SCIENCE CHINA
Information Sciences

February 2021, Vol. 64 120101:1–120101:14
https://doi.org/10.1007/s11432-020-3032-8

• RESEARCH PAPER •
Special Focus on Deep Learning for Computer Vision

Triple discriminator generative adversarial network for zero-shot image classification

Zhong JI[1*], Jiangtao YAN[1], Qiang WANG[1], Yanwei PANG[1] & Xuelong LI[2]

[1]*School of Electrical and Information Engineering, Tianjin University, Tianjin* 300072*, China;*
[2]*Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an* 710129*, China*

Received 6 March 2020/Revised 24 May 2020/Accepted 1 July 2020/Published online 20 January 2021

Abstract One key challenge in zero-shot classification (ZSC) is the exploration of knowledge hidden in unseen classes. Generative methods such as generative adversarial networks (GANs) are typically employed to generate the visual information of unseen classes. However, the majority of these methods exploit global semantic features while neglecting the discriminative differences of local semantic features when synthesizing images, which may lead to sub-optimal results. In fact, local semantic information can provide more discriminative knowledge than global information can. To this end, this paper presents a new triple discriminator GAN for ZSC called TDGAN, which incorporates a text-reconstruction network into a dual discriminator GAN (D2GAN), allowing to realize cross-modal mapping from text descriptions to their visual representations. The text-reconstruction network focuses on key text descriptions for aligning semantic relationships to enable synthetic visual features to effectively represent images. Sharma-Mittal entropy is exploited in the loss function to make the distribution of synthetic classes be as close as possible to the distribution of real classes. The results of extensive experiments over the Caltech-UCSD Birds-2011 and North America Birds datasets demonstrate that the proposed TDGAN method consistently yields competitive performance compared to several state-of-the-art ZSC methods.

Keywords zero-shot classification, generative adversarial nets, text reconstruction, Sharma-Mittal entropy

Citation Ji Z, Yan J T, Wang Q, et al. Triple discriminator generative adversarial network for zero-shot image classification. Sci China Inf Sci, 2021, 64(2): 120101, https://doi.org/10.1007/s11432-020-3032-8

1 Introduction

Visually distinguishing unseen classes based only on their text descriptions is an attractive characteristic of human beings, which is a desired learning and generalization capability in the machine learning domain [1–3]. For example, humans can recognize a zebra when they first see its appearance based only on the description that a zebra is a striped horse. The concept of zero-shot classification (ZSC) has been introduced to emulate this capability [4, 5].

Generally, ZSC is achieved by transferring knowledge from seen classes to unseen classes using some additional information such as attributes, word vectors, and descriptions. Early ZSC methods rely on semantically meaningful attributes to transfer knowledge [4, 6]. The majority of these methods transfer cross-modal information [5, 7] through the joint embedding of image visual features and attributes [8–11]. As intermediate representations, attributes share properties across multiple classes, indicating whether some predefined properties exist. However, such methods require experts in the field to annotate a potentially huge amount of data attributes, which is a tedious and costly task. To overcome this problem, word vectors and text descriptions requiring no predefined annotation are employed in ZSC. In particular, word vectors are typically obtained using neural networks trained on a large language corpus, while text descriptions are directly acquired from the Internet, e.g., from Wikipedia articles. Descriptions can provide richer and more detailed information than word vectors can. However, the problem is that these textual descriptions often contain a large amount of redundant information.

* Corresponding author (email: jizhong@tju.edu.cn)

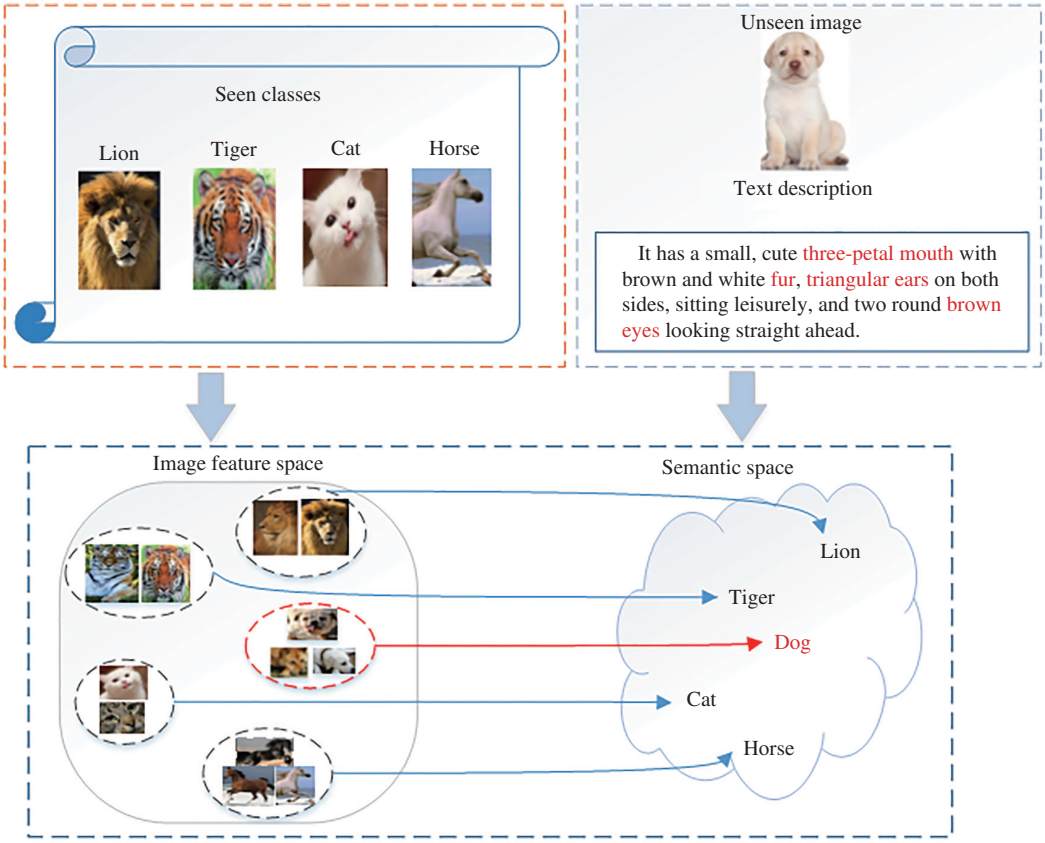

Figure 1 (Color online) Motivation behind this study. Suppose that the tiger, cat, horse, and lion classes belong to the training set, while the dog class belongs to the test set. The dog's text description is noisy; the discriminative description is marked in red. The purpose of this study is to generate fake images of the corresponding test classes using semantic textual descriptions. Each fake image is semantically associated with the real images of seen classes, while their related information is transferred to unseen classes.

