clustered as a hub, as shown in Figure 2. In order to alleviate the hubness problem, some works [19–21] proposed to map semantic features into the visual space as illustrated in Figure 1(b). However, this will lead to another problem called domain shift. For example, the tail of a pig and the tail of a horse are similar in semantic space, but they are quite different in visual space, as shown in Figure 3. Then, the concept of a shared latent space was put forward. People mapped visual features and semantic attributes into a latent space at the same time, as shown in Figure 1(c), and performed nearest neighbor search to calculate the average per-class top-1 accuracy. This shared latent space was considered to alleviate the hubness and shifting problems, but the generalization ability of this method is poor. When using mapping methods for GZSL, the performance will be significantly degraded. Our dual generation model combines the advantages of improved WGAN and conditional VAE, which can alleviate hubness and shifting problems, thus effectively achieving the goal of zero-shot learning and generalized zero-shot learning.

In contrast, most recent ZSL and GZSL approaches [8, 22–25] are based on generative adversarial network [26], which aims at directly optimizing the divergence between real and generated data distributions. The work of Xian et al. [8] learns a GAN by using the seen class visual features and the corresponding semantic embedding that are manually annotated attributes or word vector [27] representations. Fake visual features of the unseen categories are synthesized using the trained generator and then used together with the real visual features of seen classes to train ZSL classifiers in a fully-supervised setting. But GANs are often suffering from mode collapse and unstable training issues. Inspired by the idea of generative adversarial networks, our proposed dual generative framework combines the advantages of conditional variational auto encoder network and improved WGAN, with the discriminative information by using an additional classifier trained on the seen classes to increase the diversity and distinguishability of samples that are generated by the generator. Among them, the improved WGAN can overcome the mode collapse problem, and VAE can alleviate the unstable problem of GAN training, so that our model can stably and quickly generate visual features corresponding to categories according to semantic embedding.

As described above, we combine the advantages of improved WGAN and conditional VAE together with intrinsic characteristics of visual feature representation itself by using an additional classifier to propose a new model called dual generative network with discriminative information (DGDI). Compared with the previous generative methods for ZSL whose models suffer from mode collapse problems [28, 29], our model is more stable by using conditional VAE to assist GAN in generating visual features. In this work, our main task is to obtain a robust generator to synthesis visual features of the unlabeled classes. In particular, if the generator learns discriminative visual feature data with sufficient variation, the generated data should be useful for implementing supervised learning. Moreover, we consider our dual generative framework that was composed by improved WGAN and conditional VAE can learn the complementary information of semantic space, so we believe that our model can produce higher quality visual features from semantic embeddings.

Our main contributions are summarized as follows: (1) we propose a novel generative model named DGDI with combining the advantages of improved WGAN and conditional VAE, which can learn complementary information from semantic embeddings. (2) In contrast to previous zero-shot learning works, we add an additional classifier loss to train the generator by using the intrinsic characteristics of visual feature representation, which makes the synthesized visual features more diverse and distinguishable. (3) We conduct extensive experiments that demonstrate the effectiveness of our proposed model and the results maintain high accuracy for both ZSL and GZSL on six widely used benchmark datasets. In addition, in order to make better use of the discriminative information expressed by visual features, we also analyze the effects of classifiers with different structures. (4) We also conduct visual experiments on synthetic visual features from unseen classes by t-SNE [30], which intuitively proves the effective generation ability of our model.

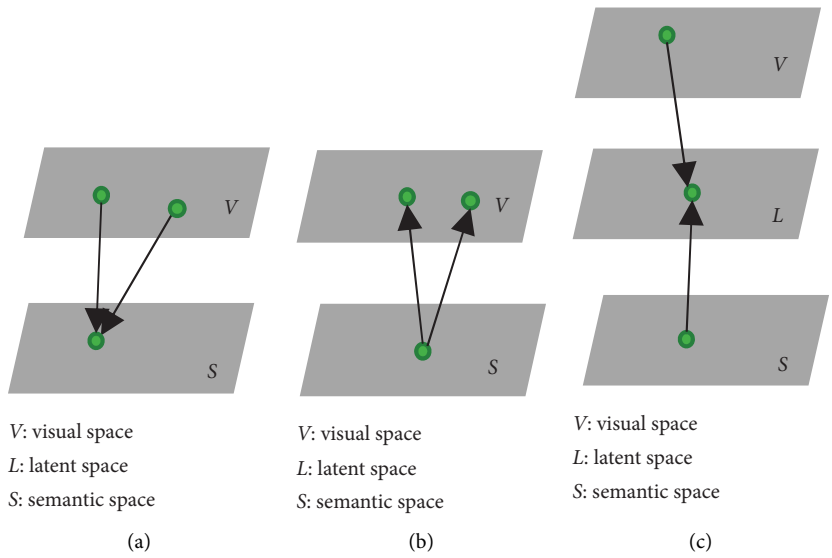


Figure 1: Three mapping methods commonly used in zero-shot learning. (a) Mapping from the visual space to the semantic space. (b) Mapping from the semantic space to its visual space. (c) Mapping both semantic features and visual features to a shared latent space.

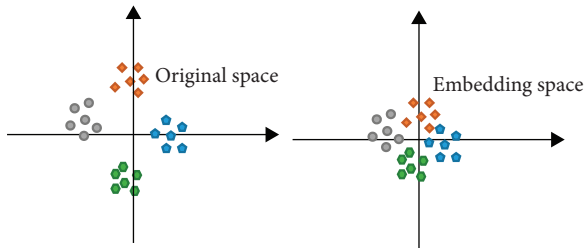


Figure 2: Visual explanation of hubness phenomenon. When a sample point is projected from the original space to the embedding space, the discriminative information in the original space is very likely to be lost, so the sample points belonging to the same class will be closer to other classes, which is especially obvious when mapping from high-dimensional space to low-dimensional space.

2. Related Work

In this section, we will discuss some relevant works on (generalized) zero-shot learning as well as generative models.

