Front.Comput.Sci.
DOI

RESEARCH ARTICLE

CRD-CGAN: Category-Consistent and Relativistic Constraints for Diverse Text-to-Image Generation

Tao Hu [1,2,3], **Chengjiang Long** [4], **Chunxia Xiao** (✉)[2]

1 College of Intelligent Systems Science and Engineering, Hubei Minzu University,Enshi 445000, China
2 School of Computer Science, Wuhan University, Wuhan 430072, China
3 Key Laboratory of Performing Art Equipment & System Technology, Ministry of Culture and Tourism, Beijing 100007, China
4 Meta Reality Labs, Burlingame, CA, 94010, USA

Abstract Generating photo-realistic images from a text description is a challenging problem in computer vision. Previous works have shown promising performance to generate synthetic images conditional on text by Generative Adversarial Networks (GANs). In this paper, we focus on the category-consistent and relativistic diverse constraints to optimize the diversity of synthetic images. Based on those constraints, a category-consistent and relativistic diverse conditional GAN (CRD-CGAN) is proposed to synthesize K photo-realistic images simultaneously. We use the attention loss and diversity loss to improve the sensitivity of the GAN to word attention and noises. Then, we employ the relativistic conditional loss to estimate the probability of relatively real or fake for synthetic images, which can improve the performance of basic conditional loss. Finally, we introduce a category-consistent loss to alleviate the over-category issues between K synthetic images. We evaluate our approach using the Birds-200-2011, Oxford-102 flower and MSCOCO 2014 datasets, and the extensive experiments demonstrate superiority of the proposed method in comparison with state-of-the-art methods in terms of photorealistic and diversity of the generated synthetic images.

Keywords Text-to-Image, diverse conditional GAN, relativistic category-consistent.

Received month dd, yyyy; accepted month dd, yyyy

E-mail: *cxxiao@whu.edu.cn*

1 Introduction

Text-to-image generation has wide range of applications in computer vision and graphics [1–9], and many methods have been proposed for this research topic. Various conditional Generative Adversarial Networks (GANs) [1, 10–33] have been developed to generate photo-realistic images conditional on text with a random noise. However, it is still challenging to simultaneously generate a set of photo-realistic as well as significantly diverse synthetic images conditional on text description.

Existing text-to-image GANs mainly focus on improve the synthesizing performance to generate high-quality and resolution images by tree-liked stacked GANs [15, 16], word-region attention guided GANs [17, 18, 34], object-driven attentive GANs [24], or the mode seeking GANs [19,35]. However, all those methods ignore the category attributes of the real image corresponding to the text description. Such as Liu *et al.* [36] used the category information to generate text. It means that we expect the synthetic images generated by GANs should have category attributes corresponding to real images. In other words, we expect the synthetic images have the main visual feature of the same category.

When given a text description (middle), the synthetic images generated by StackGAN++ [16] (in the blue rectangle in Fig. 1) are highly realistic. But there is a clear visual difference between synthetic and real images. The color of *"American_Crow"* bird category is black, but the color of synthetic

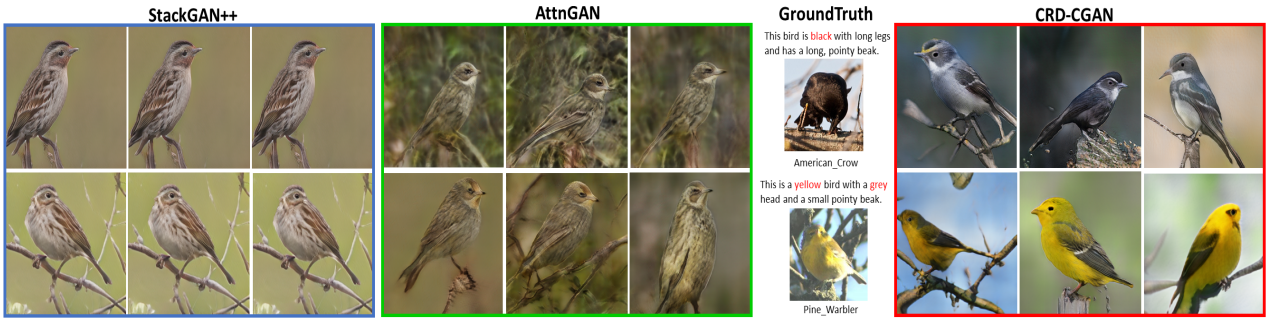

Fig. 1 Illustration of three methods to generate a synthetic images conditional on a text. Our goal is to generate a set of diverse and high-quality synthetic images that are as consistent as possible with the text description and consistent with the category visual feature of real image category.

bird generated by StackGAN++ is grey. The same problem appears on the *"Pine_Warbler"* category. The synthetic images generated by AttnGAN [18] also have not consistent of color with the real image, which are shown in green rectangle in Fig. 1. And there are less diversity between them. Therefore, it is desirable to ensure the synthetic images to retain the main visual feature of real image, as well as preserving the category-consistent visual concept and keeping diversity.

To address this issue, we propose a novel diverse photo-realistic and category-consistency text-to-image generation method that effectively exploits the relative relationship between synthetic and real images, and the category information of the real images correspondingly within the generation procedure, named as category-consistent and relativistic diverse conditional GAN (CRD-CGAN). Inspired by the advantages of tree-liked stacked GANs [1, 15, 16, 18, 19, 37], we propose a basic diverse conditional GAN (D-CGAN) to generate K synthetic images with K different generators and one shared discriminator firstly. Then we propose a relativistic discrimination regularization to improve the estimation performance of synthetic image is real, which can effectively generate more diverse K photo-realistic images. To ensure that the K synthetic images have the main visual feature of category correspondingly, we use the category consistency regularization to constrain the main visual feature of synthetic images. From Fig. 1, we can see the synthetic images generated by CRD-CGAN retain the main visual feature of real images with a diversity of photo-realistic appearance.

To sum up, our contributions are three-fold as follows:

(1) We propose a new framework CRD-CGAN which contains K generators and a shared discriminator to improve the diversity of K high-quality synthetic images simultaneously.

(2) We incorporate category-consistent and relativistic diverse conditional constraints, which effectively improve the quality of photo-realistic synthetic images and ensure that K synthetic images retain the main visual feature of the corresponding category of real images.

(3) The proposed CRD-CGAN achieves the state-of-the-art performance on the Caltech-UCSD Birds-200-2011 dataset [38], the Oxford 102 Category Flower dataset [39] and the MS COCO 2014 dataset [40] for text-to-image generation.