Generative adversarial networks (GANs) have been widely applied to ZSC. In GANs, the idea is to synthesize pseudo visual features for unseen classes by projecting the class semantic prototypes into the visual space. However, when applying descriptions as auxiliary information, the noisy information among them seriously degrades their effect. To address this problem, a triple discriminator GAN model is proposed in this study to mine effective textual parts and synthesize discriminative visual features for images in unseen classes. The motivation behind this proposal is illustrated in Figure 1. In particular, the proposed model is built upon a dual discriminator GAN (D2GAN) [12], which incorporates the Lipschitz constraint to make the weights follow the Gaussian distribution, thereby improving the network stability for ZSC. A text-reconstruction network is further introduced to utilize local text parts and highlight the key information from text descriptions. Sharma-Mittal (SM) entropy is employed to impose constraints on weights for different textual features with the purpose of reducing irrelevant text information so that the sample local visual features synthesized by GAN are more discriminative.

In summary, the contributions of this work are twofold:

• We propose a novel triple discriminator GAN (TDGAN) for ZSC, which employs a GAN to synthesize visual features for images of unseen classes. In particular, TDGAN employs three discriminators for emphasizing infomation of real samples, focusing on synthetic samples, and discriminating the visual features generated by reconstructed text respectively. It can progressively improve the stability of network training and the diversity of generated samples.

• By exploiting class-level text description information, we develop a text-reconstruction network for guiding the process of generating features and highlighting key information of text descriptions. Besides, we introduce the SM entropy as a constraint for the embedding of text information to generate common representation.

The performance of the proposed TDGAN is tested on the Caltech-UCSD Birds-2011 (CUB) [13] and

North America Birds (NABirds) [14] ZSC datasets under two different settings.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 presents the proposed TDGAN approach, including the details on the employed D2GAN and the text-reconstruction network. Section 4 outlines the experiments and ablation studies. Section 5 concludes the paper.

2 Related work

2.1 Description-based zero-shot classification

As easily accessible side information for ZSC, class descriptions offer a rich knowledge about both seen and unseen classes. They can be obtained from various sources such as Wikipedia articles [15, 16], and sentence descriptions [17]. For example, Ba et al. [16] proposed to utilize Wikipedia articles for ZSC to avoid the problem of having to explicitly define attributes. The authors mapped raw descriptions and images to a shared embedding space, where the dot product was utilized to learn their matching relations. Qiao et al. [18] presented an $L_{2,1}$-norm based objective function to simultaneously suppress noise in descriptions and learn to match images and their descriptions. The authors demonstrated that inherent noise in text descriptions has a significant impact on ZSC. Elhoseiny et al. [19] proposed a new learning framework associating text descriptions with the relevant parts of unseen images and suppressing text noise in unseen information without any text annotation. Zhu et al. [20] used GAN to form a loop network, and introduced cycle consistency loss to generate features for adapting to another domain space, so that the generated features are retained in the original distribution of ZSC, namely, most of the content information of the original image is retained. Similarly, this study utilizes a GAN for description-based ZSC. However, in contrast to [20], a text-reconstruction network is additionally incorporated into the GAN framework to suppress noisy signals present in descriptions.

2.2 Generative zero-shot classification

The GAN model firstly proposed in 2014 [21] suffers from unstable training due to the Jensen-Shannon divergence between the generated and true distribution during the training process. Many attempts have been made to address this problem [22–24]. For example, Arjovsky et al. [24] proposed a Wasserstein-GAN (WGAN) model utilizing the Wasserstein distance, while Gulrajani et al. [23] further developed this model by introducing a gradient penalty based on the Lipschitz constraint. However, both models do not consider image features. In contrast, Xian et al. [25] utilized three conditional GANs combining the f-GAN with the Wasserstein GAN and classification loss to generate embedding features step by step and employ them in training a good classifier to achieve ZSC. Schonfeld et al. [26] used two variational autoencoders with the same structure, one encoding images, while the other decoding class embeddings, ensuring consistency of the space, onto which the images and class embeddings are projected in ZSC. Unlike the above approaches, this study tackles ZSC by generating features for unseen classes via a novel GAN model. The proposed model combines a GAN with a discriminator to employ unlabeled data of unseen classes in generating discriminative features.

Many existing generative models [25, 27, 28] have been applied to solve ZSC by generating features of unseen classes from semantic embeddings. For example, Bucher et al. [27] employed a generative moment matching network [29], while [25, 28] used GANs. This study proposes a novel triple discriminator GAN architecture to directly generate features from text description. In particular, the proposed architecture combines the powerful D2GAN [12] with a text-reconstruction network and a classification loss function to achieve effective ZSC.

3 TDGAN framework for zero-shot classification

The main notations are listed in Table 1. The proposed TDGAN framework employs two networks, a D2GAN and a text-reconstruction network, which exploit global and local semantic features to synthesize images, respectively (Figure 2). In particular, the D2GAN is used to train a stable and diverse generator with a dual discriminator GAN structure, while the text-reconstruction network applies another GAN with discriminative local text knowledge. In this way, TDGAN realizes an accurate cross-modal mapping from text descriptions to their visual representations. Training is performed using labeled seen data

Table 1 Main notations

Symbol	Meaning
N	Number of instances
s	Number of seen categories
u	Number of unseen categories
V	Dimensionality of visual space
Q	Dimensionality of textual space
M	Dimensionality of noise
$x \in \mathbb{R}^V$	Visual representation vector
$t \in \mathbb{R}^Q$	Textual representation vector
$y \in \mathbb{R}^{s+u}$	Label vector
$z \in \mathbb{R}^M$	Noise representation vector
θ	Generator network parameters
ω_1, ω_2	Dual discriminator network parameters

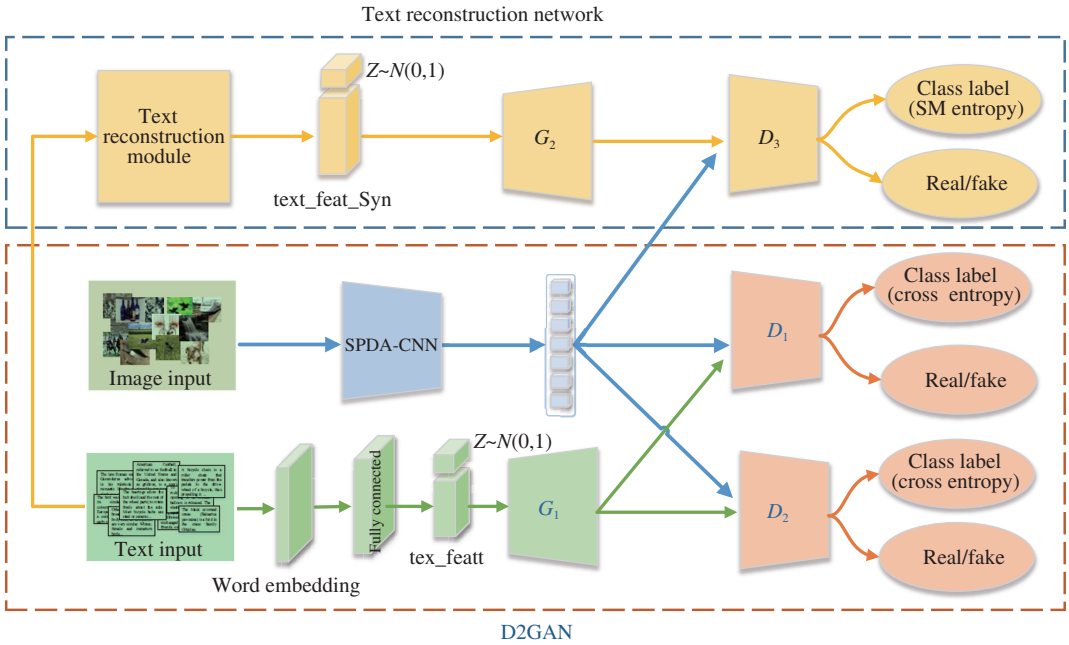

Figure 2 (Color online) Overall framework of the proposed TDGAN model, where the arrows indicate data flow. It consists of two parts: D2GAN network and text reconstruction network.