We are interested in both ZSL and GZSL tasks, in which the former aims at predicting the labels of unseen classes, while the latter tries to predict labels of both seen and unseen classes. Visual feature representation itself has strong distinguishability, but this is often ignored by previous researchers, so it is not reused. In this paper, a discriminative classifier is added to study the intrinsic distinguishable information of visual features, and it is applied to the dual generation module to synthesize more distinctive feature representations according to the corresponding semantic attributes of categories.

Early works [31, 32] associated seen and unseen classes by directly learning attribute classifiers. However, most recent works either learn a compatibility function between the image feature and class embedding spaces [7, 11, 16, 17, 21] or learn unseen classes, which are the mixture of visible classes [33–35]. For example, SYNC [33, 36, 37] try to predict the labels of unseen classes by learning linear classifiers. Wang et al. [38] proposed to combine the knowledge graph with graph convolutional network [39] and semantic embeddings. Rohrbach et al. [40] and Ye and Guo [9] project image features to the semantic embedding space followed by label propagation. Verma and Rai [41] treat unknown labels of unseen class images as latent variables and apply expectation-maximization (EM). All the abovementioned models are nongenerative and suffer from the problems of hubness as well as domain-shifting, but our proposed method uses a dual generative model to transform ZSL or GZSL into traditional supervised learning by generating fake visual features of unseen classes, which is considered to alleviate the problems of embedding methods.

In recent years, generative models have been widely used. Generative adversarial network [26] was originally proposed as an image synthesis method based on a particular image data distribution [42] and has achieved the state-of-the-art results. Generative adversarial network [26, 42, 43] is composed of a generator that synthesizes fake data distribution and a discriminator that distinguishes fake data from real data. However, GANs are suffering from the problems of unstable training and mode collapse [44, 45]. In order to alleviate these problems and improve the quality of synthesized features, many researches have put forward their own methods. Arjovsky et al. [44] proposed WGAN to optimize GAN on an approximate Wasserstein distance by enforcing 1-Lipschitz smoothness. Although WGAN has obtained better theoretical performance than the original GAN, it still has the problems of disappearance and explosion gradient due to weight clipping to enforce the 1-Lipschitz constraint on the discriminator, and then, Gulrajani et al. [45] proposed an improved version of WGAN which is called WGAN-GP enforcing the Lipschitz constraint [3] through gradient penalty. Therefore, our method draws lessons from the idea of the improved

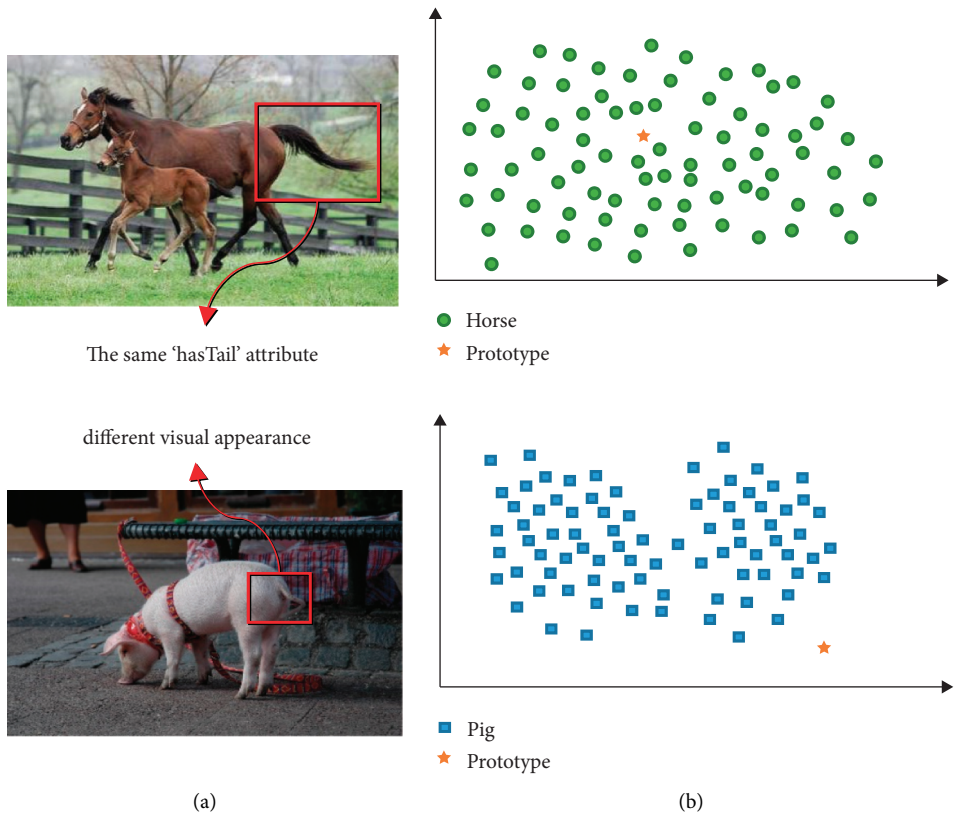


Figure 3: An illustration of the domain shift problem in zero-shot learning image classification. As we can see from the picture, both horses and pigs have tail attributes, but the visual characteristics of their tails are far apart. It is difficult to correctly identify pigs if the model is trained by horses. (a) Visual space. (b) Attribute space.

WGAN. Different from the existing works that directly generate image itself, our proposed model chooses to generate visuals features instead, which can be directly used to train a discriminative classifier for zero-shot learning.

Further, Zhu et al. [46] proposed an interesting application of GANs named CycleGAN that translates an image from one domain to another domain and then back to the original domain to form a closed loop. Schonfeld et al. [47] proposed an approach where cross and distribution alignment losses are introduced for aligning the visual features and corresponding embeddings in a shared latent space, by using two variational auto encoders [48]. The work of [25] is similar to our model, which introduces a f-VAEGAN framework that combines a VAE and a GAN by sharing the decoder of VAE and generator of GAN for feature synthesis. Xian et al. [8] used a conditional Wasserstein GAN [44] along with a seen category classifier to learn the generator for unseen class feature synthesis. Our proposed model combines the idea of VAEGAN of [25] and the seen classes classifier of [8] to encourage the generator to synthesize more discriminative features, which will improve the performance of zero-shot learning and generalized zero-shot learning to a certain extent.