2 Related Works

Generative Adversarial Networks (GANs) and attention mechanisms have been successfully applied to various visual applications [41–49], such as visual inpainting [50] and outpainting [51]. For example, Yang *et al.* [51] applied Skip Horizontal Connection and Recurrent Content Transfer in GANs to obtain essential information form regions. Zheng *et al.* [52] introduced to use unconditional GAN to generate images from random vectors for person re-identification. GD-Net [53] is designed as a single unified network for supervised person-id, which separately learns the appearance and structure codes to improve the image generation quality.

Especially, to translate the visual concepts from characters to pixels, Reed *et al.* [12] designed a novel GAN to effectively bridge text and image modeling, while the size of synthetic image is 64×64 pixels. Due to the real image distribution and GAN's distibution may not overlap in high dimensional pixel space, Zhang *et al.* [15, 16] proposed StackGAN and StackGAN++ to improve the quality of generated 256×256images by jointly approximating multiple distributions, while the multiple generators and discriminators arranged in a tree-like structure. To better generate fine-grained texture features based on the input words, Xu *et al.* [18] proposed an AttnGAN to synthesize fine-grained details at different sub-regions of the generated image by automatically attending to the relevant words with a deep attentional mul-

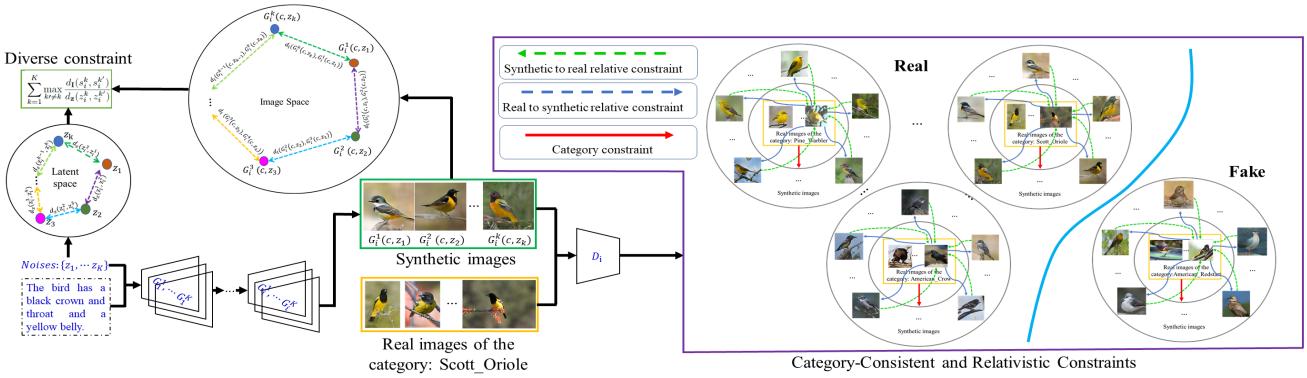

Fig. 2 Overview of our proposed framework CRD-CGAN. We use the generators $G_i^1, ..., G_i^k$ to generate K synthetic images. Based on the difference between image space and latent noise space, we use the diverse constraints to improve the diversity of K synthetic images. The proposed category-consistent and relativistic constraints are reflected in the discriminator at each stage. In this figure, the green arrow is the relativistic loss between synthetic image with the corresponding real images based on the true label, and the blue arrow is the relativistic loss between real image with the corresponding synthetic images based on the fake label. Among these four groups of images (inner circle are real images and outsider are the corresponding synthetic images), the three ground above the decision boundary have better category consistency.

timodal similarity model. And Wang *et al.* [54] proposed a bidirectional caption head to enhance the visual presentation and alleviate the cross-modal gap between video and text. Li *et al.* [24] proposed object-driven attentive GANs to synetesize objects by paying attention to the most relevant words and the pre-generated layout. simGAN [55] used an adversarial loss and a self-regularization loss to imporve the realism of synthetic images using unlabeled real data. VTN [31] learned the margins, alignments, and other global design rules to generate the layout of synthetic images based on Transformer. The above-mentioned related work are focused on generating higher resolution and realistic images conditioned on text. In contrast, our work is going to synchronously generate more diverse and photo-realistic synthetic images with the same resolution.

Some state-of-the-art GANs are designed to generate significantly diverse synthetic images by optimizing the generator or discriminators. For example, Mao *et al.* [19] proposed a mode seeking regularization method to minimize the generator's loss by maximize the ratio between the distance of two synthetic images in visual space and the distance of two noises correspondingly, and they used different noise inputs to generate diverse synthetic images through the optimized generator. To generate more diverse images, Cha *et al.* [21] used triplets (*i.e.*, a positive image, a text, and a negative-image) to train the generator and discriminator, and selected the negative image by the semantic distance from a positive example in the class. Hu *et al.* [35] used the hierarchical model in conditional GANs to imporve the diversity of synthetic images. Contrastive learning [56] is another novel method to generate scene graph, such as SMC-GAN [57], which has used several text-image contrastive losses in a one-stage GAN. However, our method focus on the relative relation between images and the category attribute of images, which is different from contrastive learning.

It is well known that the real image corresponding to the text has very specific category information. For example, each image in the Birds-200-2011 dataset [38] has a specific bird category, such as "American_Crow", and "Scott_Oriole". However, all those methods ignore the category attributes of the real image corresponding to the text description. The category attributes can effectively improve the GAN's performance, for example, Liu *et al.* [36] used the category information to generate text. It means that we expect the synthetic images generated by GANs should have the same category attributes as the corresponding real images. In other words, we expect the synthetic images have the main visual feature of the same category.

Unlike previous approaches that generate more realistic or diverse images by cyclic inputing set of noise, our proposed method extend one generator to K generators in a certain resolution. In this paper, we propose a novel GAN framework named CRD-CGAN which incorporates category-consistent and relativistic constraints for diverse image generation. It exploits the relative relationship between synthetic and real images, and the category information of the real images correspondingly in the generation procedure. Inspired by the advantages of tree-liked stacked GANs [15, 16, 18, 19, 37], we first design a basic diverse conditional GAN (D-CGAN) to generate K synthetic images with K different generators and one shared discriminator. Then we propose a relativistic discrimination constraint to improve the photo-realistic

performance of synthetic images. To ensure that the K synthetic images have the main visual feature of category correspondingly, we use the category consistency regularization to constrain the main visual features of synthetic images. From Fig. 1, we can see the synthetic images generated by CRD-CGAN retain the main visual appearance feature of real images with a diversity of photo-realistic.