(visual images and their corresponding text descriptions). At the testing stage (i.e., classification of unseen images), unseen images are recognized by training a classifier with the synthesized data.

3.1 Dual discriminator generative adversarial networks

Compared with the traditional GAN model, the advantage of the D2GAN [12] is its employment of one generator and two discriminators formulating a mini-max game of three players. The D2GAN structure has three main modules (Figure 3): (1) generative network G_1; (2) discriminator network D_1; and (3) discriminator network D_2.

The generator G_1 inputs the prior noise $z \sim p_z(z)$ (a ramdom noise vector following a standard normal distribution) and text feature $t \sim p_t(t)$ (represented by term frequency-inverse document frequency (TF-IDF)) to generate synthetic data for the corresponding class. The discriminators D_1 and D_2 output probability values for the considered classes. There are two types of input data for the discriminators D_1 and D_2, namely, labeled real image data and image data generated for from G_1. Real image data are considered by the SPDA-CNN [30], which extracts fine-grained visual features. The purpose of employing the two discriminators is to utilize the complementary statistical properties of two divergences in elevating the quality and diversity of samples output by the generator G_1. In particular, D_1 provides high marks for the data sampled from p_{data} and low marks to the data sampled from G_1; D_2 acts in the opposite

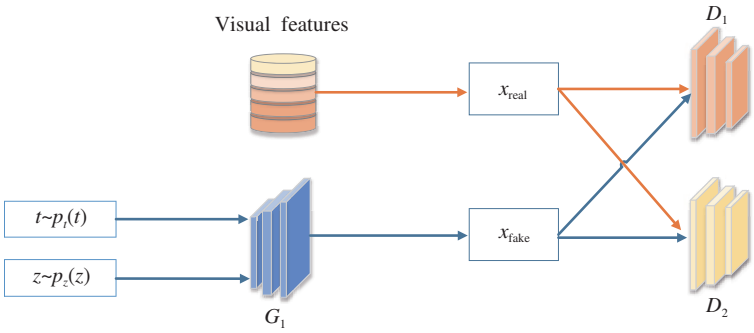

Figure 3 (Color online) Illustration of the D2GAN structure.

way. The generator G_1 generates data to fool both D_1 and D_2.

The design of D_1 is the same to that in the traditional GAN, whose loss function is

$$\begin{aligned} L_{D_1} &= \min_{G_1} \max_{D_1} v(D_1, G_1) \\ &= \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D_1(x)] + \mathbb{E}_{z \sim p_z(z), t \sim p_t(t)}[\log(1 - D_1(G_1(z,t)))], \end{aligned} \tag{1}$$

where $v(D_1, G_1)$ represents the divergence between the D_1 distribution $p_{\text{data}}(x)$ and the G_1 distribution produced by $p_z(z)$ and $p_t(t)$.

The loss function of D_2 is designed as

$$\begin{aligned} L_{D_2} &= \max_{G_1, D_2} v(D_2, G_1) \\ &= \mathbb{E}_{(x,y) \sim p_{\text{data}}(x_1, y_1)}[\log D_2(y|x)] + \mathbb{E}_{z \sim p_z(z), t \sim p_t(t)}[\log(D_2(t|G_1(z,t)))] - L_c, \end{aligned} \tag{2}$$

where L_c is applied to supervised G_1 to generate the corresponding label for category c. Its loss function is defined as

$$\begin{aligned} L_c &= \mathbb{E}_{x \sim p_{\text{data}}(x)} S[D_2(n|x)] \\ &= \frac{1}{N} \sum_{n=1}^{N} S[D_2(n|x)] \\ &= \frac{1}{N} \sum_{n=1}^{N} \left[- \sum_{n=0}^{c} D_2(n|x) \log D_2(n|x) \right], \end{aligned} \tag{3}$$

where $S[D_2(n|x)]$ represents Shannon entropy of the discriminator D_2, and n indicates the number of instances per category.

Combining the three parts of D_1, D_2, and G_1, the objective function is

$$\min_{G_1, D_2} \max_{D_1} v(D_1, D_2, G_1) = L_{D_1} - L_{D_2}. \tag{4}$$

Note that there is a fully connected layer behind the D_1 and D_2, respectively. It predicts the class information with cross entropy loss function. We naturally adopt the Lipschitz constraint to avoid the gradient disappearing problem, so that the objective function is

$$L_{\text{D2GAN}} = L_{D_1} - L_{D_2} + L_{\text{GP}}, \tag{5}$$

where $L_{\text{GP}} = (\|\nabla_{\hat{X}} D_\omega(\hat{X})\|_2^2 - 1)^2$ is the gradient penalty to enforce the Lipschitz constraint [23] with $\hat{X}$ being the linear interpolation of the real feature X and the synthetic feature $\tilde{X}$, D_ω is a discriminator for the optimization parameter ω.

3.2 Text-reconstruction network

A text-reconstruction network comprising a text-reconstruction module (Figure 4) and a GAN module is employed in this study to extract useful text semantic features and match them to image features. The

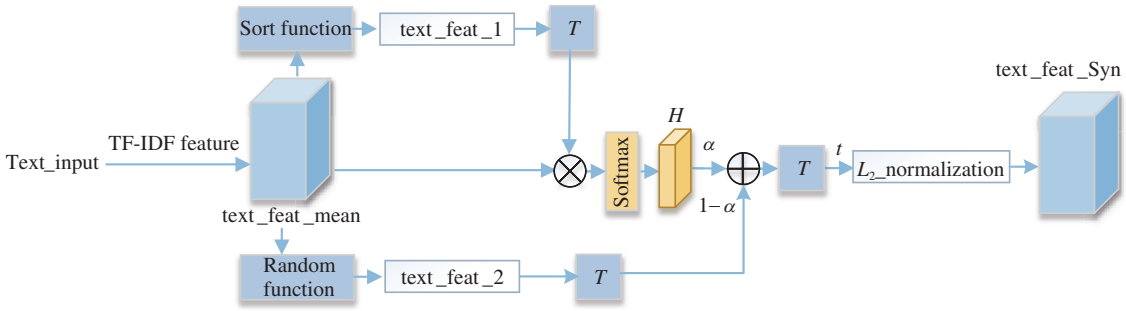

Figure 4 (Color online) The proposed text reconstruction module.

text-reconstruction module pays partial attention to texts, thus helping to extract local discriminative features. In essence, the module combines global text features with local key text features, with local text features being used to distinguish the differences across the considered classes, while global text features being used to make up for the missing information captures by local key features.