The abovementioned generative methods of zero-shot learning and generalized zero-shot learning almost ignore the inherent distinguishability of visual feature representations between categories, which is actually very important to classification. Therefore, we apply the key discriminative information of visual feature representations to the proposed dual generation framework, which promotes the synthesized visual feature representations generated by the learned generator to be more easily distinguished from each other. In this paper, we also analyze the role of the additional classifier with different structures in the transmission of discriminative information.

3. Proposed Model

In this section, we first formally define the zero-shot learning generalized zero-shot learning problems, give an overview of our proposed dual generative model with using the discriminative information of visual feature representation by an additional classifier, and then introduce each component of our model in detail.

3.1. ZSL and GZSL Problem Formulation. In this paper, we study both the conventional and generalized zero-shot learning. Specifically, let the source dataset be defined as $S = \{v, y, s_s | v \in v_s, y \in \gamma_s, s_s \in A\}$, where S stands for the training data of seen classes, $v = R^{d_v}$ is the image's visual feature produced by a pretrained neural network which is usually ResNet101 trained on ImageNet1K, v_s is the set of visual features from seen classes, y is the label of image visual

feature v, γ_s is the set of labels for seen classes, and $s_s$is the semantic embedding for the class y. Similarly, we can define the test set, i.e., the target dataset as $T = \{v, y, s_u | v \in v_u, y \in \gamma_u, s_u \in A\}$ where the v_u represents the set of image features from unseen classes, γ_u represents the set of labels for unseen classes, and that $\gamma_u \cap \gamma_s = \varnothing$. The tasks in ZSL and GZSL are to learn the classifiers f_{zsl}: $v \longrightarrow \gamma_u$ and f_{gzsl}: $v \longrightarrow \gamma_u \cup \gamma_s$, respectively.

3.2. Model Overview. The overall framework of our proposed framework is illustrated in Figure 4. There are four main components in our model, i.e., an encoder, a generator/decoder, a discriminator, and a pretrained classifier, in which the encoder, the generator/encoder, and the discriminator form a dual generative framework, i.e., VAE-GAN. Our proposed method is based on the recently introduced f-VAE-GAN [25] that combines the advantages of the VAE [48] and GAN [26] which is the same as our proposed method and has achieved impressive results for ZSL classification. Referring to the idea of [25], we add an extra classifier which is the utilization of discriminative information to classify the generated visual features of the seen classes, in which the classifier is pretrained on seen classes. We believe that the additional classifier loss can make the generator learn to synthesize more discriminative visual features which is helpful. The core component of our model is the dual generative framework whose role is to generate various visual features conditioned on certain class semantic embedding. In this paper, we make full use of the inherent discriminative information of visual feature representations and apply this inherent feature to the dual generation module to encourage the generator to synthesize visual feature representations that are easier to be classified based on the corresponding category semantic attributes. In the following, we will introduce the main components dual generative network, the additional classifier, and their loss functions of the proposed model in detail.

3.3. Dual Generative Framework. In this work, we propose a dual generative framework to synthesize visual feature representations of unseen classes stably and efficiently. The dual generative network combines the strengths of improved WGAN and conditional VAE, which can deal with the mode collapse and unstable training problems well.

As we can see from Figure 4, the conditional VAE network is composed of a latent noise encoder $p(z|v, s)$ and a visual feature representation decoder $p(v|z, s)$, and the conditional VAE is proposed as a generative method that maps a random noise vector $z = R^{d_v}$ drawn from $p(z|v, s)$to a data point v in the data distribution conditioning on the semantic embeddings. We train conditional VAE by minimizing the following loss function L_{CVAE}:

$$L_{\mathrm{CVAE}} = \mathrm{KL}(p(z|v, s) \| p(z|s)) - E(\log p(v|z, s)), \quad (1)$$

where $\mathrm{KL}(p(z|v, s)||p(z|s))$ represents the L_{KL}, i.e., the Kullback–Leibler divergence between $p(z|v, s)$and $p(z|s)$, the conditional distribution $p(z|v, s)$ is modeled as $E(v, s)$,$p(z|v, s)$ is equal to $G(z, s)$, and $p(z|s)$ is treated as a unit Gaussian distribution.

As shown in Figure 4, the improved WGAN is composed of a generator G and a discriminator D. We aim to learn a generator$G$: $z \times c \longrightarrow v$ conditioned on semantic embeddings. The generator takes class embedding $s \in A$ and random Gaussian noise $z = R^{d_v}$as inputs and then outputs a fake visual feature $\tilde{v}$ of the class y. The loss function of our improved WGAN is

$$L_{\mathrm{WGAN}} = E[D(v, s_s)] - E[D(\tilde{v}, s)] - \lambda E\left[\left(\left\|\nabla_{\hat{v}} D(\hat{v}, s)\right\| - 1\right)^2\right], \quad (2)$$

where $\tilde{v} = G(z, s_s)$, $\hat{v} = \alpha v + (1 - \alpha)\tilde{v}$, with $\alpha \sim U(0, 1)$, and λ is the penalty coefficient, initialized to 10. Different from the pure GAN, the discriminative network of WGAN is defined as D: $v \times c \longrightarrow R$which eliminates the sigmoid layer and outputs a real value. The first two terms of Equation (2) are considered as Wasserstein distance, and the third term is the gradient penalty to enforce the gradient of D to have unit norm along the straight line between real and generated visual feature pairs. We also calculate the value of the gradient penalty term in each epoch of training to adjust the super-parameter λ.

Once the dual generative model learns to generate visual features of seen classes, conditioned on the seen class semantic embeddings s_s, it can also generate $\tilde{v}_u$ of any unseen category γ_u through its class semantic embedding s_u. So, the zero-shot learning and generalized zero-shot learning problems can be transformed into traditional supervised learning.