3 Proposed Method

Our CRD-CGAN consists of two key components, *i.e.*, the diverse conditional GAN (D-CGAN) with attention diverse constraint. The CRD-CGAN with the discrimination regularization in combination of relativistic conditional constraint and category consistent constraint. The D-CGAN is the standard network for generating K diversity synthetic images. The CRD-CGAN use relativistic conditional regularization and category consistency regularization to improve the quality and diversity of synthetic images. As illustrated in Fig. 2, with K generators at the i-th stage, K synthetic images have been generated by our CRD-CGAN.

The K synthetic images are feed into the generators and discriminator optimization process. We firstly describe a diverse conditional GAN which is denoted as D-CGAN for diverse text-to-image generation, and then employ category-consistent and relativistic constraints to improve the quality and diversity of synthetic images. To better understand the proposed CRD-CGAN, we list the important parameter as Table 1.

Table 1 The parameters in CRD-CGAN

Notation	Meaning in CRD-CGAN
i	The stage of the tree-liked stacked GANs
c	text condition parameter
z_k	The $k-th$ noise
s_k	The $k-th$ synthetic image
X_i	Real image from distribution P_data at the stage i
X_f	Real Image
X_r	fake Image
G_i^K	The K generators at the stage i
D_i	The discriminator at the stage i
l_f, l_r	The symmetric labels $\in \{-1, 1\}$
$category_i$	The true category of X_i

3.1 Diverse Text-to-Image Generation

The D-CGAN is uesd to generate K photo-realistic synthetic images simultaneously with K generators and one shared discriminator. Given a text t, we follow the approach of Reed *et al.* [13] to calculate the text condition parameter c by using a text-image joint encoder and random Gaussian noise. The K generators $\{G_1, G_2, ..., G_K\}$ use K different prior noise vectors $\{z_1, z_2, ..., z_K\}$ with the same text condition parameter c to ensure the K synthetic images have high diversity, correspondingly. The synthetic image s_k is defined as $s_k = G_k(z_k, c)$. The discriminator D and generators $G_1, \ldots, G_K$ can be optimized in a joint form by alternatively maximizing $\mathcal{L}_D$ and minimizing $\mathcal{L}_{G_1,\ldots,G_K}$ until convergence.

AttnGAN [18] proposed a deep attentional multimodal similarity model to compute a fine-grained image-text matching loss for training the generator. In our work, we use a similar image-text similarity loss, denoted as $\mathcal{L}_{sim}(s_1, \ldots, s_K)$ to make sure that the generated synthetic images cover the text description. The $\mathcal{L}_{sim}(s_1, \ldots, s_K)$ is used to estimate the probability of the matching level between each word and a subregion of synthetic image, and the matching level between the input sentence and synthetic image. The $\mathcal{L}_{sim}(s_1, \ldots, s_K)$ is defined as:

$$\mathcal{L}_{sim}(s_1, \ldots, s_K) = \sum_{k=1}^{K} (\mathcal{L}_{1,k}^{w} + \mathcal{L}_{2,k}^{w} + \mathcal{L}_{1,k}^{s} + \mathcal{L}_{2,k}^{s}) \quad (1)$$

where $\mathcal{L}_{1,k}^{w}$ is the negative log posterior probability that measure matching level between the synthetic images and the corresponding inputting word-level description of the k-th generator, and $\mathcal{L}_{2,k}^{w}$ is the negative log posterior probability measure matching level between the word-level description with the corresponding image. Similarly, the $\mathcal{L}_{1,k}^{s}$ and $\mathcal{L}_{2,k}^{s}$ are the negative log posterior probabilities of matching level between image and inputting sentence-level description of the k-th generator. The objective function of attention loss is $min\ \mathcal{L}_{sim}(s_1, \ldots, s_K)$.

To ensure the diversity among the generated synthetic images, we introduce a diversity loss, denoted as $\mathcal{L}_{div}(s_1, \ldots, s_K)$, to measure the diversity of the K synthetic images. Inspired by [19], we define it as:

$$\mathcal{L}_{div}(s_1, \ldots, s_K) = \sum_{k=1}^{K} \max_{k' \neq k} \frac{d_{\mathbf{I}}(s_k, s_{k'})}{d_{\mathbf{z}}(z_k, z_{k'})} \quad (2)$$

where $d_{\mathbf{I}}$ is the distance between synthetic image features, and $d_{\mathbf{Z}}$ means the distance between noise vector. We use the L_1 norm distance to calculate the distance metrics for $d_{\mathbf{I}}$ and $d_{\mathbf{Z}}$. The objective function of diversity loss is $min\ \mathcal{L}_{div}(s_1, \ldots, s_K)$.

At each stage i, given the text condition parameter c, the loss of generators G_i^K and discriminator D_i of D-CGAN can be defined as Eq. 3 and Eq. 4 generally, where s_i^k is from the synthetic image distribution $p_{G_i^k}$, and X_i is

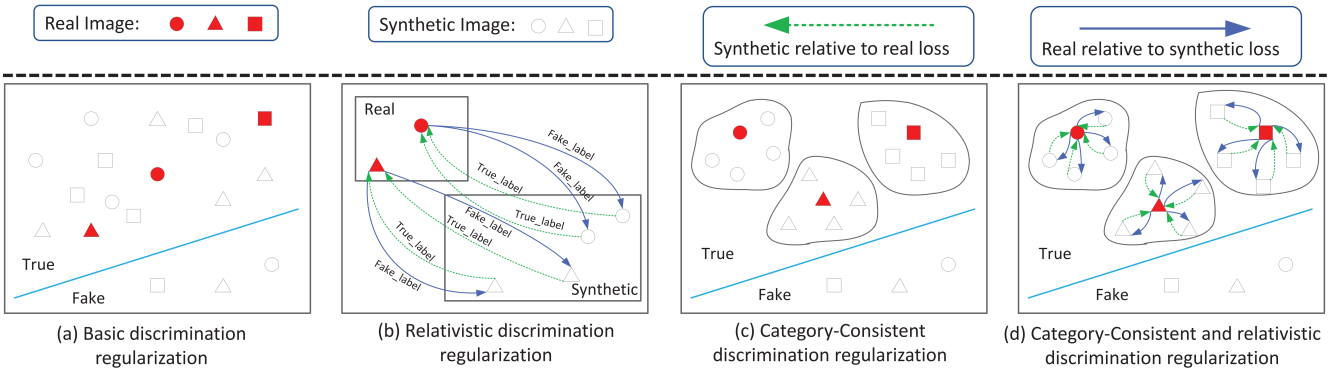

Fig. 3 Illustration of four different kinds of discrimination regularization. (a) is the basic discrimination regularization, which just discriminates whether the synthetic image is true. (b) is the proposed relativistic discrimination regularization, which adds the relativistic average conditional loss on the basis of (a). (c) is the proposed category-consistent discrimination regularization, which combines the category consistency loss on the basis of (a). (d) is the proposed category-consistent and relativistic discrimination regularization, which introduce the category consistency loss and the relativistic average conditional loss on the basis of (a). The green arrow is the relativistic loss from synthetic image to the corresponding real images based on the true label. The blue arrow is the relativistic loss from real image to the corresponding synthetic images based on the fake label.