The reconstruction steps include the following. The text features (the TF-IDF feature vector [31]) is first averaged to obtain the mean text feature, text_feat_mean, by averaging the features in the same class. Then, two text features are extracted to obtain text_feat_1 and text_feat_2. Our motivation is to highlight discriminative text words. Considering that each word makes a different contribution to each class, a simple sorting method is applied to weight each word. Intuitively, the larger is the number of features, the greater is their contribution to the classification task. Thus, features are sorted in the descending order, and the first 50 features are taken to construct a new vector text_feat_1. At the same time, 50 features are randomly sampled from the original feature set to form text_feat_2, which can be regarded as a down-sampling process. While the global representation ability is reduced, the text_feat_2 still contains global features. Both sorting processes can be viewed as simple feature selection approaches. Of course, other advanced feature selection methods can be more effective. Since feature selection is not the focus of this study, simple sorting functions are employed. L_2 normalization is applied to accelerate the convergence of the loss function after text feature scaling. Finally, the text feature text_feat_Syn is obtained using the text-reconstruction module.

In the Figure 4, T represents the transpose of the matrix. We use the specific formulas to illustrate some stages of text reconstruction. Firstly, by sorting function to get the partial important text features H is

$$H = \text{softmax}((f(T_M)^{\mathrm{T}}) \otimes T_M), \tag{6}$$

where T_M indicates the average text feature text_feat_mean, $f(\cdot)$ represents a descending sorting function.

Then, we combine the original text feature information and H to form a new text feature, which is formulated as

$$t = (\alpha H + (1-\alpha)(r(T_M)^{\mathrm{T}}))^{\mathrm{T}}, \tag{7}$$

where $r(\cdot)$ represents a random sampling function, the reconstruction coefficient α is formed to synthesize a new text feature, and the important labeled information is highlighted. We set α to be 0.9.

Finally, we get new synthetic text features T_S via L_2 normalization. The synthetic text feature T_S together with the Gaussian noise z are the inputs for the generator G_2, which generates a synthesized local visual feature. It is more effective than the original visual features to assist in the classification of unseen classes in the ZSC.

The GAN network with G_2 and D_3 has two outputs, namely, the probability of real features derived from generated features and the prediction of features belonging to their corresponding classes. To assign feature weight constraints to important parts in the reconstructed text and enhance knowledge transfer from seen classes to unseen classes in ZSC, SM entropy is employed instead of cross entropy as the loss function, which is defined as

$$S_{\text{cls-f}} = \sum_{c=1}^{C} p_i^r \left(\frac{p_i^k - p_j^{-k}}{2k} \right), \tag{8}$$

where k and r are the two parameters in the SM entropy. In our experiment, we set $k = 0.1$, $r = 0.01$, and use it to measure the information distribution of each category. The p_i and p_j represent the probability

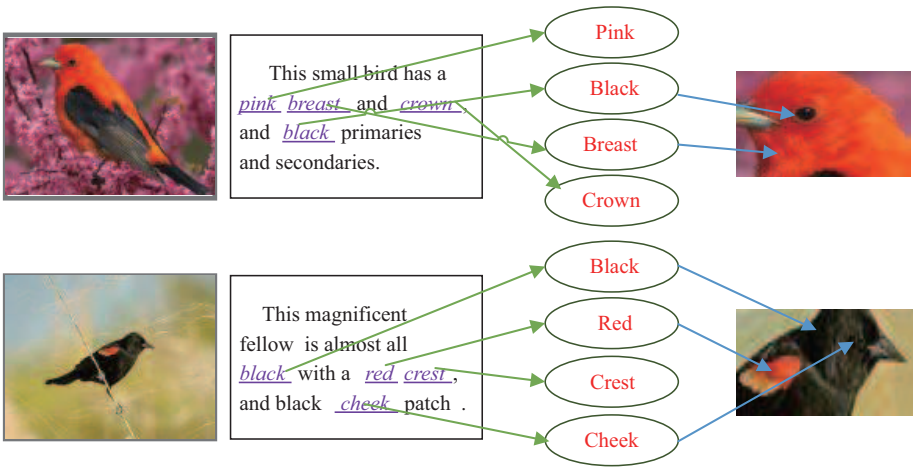

Figure 5 (Color online) The text reconstruction network focuses on important textual information, which is marked in red. The information includes color (pink, black, red and white), parts (breast, crest and patch) and textures.

distribution of categories, respectively. In this way, the loss function of D_3 is

$$L_{D_3} = -\lambda_{\mathrm{SM}} S_{\text{cls-f}} - \frac{1}{C}\sum_{c=1}^{C} D_3\big(\mathbb{E}_{\tilde{X}_c \sim p_g^c}[\tilde{X}_c]\big), \tag{9}$$

where $\tilde{X}_c$ is the synthetic feature of the category, D_3 is the discriminator for text reconstruction network. p_g^c is the conditional distributions of the generated features. The λ_{SM} is SM entropy weight hyperparameter.

As shown in Figure 5, the network could focus on more local discriminative regions, and reduces uncorrelated text noise, thereby making the synthesized visual features more representative.

3.3 Final TDGAN model

The proposed method generates visual data from text to assist the classifier in learning the information of unseen classes. However, the distribution of generated data may be scattered. To alleviate this problem, the following constraint is imposed on generated features to make them be as close to the center of the real class as possible

$$L_E = \frac{1}{C}\sum_{c=1}^{C} \left\| \mathbb{E}_{\tilde{X}_c \sim p_g^c}[\tilde{X}_c] - \mathbb{E}_{X_c \sim p_{\text{data}}^c}[X_c] \right\|^2, \tag{10}$$

where c is the number of seen categories, X_c is the mean of the visual feature of category c, $\tilde{X}_c$ is the mean of the generated feature of category c, p_g^c and p_{data}^c are the conditional distributions of the generated features and the real features, respectively.

Further, we design a classification loss function to optimize the discriminators D_1, D_2, and D_3, which is as follows:

$$L_{\text{cls}} = \frac{1}{2}(L_c(G_{\theta_1,\theta_2}(T,z)) + L_c(x)), \tag{11}$$

where G_{θ_1,θ_2} is the optimization parameter corresponding to the two generators G_1 and G_2.

The final objective function of TDGAN is

$$\text{loss}_{\text{TDGAN}} = L_{\text{D2GAN}} + L_{D_3} + L_{\text{cls}} + L_E. \tag{12}$$

In the training stage of TDGAN, the D2GAN and text-reconstruction networks are trained separately in the end-to-end manner, while the parameters in the visual features branch are shared. The D2GAN is trained by alternatively updating D_1, D_2, and G_1.