3.4. Additional Classifier for Discriminative Information. In order to ensure that the visual features generated by improved WGAN are well suited for training a discriminative classifier, we added a classifier C to make use of the discriminative information of visual feature representations, as shown in Figure 4, which is pretrained on the real features of seen classes to encourage the generator to generate distinctive features. For this purpose, module C uses the negative log likelihood to minimize the classification loss over the generated features in the following formulation:

$$L_{\mathrm{CLS}} = -E_{\tilde{v} \sim p_{\tilde{v}}}[\log P(y|\tilde{v}; \theta)], \quad (3)$$

where $\tilde{v} = G(z, s)$, y is the class label of $\tilde{v}$, and $P(y|\tilde{v}; \theta)$ denotes the probability of $\tilde{v}$ being predicted with its true class label y. The conditional probability is computed by a linear softmax classifier parameterized by θ. The classification loss can be regarded as a regularization that enforces the generator to construct discriminative features. In the next section, we carry out experiments to analyze the performance of different classifiers for zero-shot learning and generalized zero-shot learning.

In summary, our proposed model optimizes the following objective function:

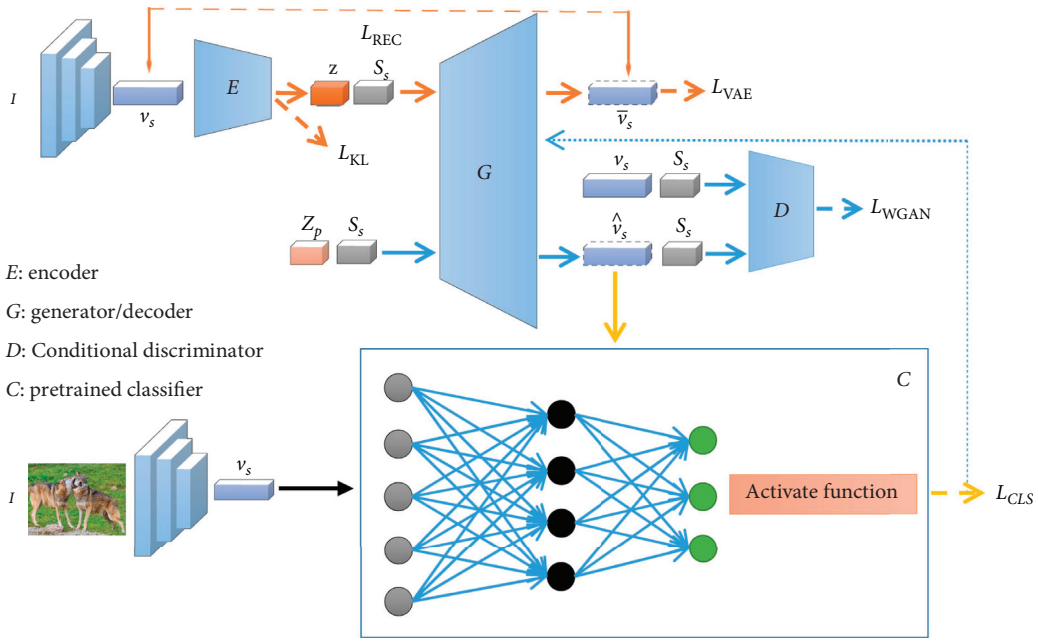


Figure 4: An overview of our proposed architecture, in which the upper part is a dual generation framework and the lower part is an additional classifier using discriminative information. Given the image samples I selected from seen classes, visual features v_s are extracted from the Resnet101 which is pretrained on ImageNet1K and input to the netE (encoder), along with the corresponding semantic embeddings s_s. The latent noise vector z output from netE is then input together with semantic embeddings s_s to the netG (generator) that synthesizes fake visual features $\bar{v}_s$. The netD (conditional discriminator) learns to distinguish real features v_s from synthesized $\bar{v}_s$. Both netE and netG together constitute the so-called conditional VAE, which is training by L_{KL} (KL divergence) and L_{BCE} (binary cross-entropy loss). Similarly, both netG and netD are trained using L_{WGAN}. The additional classifier is a multilayer fully-connected neural network, and we also discuss the influence of different classifier structures in the following.

$$\min_{G,E} \max_{D} L_{CVAE} + \gamma L_{WGAN} + \alpha L_{CLS}. \tag{4}$$

As shown in Figure 5, once the model has been trained, in order to predict the label of unseen classes, we can first generate pseudovisual features for each unseen class using the learned generator. Then, we construct a new dataset by combining these pseudovisual features with the real features of the seen classes for GZSL. After that, we can train any classifier based on this new dataset containing the visual features of the seen classes and unseen classes. Therefore, the GZSL task is transformed into a supervised learning problem. Here, we use a self-learning classifier to fine-tune the accuracy as in [24].

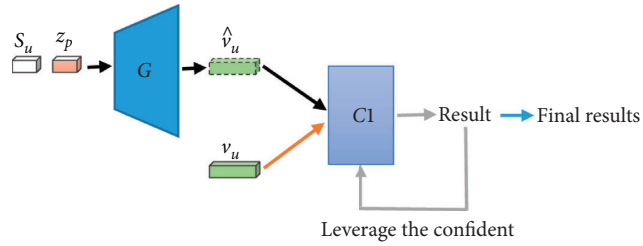


Figure 5: The test process of our method for ZSL: firstly, we use the learned generator to synthesize the visual features of unseen categories conditioned on semantic embeddings; then, we train the classifier by utilizing these synthesized visual features; finally, the real unseen samples are used for testing. The test part is divided into two steps, the first step is rough classification, and then the second step is to fine-tune the classifier by using the confident classification results in the first step.

4. Experiments

In this section, we have conducted a lot of experiments on six public benchmark datasets for both ZSL and GZSL. The detailed information of the experimental setup is provided in the respective chapters, and in order to make better use of the discriminative information, we discuss the influences of classifiers with different structures by conducting comparative experiments and comprehensively analyze the corresponding experimental results.

4.1. Datasets and Settings. We compare our proposed model with several baselines on six widely used datasets, i.e., Oxford Flowers (FLO) [49], Animals with Attributes 2 (AWA2) [14], Caltech-UCSD-Birds (CUB) [50], SUN Attribute (SUN) [51], and APascal-a Yahoo (APY). Among these datasets, APY contains 32 categories from both PASCAL VOC 2008 and YahooL that contain 15339 images. AWA2 is a coarse-grained and medium-size dataset which contains 30,475 images, 50 classes, and 85 attributes. CUB, FLO, and SUN are medium scale but fine-grained datasets, in which SUB contains 11788 images from 200 different types of birds annotated with 312 attributes. FLO dataset contains 8189 images from 102 different types of flowers without attribute annotations. However, we use the fine-

grained visual descriptions collected by [27]. SUN contains 14340 images from 717 scenes annotated with 102 attributes. Statistics of the datasets are presented in Table 1.