from the true image distribution p_{data_i}. The G_i^K contains K generators $\{G_1, G_2, ..., G_K)\}$ at stage i. The term of $\sum_{k=1}^{K} \mathbb{E}_{s_i^k \sim p_{G_i^k}} \left[-\log(D_i(s_i^k, c)\right]$ in Eq. 3 is the conditional loss of generators G_i^k, which represents the approximate expected distribution of K synthetic images $\left\{s_i^1, s_i^2, ..., s_i^K\right\}$ matching the condition parameters c. The images from interpolated text embedding can fill in the gaps in the data manifold, which were presented during training. With the diverse term in Eq. 2 included, we are able to ensure the diversity of the synthetic images.

$$\mathcal{L}_{G_i^K}^{DIV} = \sum_{k=1}^{K} \mathbb{E}_{s_i^k \sim p_{G_i^k}} \left[-\log(D_i(s_i^k, c)\right] + \mathcal{L}_{sim(...)} + \mathcal{L}_{div(...)} \quad (3)$$

$$\mathcal{L}_{D_i}^{DIV} = K\mathbb{E}_{X_i \sim p_{data_i}} [-\log(D_i(X_i, c))] + \sum_{k=1}^{K} \mathbb{E}_{s_i^k \sim p_{G_i^k}} \left[-\log(1 - D_i(s_i^k, c))\right] \quad (4)$$

The generators $\mathcal{L}_{G_i}^{DIV}$ are trained to combine K different prior noise vectors, and the text embedding vector is used to interpolate K different synthetic images. The discriminator $\mathcal{L}_{D_i}^{DIV}$ has been trained to predict whether the synthetic images and the text match or not.

3.2 Category-Consistent and Relativistic Constraints

To better explore the visual feature representation of text, we design CRD-CGAN to generate K diverse synthetic images by alternative optimizing generators and discriminators with category-consistent and relativistic constraints. And Based on the two constraints, we proposed the category-consistent and relativistic discrimination regularization. To understand the meaning of a specific discrimination regularization, we discuss each regularization and the corresponding variant of CRD-CGAN as follow:

(a) **Basic discrimination regularization**(Fig. 3(a)) with the standard conditional loss is used to estimate the probability that the synthetic image is real in the variant D-CGAN.

(b) **Relativistic discrimination regularization**(Fig. 3(b)) with relativistic conditional loss can be used to estimate the probability that the synthetic image is enhanced realistic than a randomly sampled synthetic image, which can improve the performance of conditional loss. The variant with this regularization can be denoted as RD-CGAN.

(c) **Category-consistent discrimination regularization**(Fig. 3(c)) with category consistency loss is proposed to alleviate the over-category issue between K generators based on the image category. There is a performance imbalance between K generators $\{G_1, G_2, ..., G_K\}$, for example, some synthetic images is significantly different in shape or color from the corresponding real image. The corresponding variant with this regularization can be denoted as CD-CGAN.

(d) **Category-consistent and relativistic discrimination regularization**(Fig. 3(d)) in term of combining both a relativistic conditional loss and a category consistency loss is exploited in our proposed CRD-CGAN to jointly improve synthetic image's quality and diversity.

3.2.1 Relativistic conditional loss

Based on the Theorems 2.1, 2.2 and 2.4 of [58], If two distributions are disjoint or lie on low dimensional manifolds, the optimal discriminator will be perfect and its gradient will be zero almost everywhere. On the other words, the discriminator of GANs gets better, the gradient of the generator van-

ishes under certain conditions. If real and fake data are perfectly classified, the saturating loss has zero gradient and the non-saturating loss has non-zero but volatile gradient, which means that the discriminator in GANs often cannot be trained to optimally or with a too high learning rate [59]. One prior knowledge is that half of the samples in the mini-batch are fake. And this work [59] has proven that a relativistic discriminator is necessary to make GANs analogous to divergence minimization and produce sensible predictions.

The relativistic conditional loss is used to estimate the probability of synthetic image realistic, which is compared to a randomly sampled synthetic image, thereby improving the performance of conditional loss in the process of training the discriminator. Inspired by [59, 60], when estimating the probability that the input sample is true, we need to use the real and false data synchronously. In other words, the estimated probability from absolute real to fake relative to real or fake. Given a real data X_r and fake data X_f, the relativistic conditional discriminator D_i is defined as $sigmoid((D_i(X_r, c) - D_i(X_f, c)) \times l)$, where l is either 1 or -1. If X_r is more realism relative to X_f ($D_i(X_r, c) > D_i(X_f, c)$) or X_f is more artifacts relative to X_r ($D_i(X_f, c) > D_i(X_r, c)$), we set $l = 1$. On the contrary, we set $l = -1$.

If the discriminator D_i reach optimally on GANs, the gradient completely ignores the real data. And the D_i will focus entirely on fake data rather than learning the means for data to be real. Due to the fake data will not become more realistic, the training of D_i will get stuck. To address this issue, the relativistic discrimination regularization is proposed to find out all possible combinations of real and synthetic image in the mini-batch. It compares the critic of the real image to the average critic of synthetic images with true or fake label correspondingly. The loss function of relativistic conditional generator and discriminator are defined as $\mathcal{L}_{G_i}^{RE}$ and $\mathcal{L}_{D_i}^{RE}$.