Table 2 The settings of CUB and NABirds datasets

Dataset	SCS-split		SCE-split	
	Train	Test	Train	Test
NABirds [14]	323	81	323	81
CUB [13]	150	50	160	40

3.4 Applying TDGAN to zero-shot classification

Once the TDGAN model is trained, G_1 and G_2 can generate pseudo visual features for each unseen class using corresponding text descriptions, which can be viewed as global and fine-grained features, respectively. These features can be either combined in a weighted manner or employed separately. Given similar results produced by the two methods, the features were employed separately in the experiments presented below.

Since many pseudo visual features with different noise inputs can be generated, they can be combined with other samples in the training data to train any new classifier. In this way, zero-shot learning can be considered as a traditional classification task. Optional classifiers include softmax, support vector machine, and k-nearest neighbor (k-NN). For fair comparison, a simple k-NN classifier is employed for testing.

4 Experiments

4.1 Datasets and settings

Datasets. Following [28], two bird datasets were employed in the experiments, namely, CUB [13] and NABirds [14], since only these two datasets have text descriptions. Both datasets are fine-grained, with the challenges of imperceptible inter-class variance and large intra-class variance.

Two types of splitting schemes in terms of how close seen classes are to unseen classes, namely, super-category-shared splitting (SCS) and super-category-exclusive splitting (SCE) [28], were employed to split the datasets into training and test set. For the SCS setting, the parent class of unseen classes are disjoint from those of seen classes, which is the conventional ZSL split setting for ZSC datasets. In contrast, one or more seen classes belonging to the same parent class of each unseen class can exist in the SCE setting. This division minimizes the correlation between seen and unseen classes, which poses a bigger challenge compared to the SCS setting. For brevity, the two settings are denoted as SCS-split and SCE-split, respectively.

The CUB dataset contains 200 bird classes with a total of 11788 images, each with 312 attributes. Under SCS-split, the same split ratio was employed as in [28], with 150 classes allocated for training and 50 disjoint classes allocated for testing. Under SCE-split, the splitting method used in [15] was employed in this study, where the parent classes of unseen classes are exclusive to those of seen classes. Compared to SCS-split, the relevance between seen and unseen classes in SCE-split is minimized, which brings more challenges for knowledge transfer. Under SCE-split, 160 classes were used for training and 40 classes were used for testing. The NABirds dataset is much larger than the CUB dataset; it contains 1011 classes with a total of 48562 images. In the NABirds dataset, a hierarchy of birds can be found, having a root class with 555 leaf nodes and 456 parent nodes. After merging the leaf node classes into their parent class similar to [19], 404 classes were obtained. For SCS-split, 20% of subclasses were randomly selected as unseen classes under each parent class. For SCE-split, 20% of parent classes were randomly selected, and all their-descendant classes were considered as unseen classes. In both cases, a total of 323 classes were selected as seen (training set), while the remaining 81 classes were selected as unseen (test set). Table 2 lists the numbers of classes included into the training and test sets for CUB and NABirds under the two different settings.

Textual representation. Original Wikipedia articles as described in [19] were used for both benchmark datasets. First, the articles were marked as words without considering stop words. Second, the TF-IDF feature vectors [31] were extracted, whose dimensionalities for CUB and NABirds were 7551 and 13217, respectively.

Visual representation. The SPDA-CNN [30] was employed to extract local image features and detect important regions, followed by a sub-network using 3×3 region of interests (ROIs) to pool a region for a 512-dimensional feature . For the CUB dataset, the following seven local regions were extracted

Table 3 Top-1 average accuracy (%) of traditional ZSC for CUB and NABirds datasets under two split settings[a)]

Dataset	CUB		NABirds	
	SCS	SCE	SCS	SCE
MCZSL [32]	34.7	–	–	–
$\text{WAC}_{\text{linear}}$ [15]	27.0	5.0	–	–
$\text{WAC}_{\text{kernel}}$ [15]	33.5	7.7	11.4	6.0
ESZSL [33]	28.5	7.4	24.3	6.3
SJE [34]	29.9	–	–	–
ZSLNS [18]	29.1	7.3	24.5	6.8
$\text{SynC}_{\text{fast}}$ [10]	28.0	8.6	18.4	3.8
SynC_{OVO} [10]	12.5	5.9	–	–
ZSLPP [19]	37.2	9.7	30.3	8.1
GAZSL [28]	43.7	10.3	35.6	8.6
TDGAN (ours)	**44.2**	**12.5**	**36.7**	**9.6**

a) The best results are annotated as bold font.

to represent each CUB image: "head", "back", "belly", "breast", "leg", "wing", and "tail". For the NABirds dataset, no "leg" annotations were available; hence only the features of the remaining six visual regions were extracted to represent local visual features. Thus, the visual feature dimensionalities of CUB and NABirds were 3584 and 3072, respectively.

Implementation details. TDGAN was implemented using Torch. The batch size was set to 512, while the learning rate was set to 0.0001. The Adam optimizer was used for optimization. The performance of TDGAN was measured in terms of top-one accuracy over the test images for each value of n denoting the number of novel examples per class. The results were averaged over five runs, each of which employed a different random sample of new examples during the zero-shot training phase. The testing was performed over 2000 iterations.

4.2 Traditional zero-shot classification

We compare our algorithm with 8 state-of-the-art methods, including ZSLPP [19], MCZSL [32], ZSLNS [18], ESZSL [33], SJE [34], WAC [15], SynC [10], and GAZSL [28], and report the per-class average top-1 accuracies, which is defined as follows:

$$\text{Accy}_s = \frac{1}{\|s\|}\sum_{c=1}^{\|s\|}\frac{\text{correction prediction in } c}{\text{samples in } c}, \tag{13}$$

where s indicates the number of unseen classes, c is the corresponding category. The results are show in Table 3. It should be noted that SynC [10] and GAZSL [28] are attribute-based methods and we apply textual features to replace attributes for fair comparison. The performance results of the selected approaches are all from the original papers [28].

It can be noticed from Table 3 that the proposed TDGAN method outperforms all the eight state-of-the-art methods on both the CUB and NABirds datasets. In particular, TDGAN outperforms the second best approach (GAZSL) by 0.5% and 2.2% on CUB, as well as 1.1% and 1.0% on NABirds for the SCS and SCE-split settings, respectively. Since both TDGAN and GAZSL employ GANs, the improvement can be attributed to the text-reconstruction network employed in TDGAN. In addition, the performances under SCE-split are far inferior to those under SCS-split, which demonstrates the challenges of SCE-split.

Ablation studies. Ablation studies were conducted to explore the impacts of the D2GAN and text-reconstruction network, denoted as D2ZSL and TRZSL, respectively. GAZSL [28] was chosen as the baseline in these experiments. According to the results (Table 4), D2ZSL achieved accuracy scores of 44.2% and 36.7% under the SCS setting on the CUB and NABirds datasets, respectively, which is an improvement over the baseline. However, there was no significant improvement under the SCE setting. This is mainly because while D2ZSL generates different embeddings through semantic text descriptions, it increases the diversity of generated samples to a certain range using Lipschitz regularization. It realizes global semantic information mining and improves the classification performance. However, since the center points between classes are originally far apart under the SCE setting, semantic information contributes less to the classification result, and thus, the performance does not improve significantly.