For real visual features, we extract 2048-dim top-layer pooling units of the ResNet101 [56] from the entire image. We do not do any image preprocessing such as cropping or use any other data augmentation techniques. ResNet101 is pretrained on ImageNet1K and not fine-tuned. For pseudovisual features, we generate 2048-dim features using our model. For the class semantic embeddings, we use per-class attributes for AWA (85-dim), CUB (312-dim) and SUN (102-dim), APY (64-dim). Furthermore, for dataset FLO, we extract 1024-dim character-based features from fine-grained visual descriptions by CNN-RNN [57].

At test time, in the ZSL setting, the goal is to correctly classify unseen class label, i.e., γ_u, and in the GZSL setting, the search space includes both seen and unseen classes, i.e., $\gamma_s \cup \gamma_u$. We use the unified evaluation protocol in [58]. In the ZSL setting, we first calculate the average accuracy of each category independently and then sum the average accuracy of all categories and divide by the total number of categories to get average per-class top-1 accuracy ($T1$). As for the GZSL setting, we compute the average per-class top-1 accuracy on seen classes γ_s denoted as s and the average per-class top-1 accuracy on unseen classes γ_u denoted as u; after that we calculate their harmonic mean as the final measure, i.e., $H = 2 * (s * u)/(s + u)$.

4.2. Implementation Details. In our proposed model, the encoder, the generator, and the discriminator are all implemented as multilayer perceptron (MLP). Through experiments, we find that when the dimensions of semantic embeddings s and Gaussian random noise $z \sim N(0, 1)$ are the same, the performance of zero-shot learning is the best. Therefore, we set the dimension of Gaussian random noise as the dimension of semantic embeddings of each dataset. The latent vector z and semantic embeddings s are concatenated and feed into the generator. Similarly, the discriminators take input as the concatenation of image features and class embeddings. In which, the discriminator, the encoder, and the generator are all two-layer fully-connected (FC) networks with 4096 hidden units. In addition to the output layer of G, other components use LeakyReLU as a nonlinear activation function. While for G, sigmoid activation is used to apply BCE loss. Through experiments, we prove that when this extra classifier is a single-layer perceptron, it is better to use the discriminative information by visual feature representations. The model is trained using the Adam optimizer with learning rate of 0.0001. Following the suggestion of WGAN paper [44], we update the generator once every 5 discriminator iterations. Hyperparameters α and γ are initialized to 1 and 10, respectively, and then tuned by cross-validation.

4.3. Comparing with State-of-the-Art Methods. We compare our approach with ALE [6], f-WGAN [8], SE-GZSL [52], Sycle-WGAN [22], LisGAN [24], f-VAEGAN [25], TCN [53], DVBE [55], and SAE [45] for both ZSL and GZSL, and two more approaches, CADA-VAE [54] and DVBE [55] are compared for GZSL. The above methods are either representative ones or the state-of-the-art ones published in the past few years. Following previous work [24, 25], we report the average per-class top-1 accuracy. Specifically, for ZSL, we report the top-1 accuracy of unseen samples by only searching the unseen label space. However, for the GZSL, we report the accuracy on both seen classes and unseen classes with the same settings in [58]. Some of the results reported in this paper are also cited from [5].

Table 2 reports the results of ZSL. In these experiments, the categories of test samples are only searched from γ_u . It can be seen that the classification accuracies obtained on AWA1, APY, FLO, SUN, and CUB are 71.4%, 44.9%, 73.6%, 65.1%, and 62.6%, respectively. Our proposed framework has improved the state-of-the-art performance on AWA1, APY, FLO, SUN, and CUB datasets by 0.3%, 1.8%, 3.3%, 0.4%, and 1.6%. As for AWA2, we achieve the best of previous works. From Table 2, we can also observe that the generation-based methods, e.g., LisGAN, f-CLSWGAN, and ours, generally have better results than embedding ones, e.g., ALE. The GAN method transforms ZSL into supervision problem by generating visual features of unseen classes, while the embedding methods use indirect way to deal with unseen classes. This also proves the validity of the generative model in ZSL problem. Generally speaking, our method produces one of the best performances compared to the existing methods on five of six datasets.

Table 3 summarizes the results of GZSL. From Table 3, we can observe that our proposed model has better performance than existing methods, which is similar to the conclusion to Table 2. Our method stably predicts seen and unseen classes. Although some previous methods, such as ALE, performed well in identifying unseen samples in ZSL settings, their performance in GZSL decreased significantly. When the number of unseen classes becomes larger, ZSL models always tend to be confused, resulting in performance degradation. This phenomenon is especially obvious when the number of unseen classes is much larger than that of seen classes. Moreover, in real life, the amount of seen classes that can get manual annotations is definitely far less than that of unseen classes. Therefore, the applicability of these ZSL methods in practical application is limited and GZSL is the development trend in line with the reality.

We use harmonic mean which is considered more stable than arithmetic and geometric mean to measure the mean value between the accuracy of seen and unseen classes. From the reported results from Table 3, we can find that our method is more stable than the existing methods. Our proposed method avoids the unbalanced and extreme results between $sacc_s$ and u. As far as harmonic mean H is concerned, we achieved up to 0.3%, 0.2%, 3.1%, 0.8%, and 1.1% improvements on AWA2, APY, FLO, SUN, and CUB, respectively. The average is 1.1% over the five. Although our model did not perform the best on AWA1, its performance is almost equal to the previous artistic level. It can be seen from the results that our method reduces the precision difference between known classes and unknown classes to a certain extent, which verified the effective generalization ability of our method.

Table 1: Statistics of datasets.