Based on the above analysis, relativistic conditional loss is used to estimate the probability that the synthetic image is more true than random sampling of the synthetic image. $\mathcal{L}_{G_i}^{RE}$ of generator G_i^K is defined as Eq (6). And Eq (7) is $\mathcal{L}_{D_i}^{RE}$ of the discriminator D_i. In the Eqs.(5, 6, 7), l_r and l_f are symmetric labels [61], e.g., -1 and 1. $R(s_i^k, X_i, D_i, c, l_r)$ in Eq (6) is used to calculate the probability of more realism of s_i^k relative to X_i, and $R(X_i, s_i^k, D_i, c, l_f)$ in Eq (6) is used to calculate the probability of more artifacts of X_i relative to s_i^k. Eq (7) calculates the probability more realism of X_i relative to s_i^k and the probability of more artifacts of s_i^k relative to X_i. We improve the realism of synthetic image s_i^k by estimating the probability that s_i^k is relatively true in the generator G_i using $max\ \mathcal{L}_{D_i}^{RE}$. And the $min\ \mathcal{L}_{G_i}^{RE}$ is used to estimate the probability that the true image X_i is relatively more realistic than s_i^k in D_i, as to improve the performance of the whole GAN.

$$R(X, Y, D, c, l) = \log(\text{sigmoid}((D(X, c) - D(Y, c)) \times l)) \quad (5)$$

$$\mathcal{L}_{G_i^K}^{RE} = \sum_{k=1}^{K} \mathbb{E}_{X_i \sim p_{data_i}}^{s_i^k \sim p_{G_i^k}} \left[R(s_i^k, X_i, D_i, c, l_r) \right] + \sum_{k=1}^{K} \mathbb{E}_{X_i \sim p_{data_i}}^{s_i^k \sim p_{G_i^k}} \left[R(X_i, s_i^k, D_i, c, l_f) \right] \quad (6)$$

$$\mathcal{L}_{D_i}^{RE} = \sum_{k=1}^{K} \mathbb{E}_{X_i \sim p_{data_i}}^{s_i^k \sim p_{G_i^k}} \left[R(X_i, s_i^k, D_i, c, l_r) \right] + \sum_{k=1}^{K} \mathbb{E}_{X_i \sim p_{data_i}}^{s_i^k \sim p_{G_i^k}} \left[R(s_i^k, X_i, D_i, c, l_f) \right] \quad (7)$$

3.2.2 Category consistency loss

Considering that each category of images has its unique features, such as the shape and color of objects. The conditional loss just estimates that the probability of synthetic image is real without the category attributes. To further improve the performance of D-CGAN, we propose the category consistency loss $\mathcal{L}_{G_i^K}^{CC}$. It is used to estimate the probability that the synthetic images belong to the same category of the corresponding real image.

After extracting the visual features of synthetic images and real image by the image encoder, we use a softmax layer [62] to infer the probability distributions of each visual feature. We can use the cross-entropy to estimate the probability that the real image belongs to the corresponding category. Due to the synthesized image generated by D-CGAN does not have all visual features compared to real images, estimating the category calssification performance of synthetic images is very difficult using cross-entropy directly.

We extract the real image feature X_i and synthetic image feature s_i^k using the same image encoder. To correlate the relationship between synthetic images and categories, we calculate the cosine similarity [63] $Sim(X_i, s_i^k)$ between real image feature X_i and synthetic image feature s_i^k. The $Sim(X_i, s_i^k)$ is defined as the correlation weight. Based on the correlation weight, we concatenate the real image feature X_i and synthetic image feature s_i^k as the mixed feature. We apply a linear classification to produce classification score before softmax layer. Finally, we use the softmax layer to yield the category probability of the combined visual feature. If the generated synthetic images have the same category, the combined features should enforce the correct classification. Otherwise, the combined visual feature might weaken the classification confidence and even lead to misclassification. Therefore, we define the category consistency loss as a cross-entropy between prediction probabilities and the true category $category_i$:

$$\mathcal{L}_{G_i^K}^{CC} = -\sum_{y=1}^{Y} \delta(y = category_i) log P(y \mid X_i, s_i^1, \ldots, s_i^K) \quad (8)$$

where Y is the total category number of input dataset, $\delta(y = category_i)$ is 1 when y is the true category $category_i$ of X_i and 0 otherwise, and $P(y \mid X_i, s_i^1, \ldots, s_i^K)$ is defined as follow:

$$P(y \mid X_i, s_i^1, \ldots, s_i^K) = softmax\left(\mathbf{W}^T \left[X_i, \lambda \sum_{k=1}^{K} s_i^k \otimes Sim(X_i, s_i^k)\right] + \mathbf{b}\right) \quad (9)$$

where λ is the adjustment factor to balance the importance of synthetic image feature s_i^k, $\mathbf{W}$ and $\mathbf{b}$ are the linear classification parameters. The $\mathcal{L}_{G_i^K}^{CC}$ estimates the maximum likelihood that the synthetic images and real images belongs to the same category. In other words, it can improve the semantic consistency of synthetic images compared with the corresponding real image.

Based on the Eqs. 6, 7 and 8, the final loss function for generator and discriminator in our CRD-CGAN at the stage i can be formulated as follows:

$$\mathcal{L}_{G_i^K} = \mathcal{L}_{G_i^K}^{DIV} + \mathcal{L}_{G_i^K}^{RE} + \delta \mathcal{L}_{G_i^K}^{CC} \quad (10)$$

$$\mathcal{L}_{D_i} = \mathcal{L}_{D_i}^{DIV} + \mathcal{L}_{D_i}^{RE} \quad (11)$$

where δ is the weight of category consistent constraint. We set $\delta = 1$ in the following experiments.

4 Experiments

4.1 Experiment Settings

Datasets. In this paper, we evaluate our method on the Caltech-UCSD Birds-200-2011 Dataset [38] and the Oxford 102 Category Flower Dataset [39]. To evaluate our method in the multi-objects complex scenes, we evaluate our method on the MS COCO 2014 Dataset [40].

The Birds-200-2011 Dataset consists of 11,169 bird images from 200 categories and each category has 60 images averagely. We randomly select 9,935 images for training, and use the rest 1,234 images for testing. The dataset is very challenging because it contains images with multiple objects and various backgrounds. The Oxford-102 flower Dataset [39] consists of 8,189 images with 102 categories of flowers which commonly occurs in the United Kingdom, and each category has 40 to 258 images. Each image contains 10 different text descriptions in Caltech-UCSD Birds-200-2011 and the Oxford 102 Category Flower Dataset. The MS COCO 2014 Dataset [40] contains images of 91 object categories, which contains 82783 training images, 40504 validation images and 40775 testing images. Each image contains 5 different text descriptions in COCO.