Table 4 Ablation studies (%) of different components of the method on both CUB and NABirds datasets[a)]

Method	CUB		NABirds	
	SCS	SCE	SCS	SCE
GAZSL [28]	43.7	10.3	35.6	8.6
D2ZSL	44.0	10.5	36.0	8.7
TRZSL	43.8	11.5	35.8	9.0
TDGAN (ours)	**44.2**	**12.5**	**36.7**	**9.6**

a) The best results are annotated as bold font.

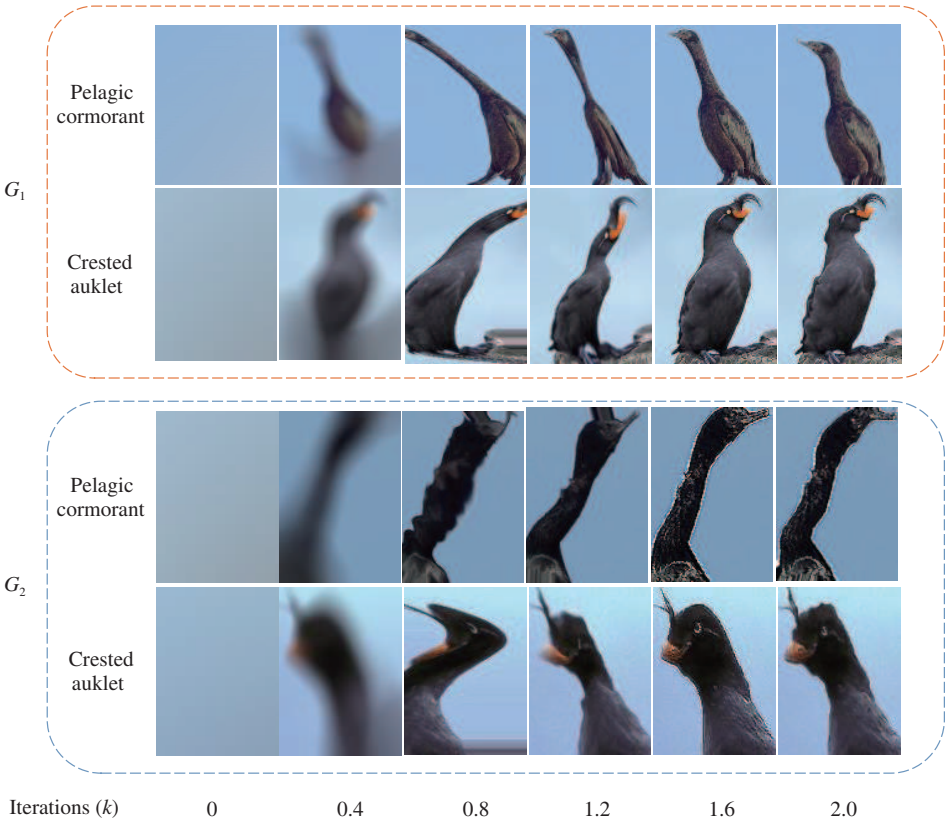

Figure 6 (Color online) The visualization of the outputs by the G_1 and G_2 in our TDGAN model by employing the two birds "pelagic cormorant" and "crested auklet" as examples. We could observe that the G_1 embodies the global information, while the G_2 captures the fine-grained local information.

The TRZSL achieved accuracy scores of 43.8% and 35.8% under the SCS setting on the CUB and NABirds datasets, respectively, which does not indicate any obvious improvement against GAZSL. In contrast, the performance of TRZSL improved by 1.2% and 0.4% under the SCE setting compared to GAZSL on the CUB and NABirds datasets, respectively. This is because local semantic information is more recognizable than global semantic information. The text-reconstruction network was employed to describe the key text so as to better highlight local key semantic information and get a better classification effect. However, the designed text-reconstruction network may pay attention to some local information that is not relatively important, thus introducing text noise and generating irrelevant samples, which can negatively affect the classification performance. Under the SCS setting, the center points of different classes are relatively concentrated, and there is a large overlap of their semantic information, which undermines the positive role of the text-reconstruction networks. In contrast, the center points of different classes are far apart under the SCE setting, and the key local information can alleviate the semantic gap between unseen and seen classes. Since the proposed TDGAN approach draws on the advantages of the two variants, its performance is further improved.

Figure 6 illustrates some qualitative results of the generators G_1 and G_2, where G_1 represents the D2GAN part of TDGAN, which explicitly focuses on holistic global features, and G_2 corresponds to the text-reconstruction part, which captures important local features. As the number of iterations increases