Dataset	att/stc	Seen classes (train + val)	Unseen classes	Images (train + val)	Images (test unseen/seen)
APY	64	15 + 5	12	5932	7924/1483
AWA1	85	27 + 13	10	19832	5685/4958
AWA2	85	27 + 13	10	23527	7913/5882
CUB	312	100 + 50	50	7057	2967/1764
SUN	102	580 + 65	72	10320	1440/2580
FLO	1024	62 + 20	20	5631	1403/1155

Table 2: Results of ZSL on six classification benchmarks. ZSL measuring per-class average top-1 accuracy (T1) on γ_u. The "-" means that there are no relevant results in the reference, while the underlined results are reproduced according to the description of references.

Method	AWA1	AWA2	APY	FLO	SUN	CUB
ALE [6]	59.9	—	—	48.5	58.1	54.9
f-WGAN [8]	68.2	—	—	67.2	60.8	57.3
SE-GZSL [52]	69.5	69.2	—	—	63.4	59.6
Sycle-WGAN [22]	66.8	—	—	70.3	59.9	58.6
LisGAN [24]	70.6	—	43.1	69.6	61.7	58.8
f-VAEGAN [25]	71.1	70.5	40.4	67.7	64.7	61.0
TCN [53]	70.3	71.2	38.9	—	61.5	59.5
Ours	71.4	71.2	44.9	73.6	65.1	62.6

Considering the fact that both f-WGAN and f-CLSWGAN leverage GANs to synthesize unseen visual features, the performance improvement of our method can be attributed to two aspects. One is that we introduce a classifier trained on seen classes to guarantee that the generated features of each class can be distinguished from each other, which is considered as the usage of the discriminative information. The other is our classifier self-learning mechanism at test time, which is able to leverage the confident results to fine-tune itself. In general, the results verify that it is beneficial to leverage the additional classifier to train VAEGAN. The correct classification of generated unseen visual features guarantee that each synthesized sample features is highly related with its category and is more distinguishable.

4.4. Discussion of the Additional Classifier. Here, we analyze the influence of the additional loss of classifiers with different structures on the performance of zero-short learning and generalized zero-shot learning. The experimental results on datasets SUN and CUB are shown in Table 4.

As we can see from Table 4, the effect of single-layer perceptron is the best among all tested classifiers, except for the accuracy of the ZSL of the SUN. The output layer of all classifiers uses sigmoid as the activation function to calculate the classification loss, thus constraining the dual generation network to synthesize the visual feature representation which is easy to classify. By comparing the experimental results from lines 2 to 4 and lines 3 to 7 in Table 4, we found that using ReLu as an activation function for the hidden layer worked best. At the same time, from the data of the last three rows and the top three rows in Table 4, it can be seen that the hidden layer uses 1024 units better than 512 for both ZSL and GZSL. Through experiments, we found that using single-layer neural network as an additional classifier to understand the discrimination information can not only get the best results, but also reduce the running time time.

4.5. Analysis of Synthetic Image Features. In order to provide an intuitive evaluation on our proposed model, we visualize the visual features of some synthetic image visual features and the corresponding real image visual features of unseen classes. The results are shown in Figure 6. For convenience, we chose 10 unseen categories of AWA2 dataset for visualization. First of all, we get the semantic embeddings and the real image features of the selected categories. Secondly, we input these semantic embeddings and Gaussian random noise into the learned generator to obtain the synthetic image features. Finally, we use t-SNE [30] to reduce the dimension of synthetic and real visual features from 2048 to 2 and plot the obtained feature data into scatter for visualization.

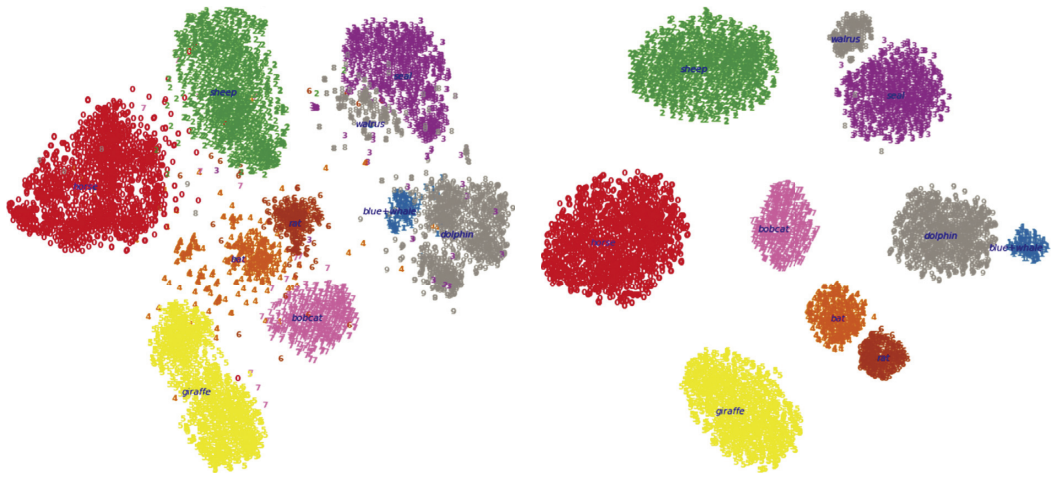
From the visualization of real feature samples in Figure 6(a), it can be seen that some categories overlap to a large extent, such as seals, walruses, blue whales, and dolphins. It is reasonable for them to overlap, because blue whales, dolphins, seals, and walruses are similar in biology and look very similar visually. The visualization of synthetic image features is shown in Figure 6(b). By comparing 6(a) and 6(b), we can clearly find that for most categories, such as seals and dolphins, the synthetic image features are very close to real samples, and some of them even overlap with real samples well, such as horses, sheep, and giraffes. One failure is rat, and we can see that the synthesized features are far from the real features. Another disadvantage is that there is almost no confusion between the categories of synthetic samples, which is contrary to the actual situation. However, the finally

Table 3: Results of GZSL on six classification benchmarks respectively. GZSL measuring the harmonic mean H of the per-class top-1 accuracy u on γ_u and the per-class top-1 accuracy s on γ_u. The "—" means that there are no relevant results in the reference, while the underlined results are reproduced according to the description of references.