Evaluation Metrics. We train the proposed models with three stages, and the resolution of synthetic image is 256×256 at the $3-th$ stage. We use Fréchet Inception Distance [64], denoted as FID, to evaluate the quality of synthetic images by calculating the distance between synthetic and real images through features extracted by Inception Network. Lower FID value indicates better quality of the synthetic image. The Inception score [65] is also used to evaluate the synthetic category consistency of images. High Inception score value means high category consistency between true image and synthetic image. In other words, we apply the Learned Perceptual Image Patch Similarity [66], denoted as LPIPS, to measure performance of diverse of GANs. Higher LPIPS value means more diverse of synthetic image. The R-Precision [18] is used to evaluate the correspondence between input text and the synthetic image.

Due to there is no category information in the process of calculating the FID value, we conduct a user-study on the testing datasets to further evaluate the similarity between synthetic images and corresponding real image. We first use our methods and baselines to generate 200 synthetic images each, which are based on the same 40 randomly selected texts from the validation set from each dataset. We define the following user-study rules to calculate the similarity between each synthetic image set and the corresponding real image. The volunteer must first choose synthetic images that are more similar to the real image. And then, they need to record the number m of the most similar synthetic images. Finally, the volunteers need to score the similarity σ between the selected synthetic images and the corresponding real image according to their first visual sense.

We invite 100 random volunteers to attend this experiment. The volunteer choose the images most similar to the corresponding real image, and they vote on the chose images for similarity. Based on the scoring and voting table of user-study, we use the equation 12 to calculate the total average similarity $Score_{similarity}$ between the selected synthetic images and the corresponding real image with each dataset for all compare methods.

$$socre_{similarity} = \left\{ \sum_{V=1}^{100} \left[\sum_{sentence=1}^{40} (m/5) * \sigma \right] /40 \right\} /100 \quad (12)$$

where V is the number of volunteer, *sentence* is the number of chose sentences from validation set, and 5 means we generate 5 synthetic images based on each method.

Baselines. We evaluate our proposed method CRD-CGAN with StackGAN++ [16], AttnGAN [18], MS-

GAN [19], as well as the variants D-CGAN, RD-CGAN, and CD-CGAN mentioned in Section 2.2. Note that there are two versions of D-CGAN [1], denoted as "D-CGAN-A" incorporating AttnGAN and "D-CGAN-S" incorporating StackGAN++. D-CGAN is our previous work to generate K synthetic images at the same time, where D-CGAN-A or D-CGAN-S share the same structure for each generator and discriminator with AttnGAN or StackGAN++. If D-CGAN-A removes $\mathcal{L}_{G_i^K}^{DIV}$, it will degenerate into K×AttnGAN. If D-CGAN-S removes $\mathcal{L}_{G_i^K}^{DIV}$, it will degenerate into K×StackGAN++. So we can describe the "D-CGAN-A" as $\mathcal{L}_{G_i^K}^{DIV}$ + K×AttnGAN, and "D-CGAN-S" as $\mathcal{L}_{G_i^K}^{DIV}$ + K×StackGAN++.

Implementation Details.

The training parameters of our CRD-CGAN involves the parameters of text encoder and image encoder of an attentional similarity model, the parameters of diversity and relative relation between real image and synthetic images wise, the parameters of category consistency of real and synthetic images, and the parameters of K generators and one discriminator in CRD-CGAN. It is worth mentioning that the hyper-parameters in AttnGAN are carefully tuned, and the K generators share the same parameters. Moreover, the parameters of all generators and discriminator are jointly updated using the Adam optimizer. Compared with AttnGAN and MSGAN, our CRD-CGAN has noticeably higher model complexity, where the amount of parameters of CRD-CGAN has increased by 52%. And the FLOPs of CRD-CGAN are up to 26.23G.

Our network is implemented in PyTorch, and the size of input image is 256×256. The BATCH_SIZE is 10 for the two datasets, and the MAX_EPOCH is 600. In the following, we use a RNN-based text-encoder to extract the feature c of input text t, which the EMBEDDING_DIM is 256 and hidden_size is 128. We feed the c and noises $\{z_1, ..., z_K\}$ to the generator $G_3^k = \{G_1, ..., G_K\}$ to generate K synthetic images with 256×256 at the third stage. Considering the comparison fair, we feed K random noises to StackGAN++ and AttnGAN to generate K synthetic images. A CNN-based image-encoder in generator and discriminator is used to extract the feature of image, which is fine-tine trained from the inception_v3 model. All generator and discriminator are trained using Adam solver with learning rate=0.0002, where the parameter β_1 and β_2 are set to 0.5 and 0.999.

Following our previous work [1], the multiple synthetic images generated by D-CGAN are complementary and differentiated to be used for extracting visual concepts embedded in the text. It means we can synthesize enough images for one text simultaneously, and the number of synthetic images depends on the GPUs. Considering the experimental conditions, we train the proposed CRD-CGAN with $K = 5$ and use it to generate five synthetic images with one input text to conduct the experimental evaluation.

Table 2 Diversity performance comparison on the Birds-200-2011 dataset.

Methods	FID ↓	LPIPS ↑	User study ↑
StackGAN++	27.90±0.02	31.37%±3.17	12.17%±1.05
MSGAN	27.48±0.38	36.87%±0.68	15.07%±5.62
AttnGAN	23.81±0.53	35.26%±0.05	12.38%±3.07
D-CGAN-S	26.41±0.48	37.12%±1.97	19.72%±3.20
D-CGAN-A	**22.61±0.13**	38.67%±0.49	26.49%±4.31
RD-CGAN	26.53±0.253	39.07%±0.31	24.73%±4.38
CD-CGAN	28.25±0.16	39.12%±0.28	31.67%±2.93
CRD-CGAN	24.59±0.35	**39.91%±0.34**	**33.24%±5.47**

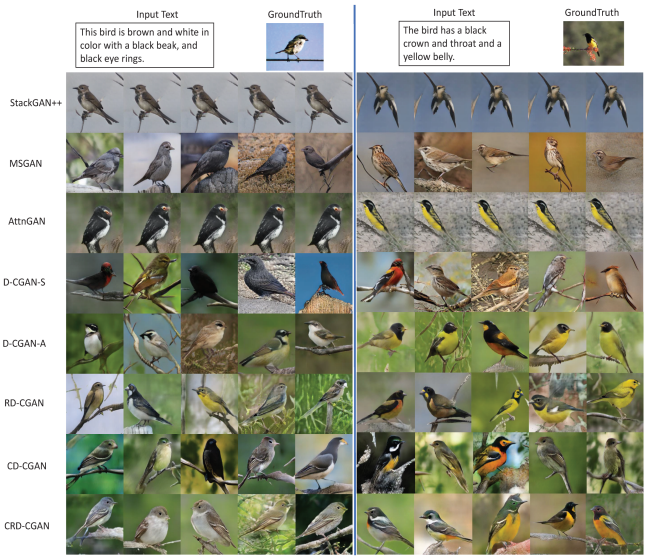

Fig. 4 Visualization of $K = 5$ high-resolution and photo-realistic synthetic images conditioned on a text, and comparison with state-of-the-art methods (top) on the Birds-200-2011 dataset.