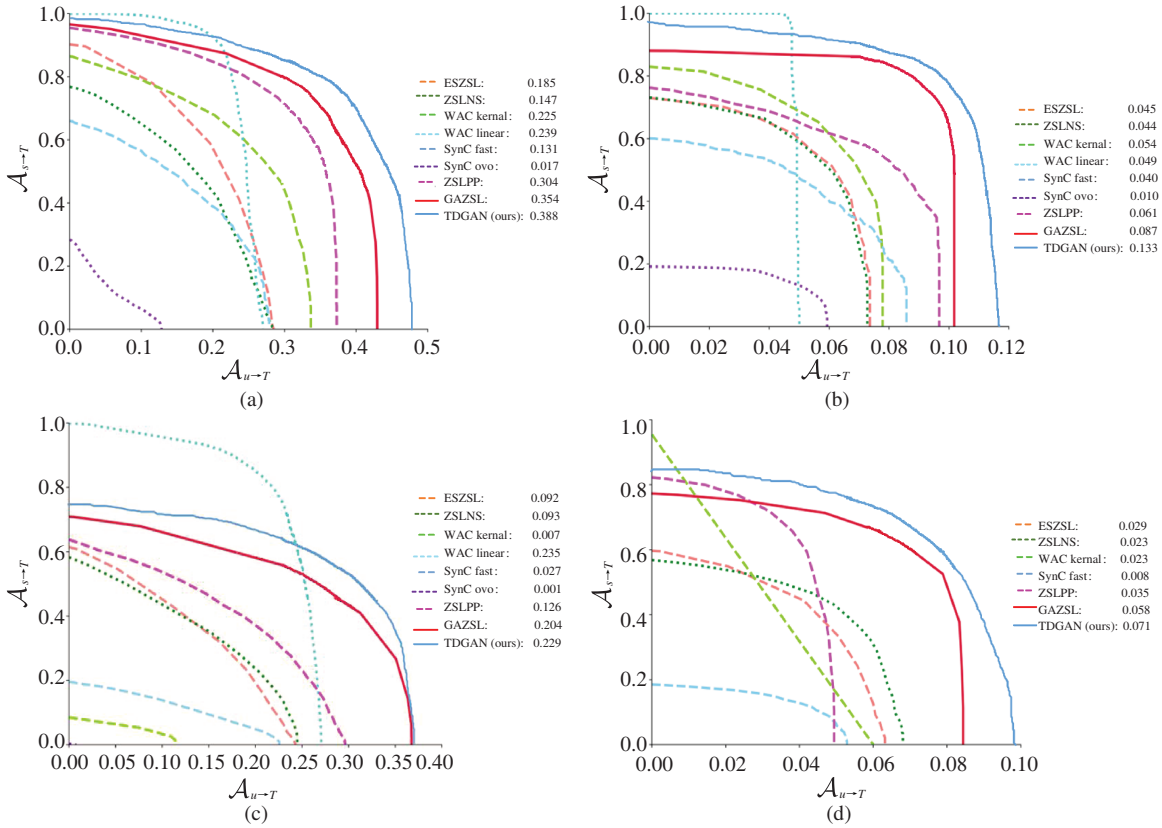

Figure 7 (Color online) The seen-unseen precision curves for each algorithm on two benchmark datasets with two split settings. (a) CUB with SCS splitting; (b) CUB with SCE splitting; (c) NAB with SCS splitting; (d) NAB with SCE splitting.

from 0 to 2000 and the generated results are output every 400 iterations, the images generated by G_1 and G_2 gradually change from blurry to relatively clear. Moreover, the fine-grained local images generated by G_2 are relatively more effective in representing the important information of the classes compared to the images generated by G_1.

4.3 Generalized zero-shot classification

Compared to ZSC, generalized ZSC (GZSC), where test data include examples from both seen and unseen classes, has proven to be more practical in real applications [26, 35]. Chao et al. [35] proposed a GZSL measurement method, which classifies a seen class S and an unseen class U into the total classes $T = S \bigcup U$, where the corresponding precisions are denoted as $A_{S\to T}$ and $A_{U\to T}$, respectively. This protocol was adopted in this study along with the seen-unseen accuracy curve (SUC) and area under SUC (AUSUC) to evaluate the knowledge transfer ability of the tested methods in GZSC.

Figure 7 shows the SUCs for all the tested methods under the considered settings. The proposed approach is superior to all the other methods in terms of the AUSUC scores, except for the $\text{WAC}_{\text{linear}}$ approach, which has significantly higher $A_{S\to T}$ and lower $A_{U\to T}$. This result indicates that TDGAN did not fully learn the knowledge transferred from seen to unseen classes and suffered from overfitting over both datasets. It is worth noting that GAZSL achieved a higher AUSUC score compared to the rest of the tested methods. However, the proposed TDGAN method performs better than GAZSL; In particular, its SUC illustrated in Figure 7 shows that the AUSUC scores of TDGAN on the CUB dataset under the SCS and SCE settings are higher than those of the GAZSL method by 0.034 and 0.046, respectively. Furthermore, the proposed method achieved higher AUSUC scores compared to GAZSL under the two setting for the NABirds dataset (by 0.025 and 0.013, respectively). In summary, the proposed TDGAN method is not only superior to other methods in the classification of unseen classes, but also achieves relatively high precision in the classification of seen classes. Moreover, TDGAN has a strong balanced performance on $A_{S\to T}$ and $A_{U\to T}$.

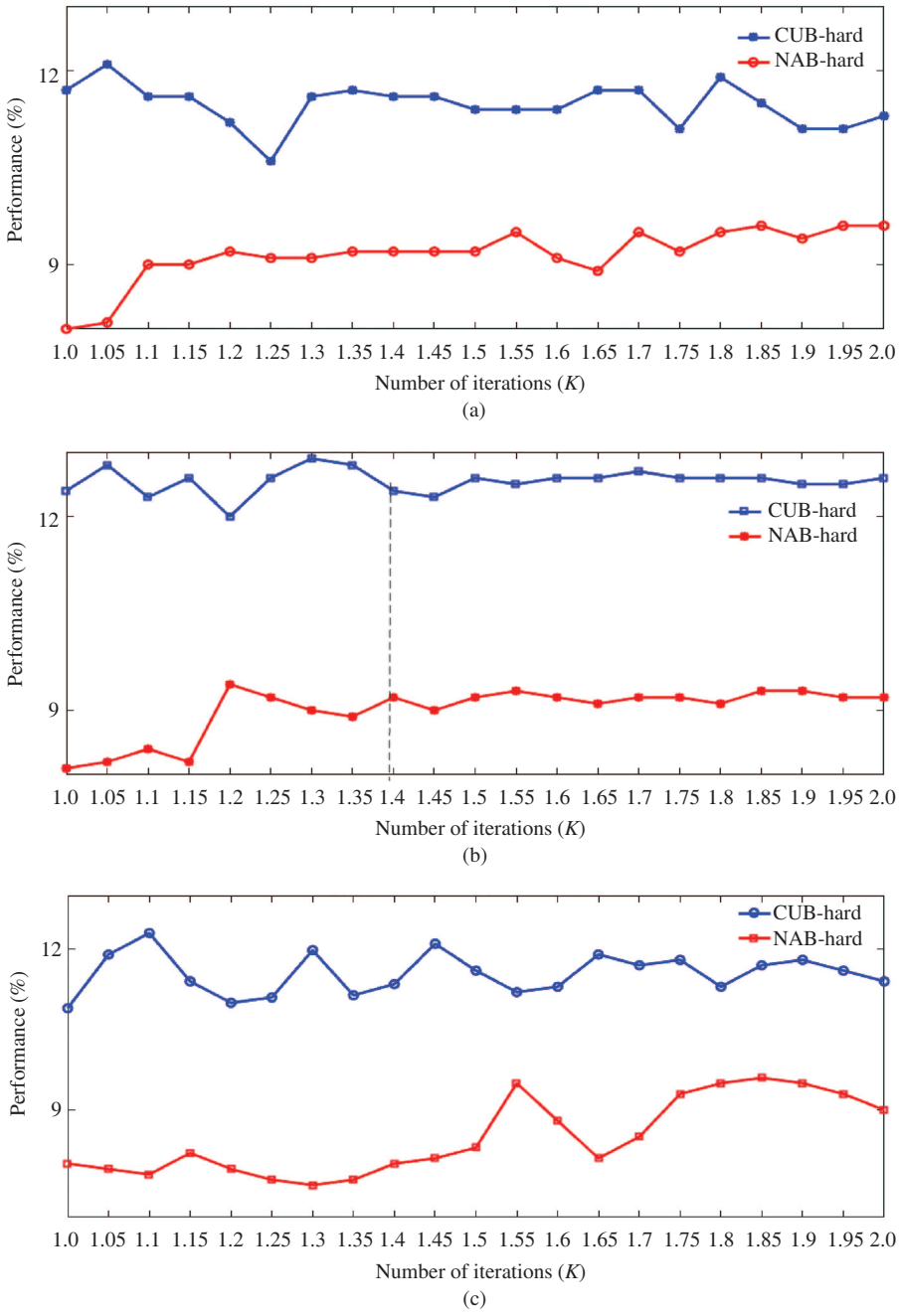

Figure 8 (Color online) The influence of different weight parameters λ_{SM} of SM entropy under SCE settings. (a) Synthetic-weight = 0.2; (b) Synthetic-weight = 0.5; (c) Synthetic-weight = 0.8.