Method	AWA1			AWA2			APY			FLO			SUN			CUB		
	u	s	H	u	s	H	u	s	H	u	s	H	u	s	H	u	s	H
ALE [6]	16.8	76.1	27.5	—	—	—	—	—	—	13.3	61.6	21.9	21.8	33.1	26.3	23.7	62.8	34.4
f-WGAN [8]	57.9	61.4	59.6	—	—	—	—	—	—	59.0	73.8	65.6	42.6	36.6	39.4	43.7	57.7	49.7
SE-GZSL [52]	58.3	67.8	61.5	58.3	68.1	62.8	—	—	—	—	—	—	40.9	30.5	34.9	41.5	53.3	46.7
Sycle-WGAN [22]	59.6	63.4	59.8	—	—	—	—	—	—	61.6	69.2	65.2	47.2	33.8	39.4	47.9	59.3	53.0
LisGAN [24]	52.6	76.3	62.3	—	—	—	34.3	68.2	45.7	57.7	83.8	68.3	42.9	37.8	40.2	46.5	57.9	51.6
f-VAEGAN [25]	57.6	70.6	63.5	55.2	73.6	63.1	30.3	58.6	39.9	56.8	74.9	64.6	45.1	38.0	41.3	48.4	60.1	53.6
TCN [53]	49.4	76.5	60.0	61.2	65.8	63.4	24.1	64.0	35.1	—	—	—	31.2	37.3	34.0	52.6	52.0	52.3
CADA-VAE [54]	72.8	57.3	64.1	75.0	55.8	63.9							36.7	47.2	40.6	53.5	51.6	52.4
DVBE [55]	—	—	—	63.6	70.8	67.0	32.6	58.3	41.8	—	—	—	45.0	37.2	40.7	53.2	60.2	56.5
Ours	58.7	70.3	64.0	60.1	76.4	67.3	36.5	61.7	45.9	62.6	83.0	71.4	48.3	37.4	42.1	53.8	61.9	57.6

Table 4: The results of different classifiers on SUN and CUB. The single-layer structure is what we use for our proposed model. The two layers mean that a latent layer is added to the single-layer structure, the _relu/_lrelu/_sigmoid indicates the activation function followed by the hidden layer, and _1024/_512 indicates the number of neurons in the added hidden layer.

	SUN				CUB			
	ZSL	GZSL			ZSL	GZSL		
	$T1$	s	u	H	$T1$	s	u	H
Single-layer (used)	65.1	37.4	48.3	42.1	62.6	61.9	53.8	57.6
two_layers_1024_relu	65.8	37.8	47.4	42.1	61.8	62.4	47.3	53.8
two_layers_1024_lrelu	65.1	37.3	48.1	41.6	61.3	60.4	47.5	53.2
two_layers_1024_softmax	65.1	38.2	46.1	41.8	61.3	61.7	47.5	53.7
two_layers_512_relu	65.3	37.6	45.6	41.2	61.4	59.0	48.3	53.1
two_layers_512_lrelu	64.5	38.0	45.6	41.5	61.1	62.3	46.0	52.9
two_layers_512_ softmax	64.7	37.3	46.1	41.2	61.2	57.5	49.2	53.0



(a) (b)

Figure 6: . SNE visualization of real (a) and synthetic (b) image features for unseen classes in AWA2 datasets.

trained softmax classifier can well predict the labels of most categories of test images.

5. Conclusion

In this paper, we discuss the generalized zero-shot learning task and propose a model called DGDI, a dual generative framework that combines the advantages of conditional VAE and improved WGAN to obtain a more robust generative model with the using of discriminative information by adding a classification loss. We make full use of the discriminative information of visual feature representation between categories to further improve our dual generative module by adding a softmax classifier pretrained on the seen classes to encourage the generator to learn the discriminative information. The experimental results on six datasets clearly

show the effectiveness of our proposed framework; our method has achieved good performance on almost all datasets, which fully proves the importance of the discriminative information between the visual feature representations of categories. It is a meaningful problem to improve the precision and generalization ability of zero-shot learning, and we will further study it.

Data Availability

The datasets used in this study can be downloaded from http://datasets.d2.mpi-inf.mpg.de/xian/xlsa17.zip.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant no. 2019YFA0706200, the National Major Research Program of China under Grant no. 2018AAA0102002, and the National Natural Science Foundation of China (NSFC) under Grant nos. 61976076 and 61632007.

References

[1] T. Dong and T. Huang, "Neural cryptography based on complex-valued neural network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 4999–5004, 2020.

[2] T. Dong and L. Xia, "Spatial temporal dynamic of a coupled reaction-diffusion neural network with time delay," *Cognitive Computation*, vol. 11, no. 2, pp. 212–226, 2019.

[3] T. Dong, X. Bu, and W. Hu, "Distributed differentially private average consensus for multi-agent networks by additive functional Laplace noise," *Journal of the Franklin Institute*, vol. 62, pp. 50–64, 2020.

[4] M. Wang, W. Fu, X. He, S. Hao, and X. Wu, "A survey on large-scale machine learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 1, 1 page.

[5] O. Russakovsky, J. Deng, H. Su et al., "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[6] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label embedding for image classification," *TPAMI-IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 7, 2016.

[7] B. Romera-Paredes and P. H. Torr, "An embarrassingly simple approach to zero-shot learning," in *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, July 2015.

[8] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," 2018, https://arxiv.org/abs/1712.00981.

[9] M. Ye and Y. Guo, "Zero-shot classification with discriminative semantic representation learning," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.

[10] J. He, R. Hong, X. Liu, M. Xu, Z.-J. Zha, and M. Wang, "Memory-augmented relation network for few-shot learning," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1236–1244, Association for Computing Machinery, New York, NY, USA, October 2020.

[11] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label embedding for attribute-based classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 819–826, Portland, OR, USA, June 2013.

[12] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition-CVPR 2009*, pp. 951–958, IEEE, Miami, FL, USA, June 2009.

[13] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha, "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," 2016, https://arxiv.org/abs/1605.04253.

[14] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly," 2018, https://arxiv.org/abs/1707.00600.

[15] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2927–2936, Boston, MA, USA, June 2015.