4.2 Comparison with the State-of-the-arts

The FID and LPIPS scores for our proposed CRDCGAN and other methods on the Birds-200-2011 dataset are summarized in Table 2. From Table 2, we can see the synthetic images generated by D-CGAN-S are more diverse than by StackGAN++ and MSGAN, and the synthetic images generated by DC-GAN-A is more diverse than by AttnGAN. We also see that the synthetic images generated by CRD-CGAN are more category similarity than other methods. The synthetic images generated by CRD-CGAN exhibits the highest quality compared with real image correspondingly, which has the

lowest FID score. In addition, the synthetic images generated by CRD-CGAN have the highest LPIPS score, which also shows the most diverse than other methods.

We visualize some synthetic images by our CRD-CGAN and other methods in Fig. 4. We can observe the synthetic images generated by StackGAN++ and AttnGAN are less diverse, and the synthetic images generated by MSGAN and D-CGAN-S have lower similarity to real images while having good diversity. The synthetic images generated by our CRD-CGAN have highest similarity to real images with highest diversity. For example, the input text is *"This bird is brown and white in color with a black beak, and black eye rings"*. The words *"brown"* and *"white"* are the main attributes of the bird, which are reflected in the synthetic images generated by our CRD-CGAN. In addition, those synthetic images are also drawn with *"black beak"* and *"black eye rings"*.

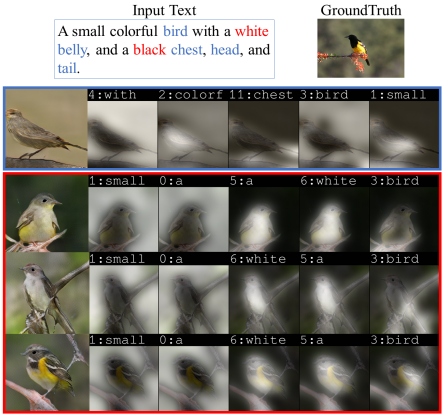

Fig. 6 $K = 3$ synthetic images generated conditioned on the text "A small colorful bird with a white belly, and a black chest, head, and tail." with the top-5 word attention maps. The results generated by AttnGAN are in blue rectangle, and the results generated by CRD-CGAN are in red rectangle, respectively.

Table 3 Diverse Performance comparison on the Oxford-102 flower dataset.

Methods	FID ↓	LPIPS ↑	User study ↑
StackGAN++	64.13±0.88	23.47%±1.63	7.47%±0.92
MSGAN	61.95±0.23	32.09%±0.29	16.07%±4.31
AttnGAN	42.41±0.19	33.04%±0.84	16.53%±1.90
D-CGAN-S	45.03±1.07	33.50%±0.13	19.36%±1.41
D-CGAN-A	**33.11±0.11**	33.16%±0.81	22.51%±4.56
RD-CGAN	42.76±0.23	33.31%±0.80	23.69%±3.57
CD-CGAN	43.71±0.19	34.82%±0.79	30.11%±3.14
CRD-CGAN	40.75±0.32	**37.56%±0.15**	**37.38%±3.07**

Fig. 5 Visualization of $K = 5$ high-resolution and photo-realistic synthtic images condtioned on a text, and compared with the corresponding real images (top) on the Oxford 102 Flower dataset.

We also evaluate the diversity performance of our proposed CRD-CGAN on the Oxford-102 flower dataset with FID and LPIPS metrics in Table 3. From Table 3, we can observe that (1) our proposed D-CGAN-S, D-CGAN-A and CRD-CGAN can effectively reduce the FID value, which means the synthetic images generated by our methods have higher category consistency with the corresponding real images; (2) the synthetic images generated by our CRD-CGAN have the highest diversity than other methods. Again, the category consistency loss and relativistic conditional loss can improve the diversity of synthetic images, while the category consistency effectively constrains the quality of synthetic images; and (3) CRD-CGAN achieves the best score in user study of similarity comparison, which means the synthetic images generated by CRD-CGAN have the highest shape and color consistency with the corresponding real images.

To better understand the effectiveness of our proposed CRD-CGAN, we also visualize the generated results of CRD-CGAN and its variants on the 102 Flower dataset. As shown in Fig. 5, the StackGAN++ just generates rough shape of flower. The MSGAN can keep a good quality and diversity of synthetic images, but it can not guarantee consistency with real images. Our D-CGAN-S can generate more diverse synthetic images, while it has lower image quality than D-CGAN-A and CRD-CGAN. The D-CGAN-A and CRD-CGAN can achieve the better diversity with higher image quality.

4.3 Analysis and Discussion

In this section, we analyze and discuss previous experimental results to further illustrate the advantages of our proposed CRD-CGAN.

To evaluate the image and text consistency of CRD-

Table 4 Inception score comparison on the CUB and the Oxford.

Methods	CUB	Oxford
StackGAN++	4.02±0.58	2.49±0.02
MSGAN	4.28±0.05	3.25±0.30
AttnGAN	4.31±0.68	3.36±0.02
HDGAN	4.15 ±0.05	3.45±0.07
CTGAN	4.23±0.05	3.71±0.06
D-CGAN-S	4.29±0.07	3.29±0.08
D-CGAN-A	4.51±0.04	3.39±0.02
RD-CGAN	4.54±0.06	3.48±0.03
CD-CGAN	**4.84±0.11**	3.50±0.02
CRD-CGAN	4.75±0.10	**3.53±0.06**

Table 5 R-Precision score comparison on the CUB and the Oxford.