4.4 SM entropy weight hyperparameter λ_{SM}

To check the impact of different weight parameters for SM entropy in the text-reconstruction network, further experiments were conducted under the SCE setting since semantic information is missing in this setting.

It can be noticed from Figure 8 that our TDGAN performances on both settings are relatively good when λ_{SM} is around 0.5; too small or too large λ_{SM} causes the performance degradation. This indicates that the text-reconstruction parameters are beneficial for different levels of key features of different classes, and information is transferred from the seen to unseen classes. Furthermore, it can be noticed from the figure that the performance is relatively stable when the number of iterations reaches around 1400. In the other words, the objective function of the TDGAN model is minimized and reaches a local optimum.

5 Conclusion

This paper presented a new ZSC method performing the generation task to compensate for data loss. The method incorporates a novel triple discriminator GAN to synthesize new data samples, along with a text-reconstruction network designed to distribute local attention to different text parts, thereby highlighting key text description information and deeply mining semantic relevance. Furthermore, a SM entropy weight constraint loss function is introduced to make the prediction class distribution of the reconstructed text feature synthesis be consistent with the original real class distribution. The proposed method achieved a better performance compared to several state-of-the-art methods on two benchmark datasets, especially under the SCE setting. It is worth noting that the text reconstructed by the proposed method allows capturing key information of text descriptions, compensating for the interference caused by redundant or irrelevant information and noise. In the future, we plan to apply the attention mechanism [36] to improve the accuracy of the text-reconstruction network in the key information extraction [37].

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61771329, 61632018).

References

1 Fu Y W, Xiang T, Jiang Y G, et al. Recent advances in zero-shot recognition: toward data-efficient understanding of visual content. IEEE Signal Process Mag, 2018, 35: 112–125
2 Guo G J, Wang H Z, Yan Y, et al. Large margin deep embedding for aesthetic image classification. Sci China Inf Sci, 2020, 63: 119101
3 Zhu X X, Anguelov D, Ramanan D. Capturing long-tail distributions of object subcategories. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2014. 915–922
4 Guo Y C, Ding G G, Han J G, et al. Synthesizing samples for zero-shot learning. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017. 1774–1780
5 Ji Z, Sun Y X, Yu Y, et al. Attribute-guided network for cross-modal zero-shot hashing. IEEE Trans Neural Netw Learn Syst, 2020, 31: 321–330
6 Long Y, Liu L, Shao L, et al. From zero-shot learning to conventional supervised classification: unseen visual data synthesis. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 1627–1636
7 Yu Y L, Ji Z, Fu Y W, et al. Stacked semantics-guided attention model for fine-grained zero-shot learning. In: Proceedings of Advances in Neural Information Processing Systems, 2018. 5995–6004
8 Akata Z, Perronnin F, Harchaoui Z, et al. Label embedding for attribute-based classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013. 819–826
9 Akata Z, Perronnin F, Harchaoui Z, et al. Label-embedding for image classification. IEEE Trans Pattern Anal Mach Intell, 2016, 38: 1425–1438
10 Changpinyo S, Chao W L, Gong B Q, et al. Synthesized classifiers for zero-shot learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 5327–5336
11 Wang X Y, Ji Q. A unified probabilistic approach modeling relationships between attributes and objects. In: Proceedings of IEEE International Conference on Computer Vision, 2013. 2120–2127
12 Nguyen T D, Le T, Vu H, et al. Dual discriminator generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 2670–2680
13 Wah C, Branson S, Welinder P, et al. The caltech-ucsd birds-200-2011 dataset. 2011. http://www.vision.caltech.edu/visipedia/CUB-200-2011.html
14 van Horn G, Branson S, Farrell R, et al. Building a bird recognition app and large scale dataset with citizen scientists: the fine print in fine-grained dataset collection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015. 595–604
15 Elhoseiny M, Saleh B, Elgammal A. Write a classifier: zero-shot learning using purely textual descriptions. In: Proceedings of IEEE International Conference on Computer Vision, 2013. 2584–2591
16 Ba J L, Swersky K, Fidler S. Predicting deep zero-shot convolutional neural networks using textual descriptions. In: Proceedings of IEEE International Conference on Computer Vision, 2015. 4247–4255
17 Reed S, Akata Z, Lee H, et al. Learning deep representations of fine-grained visual descriptions. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 49–58
18 Qiao R Z, Liu L Q, Shen C H, et al. Less is more: zero-shot learning from online textual documents with noise suppression. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2249–2257
19 Elhoseiny M, Zhu Y Z, Zhang H, et al. Link the head to the "beak": zero shot learning from noisy text description at part precision. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 6288–6297
20 Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of IEEE International Conference on Computer Vision, 2017. 2223–2232
21 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems, 2014. 2672–2680
22 Martin A, Bottou L. Towards principled methods for training generative adversarial networks. 2017. ArXiv:1701.04862
23 Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 5767–5777
24 Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning, 2017. 214–223
25 Xian Y Q, Lorenz T, Schiele B, et al. Feature generating networks for zero-shot learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 5542–5551
26 Schonfeld E, Ebrahimi S, Sinha S, et al. Generalized zero-and few-shot learning via aligned variational autoencoders. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 8247–8255

27 Bucher M, Herbin S, Jurie F. Generating visual representations for zero-shot classification. In: Proceedings of IEEE International Conference on Computer Vision, 2017. 2666–2673
28 Zhu Y Z, Elhoseiny M, Liu B C, et al. A generative adversarial approach for zero-shot learning from noisy texts. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 1004–1013
29 Li Y J, Swersky K, Zemel R. Generative moment matching networks. In: Proceedings of International Conference on Machine Learning, 2015. 1718–1727
30 Zhang H, Xu T, Elhoseiny M, et al. SPDA-CNN: unifying semantic part detection and abstraction for fine-grained recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 1143–1152
31 Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manage, 1988, 24: 513–523
32 Akata Z, Malinowski M, Fritz M, et al. Multi-cue zero-shot learning with strong supervision. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 59–68
33 Romera-Paredes B, Torr P. An embarrassingly simple approach to zero-shot learning. In: Proceedings of International Conference on Machine Learning, 2015. 2152–2161
34 Akata Z, Reed S, Walter D, et al. Evaluation of output embeddings for fine-grained image classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015. 2927–2936
35 Chao W L, Changpinyo S, Gong B, et al. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In: Proceedings of European Conference on Computer Vision, 2016. 52–68
36 Ji Z, Xiong K L, Pang Y W, et al. Video summarization with attention-based encoder-decoder networks. IEEE Trans Circ Syst Video Technol, 2020, 30: 1709–1717
37 Wang Z H, Liu X, Lin J W, et al. Multi-attention based cross-domain beauty product image retrieval. Sci China Inf Sci, 2020, 63: 120112