[16] A. Frome, G. S. Corrado, J. Shlens et al., "A deep visual-semantic embedding model," in *Proceedings of the NIPS-26th International Conference on Neural Information Processing Systems*, Red Hook; NY, USA, December 2013.

[17] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *Proceedings of the CVPR-Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 2017.

[18] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," 2016, https://arxiv.org/abs/1603.08895.

[19] G. Dinu, A. Lazaridou, and M. Baroni, "Improving zero-shot learning by mitigating the hubness problem," 2014, https://arxiv.org/abs/1412.6568.

[20] Y. Shigeto, I. Suzuki, K. Hara, M. Shimbo, and Y. Matsumoto, "Ridge regression, hubness, and zero-shot learning," in *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 135–151, Springer, Berlin, Germany, September 2015.

[21] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," 2017, https://arxiv.org/abs/1611.05088.

[22] R. Felix, I. Reid, and G. Carneiro, "Multi-modal cycle-consistent generalized zero-shot learning," in *Proceedings of the ECCV-European Conference on Computer Vision*, Munich, Germany, September 2018.

[23] He Huang, C. Wang, S. Yu Philip, and C.-D. Wang, "Generative dual adversarial network for generalized zero-shot learning," in *Proceedings of the CVPR-Conference on Computer Vision and Pattern Recognition*, Salt Lake, UT, USA, June 2019.

[24] J. Li, M. Jing, Ke Lu, Z. Ding, L. Zhu, and Zi Huang, "Leveraging the invariant side of generative zero-shot learning," in *Proceedings of the CVPR-Conference on Computer Vision and Pattern Recognition*, Salt Lake, UT, USA, September 2019.

[25] Y. Xian, S. Sharma, Bernt Schiele, and Z. Akata, "f-vaegan-d2: a feature generating framework for any-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake, UT, USA, June 2019.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial nets," in *Proceedings of the NIPS-International Conference on Neural Information Processing Systems*, Washington; DC, USA, December 2014.

[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the NIPS-Neural Information Processing Systems*, Lake Tahoe, NV, USA, December 2013.

[28] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," 2017, https://arxiv.org/abs/1701.04862.

[29] T. Dong and Q. Zhang, "Stability and oscillation analysis of a gene regulatory network with multiple time delays and diffusion rate," *IEEE Transactions on NanoBioscience*, vol. 19, no. 2, pp. 285–298, 2020.

[30] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[31] D. Jayaraman and K. Grauman, "Zero-shot recognition with unreliable attributes," in *Proceedings of the NIPS International Conference on Neural Information Processing Systems*, Washington; DC, USA, December 2014.

[32] C. Lampert, H. Nickisch, and S. Harmeling, "Attributebased classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, 2013.

[33] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," 2016, https://arxiv.org/abs/1603.00550.

[34] M. Norouzi, T. Mikolov, S. Bengio et al., "Zero-shot learning by convex combination of semantic embeddings," 2014, https://arxiv.org/abs/1312.5650.

[35] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," 2015, https://arxiv.org/abs/1509.04767.

[36] M. Elhoseiny, B. Saleh, and A. Elgammal, "Write a classifier: zero-shot learning using purely textual descriptions," in *Proceedings of the ICCV-International Conference on Computer Vision*, Sydney, Australia, December 2013.

[37] J. Lei Ba, K. Swersky, S. Fidler et al., "Predicting deep zeroshot convolutional neural networks using textual descriptions," in *Proceedings of the ICCV-2015 International Conference on Computer Vision*, Santiago, Chile, December 2015.

[38] X. Wang, Y. Ye, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," 2018, https://arxiv.org/abs/1803.08035.

[39] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proceedings of the ICLR-International Conference on Learning Representations*, Toulon, France, April 2017.

[40] M. Rohrbach, S. Ebert, and B. Schiele, "Transfer learning in a transductive setting," in *Proceedings of the NIPS-Neural Information Processing Systems*, Sierra Nevada, Spain, December 2013.

[41] V. K. Verma and P. Rai, "A simple exponential family framework for zero-shot learning," 2017, https://arxiv.org/abs/1707.08040.

[42] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2016, https://arxiv.org/abs/1511.06434.

[43] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, https://arxiv.org/abs/1411.1784.

[44] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," 2017, https://arxiv.org/abs/1701.07875.

[45] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," 2017, https://arxiv.org/abs/1704.00028.

[46] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired imageto-image translation using cycle-consistent adversarial networks," 2017, https://arxiv.org/abs/1703.10593.

[47] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero-and few-shot learning via aligned variational autoencoders," 2019, https://arxiv.org/abs/1812.01784.

[48] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the ICLR-International Conference on Learning Representations*, Banff, AB, Canada, April 2014.

[49] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Bhubaneswar, India, December 2008.

[50] P. Welinder, S. Branson, T. Mita et al., "Caltech-UCSD birds 200," Technical Report CNS-TR-2010-001, Caltech, Pasadena, CA, USA, 2010.

[51] G. Patterson and J. Hays, "Sun attribute database: discovering, annotating, and recognizing scene attributes," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, June 2012.

[52] V. Kumar Verma, G. Arora, A. Mishra, and P. Rai, "Generalized zero-shot learning via synthesized examples," in *Proceedings of the CVPR-Conference on Computer Vision and Pattern Recognition*, Salt Lake, UT, USA, June 2018.

[53] H. Jiang, R. Wang, S. Shan, and X. Chen, "Transferable contrastive network for generalized zero-shot learning," 2019, https://arxiv.org/abs/1908.05832.

[54] E. Schönfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero- and few-shot learning via aligned variational autoencoders," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8239-8247, Long Beach, CA, USA, June 2019.

[55] S. Min, H. Yao, H. Xie, C. Wang, Z.-J. Zha, and Y. Zhang, "Domain-aware visual bias eliminating for generalized zero-shot learning," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12661–12670, Seattle, WA, USA, September 2020.

[56] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *Proceedings of the European Conference on Computer Vision*, vol. 3–19, Springer, New York, NY, USA, August 2016.

[57] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," 2016, https://arxiv.org/abs/1605.05395.

[58] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning - the good, the bad and the ugly," 2017, https://arxiv.org/abs/1703.04394.