Methods	CUB	Oxford
StackGAN++	10.57±4.83	13.66±1.44
MSGAN	16.08±5.12	18.67±1.73
AttnGAN	67.82±4.43	45.50±1.25
HDGAN	68.59±1.33	44.46 ±1.54
CTGAN	69.07±1.50	45.99±1.62
D-CGAN-S	67.33±4.85	20.13±0.98
D-CGAN-A	68.96±3.17	46.54±1.56
RD-CGAN	70.41±3.28	**56.88±2.72**
CD-CGAN	70.62±2.92	47.12±1.94
CRD-CGAN	**71.17±2.36**	47.70±2.22

CGAN, we visualize the word attention map of CRD-CGAN with AttnGAN in Fig. 6, which shows the top-5 word that were attended to by AttnGAN and CRD-CGAN. We can see that the color attribute word *"white"* has a lower attention in AttnGAN, while it is the top-5 word in CRD-CGAN.

Compared with StackGAN++, MSGAN and AttnGAN, our proposed CRD-CGAN effectively reduces the FID value in Tabel 2. At the same time, we also see that the synthetic images generated by D-CGAN-A have a higher FID value than our CRD-CGAN, which means that D-CGAN-A further improves the diversity of synthetic images. This result illustrates that CRD-CGAN reduces the quality of synthetic images compared with D-CGAN-A, this is because the importance of category visual features is emphasized in CRD-CGAN. The same results are also appears in Fig. 4, where the birds generated by D-CGAN-A have a realistic shape and reasonable color. However, the appearance difference between the generated birds and real birds is still relatively large. On the other hand, the synthetic images generated by CRD-CGAN have the highest scores in user study in Tabel 2 and Tabel 3, which means that they have the best shape and color consistency with the corresponding real images.

To improve the performance of RD-CGAN, we introduce the category consistency loss into D-CGAN-A and RD-CGAN. Firstly, the appearance consistency and diversity of CD-CGAN have been improved compared with D-CGAN-A. The CRD-CGAN also effectively improves the performance of RD-CGAN, and it has the best diversity compared than other methods in Table 2.The same result is also illustrated in Table 3. So we can experimentally confirm that the category consistency loss can effectively constrain the shape and color of the generated images. In other words, the proposed CRD-CGAN can generate synthetic images that are highly realistic and highly consistent with real images.

To further explore the contribution of the proposed CRD-CGAN with higher consistency between synthetic and real images. We firstly use the The Inception score [65] to evaluate the Synthetic quality of synthetic images. The Inception score performance of proposed CRD-CGAN is described in Table 4, while the StackGAN++ [16], AttnGAN [18], MSGAN [19], HDGAN [67] and CTGAN [68] are used for the baselines. From Table 4, we can confirm the robustness of our work. And the, we calculate the R-Precision score on the two datasets to evaluate the correspondence between input text and the synthetic image. Because of there are K synthetic images, we first calculate the R-Precision score between each synthetic image and the same text correspondingly.

Then, we use the mean of K R-Precision scores as the final R-Precision result of K synthetic images. The R-Precision score performance of proposed CRD-CGAN is described in Table 5. From the Table 5, our CRD-CGAN has the best R-Precision score on the CUB dataset. And our RD-CGAN has the best R-Precision score on Oxford dataset. In another word, the synthetic images generated by our proposed methods are better able to match the input text.

4.4 Experiments on the MS COCO 2014 Dataset

To verify the performance of the proposed CRD-CGAN in complex scenes, we conduct the experiments on the MS COCO 2014 Dataset in this section.

We use the cross entropy to estimate the category probability in Eq. (8), which can calculate the general category of a single object in synthetic image. However, there are muilt-objects in a single image on the MS COCO 2014 Dataset. So we use the multi-class multi-classification hinge loss as a criterion to estimate multi-category probability. We extend the

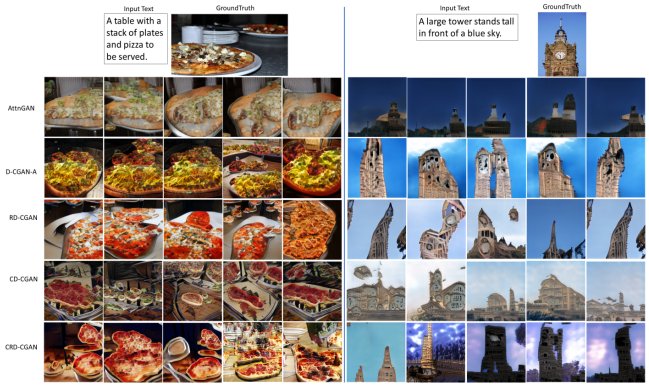

Fig. 7 Visualization of $K = 5$ high-resolution and photo-realistic synthtic images condtioned on a text, and comparing with the corresponding real images (top) on the MS COCO 2014 dataset.

Table 6 Diversity performance comparison on the MS COCO 2014 dataset.

Methods	FID ↓	LPIPS ↑	User study ↑
AttnGAN	42.16±0.01	42.06%±0.21	12.41%±0.70
CPGAN	55.82±0.52	–	16.24%±0.50
PPGAN	43.77±0.13	–	14.86%±0.34
D-CGAN-A	39.35±0.02	41.49%±0.23	16.05%±0.23
RD-CGAN	**38.61±0.10**	42.16%±0.14	14.64%±0.47
CD-CGAN	43.31±0.02	42.18%±0.43	12.85%±0.49
CRD-CGAN	41.79±0.07	**42.52%±0.46**	**19.45%±0.11**

Eq. (8) as follows:

$$\mathcal{L}_{G_i^K}^{CC} = -\sum_{y=1}^{Y} \frac{\max(0, 1 - logP(y \mid X_i, s_i^1, \ldots, s_i^K) - \delta(y \in \mathbf{c}_i)))}{|\mathbf{c}_i|} \tag{13}$$

where Y is the total category number of COCO dataset, $\delta(y \in \mathbf{c}_i)$ is 1 when y is one of the true multi-labels $\mathbf{c}_i$ of the real image X_i and 0 otherwise, and $|\mathbf{c}_i|$ is the total number of multi-labels for X_i.

We use two Nvidia Titan RTX GPUs to train the proposed CRD-CGAN on the MS COCO 2014 dataset, and the GPU memory is up to 48GB. Due to complex loss calculations and large-scale dataset, we set $K = 5$ to train the proposed RD-CGAN, CD-CGAN and CRD-CGAN. The FID, LPIPS and User-study scores for our proposed CRD-CGAN with AttnGAN [18], CPGAN [69] and PPGAN [70] on the MS COCO 2014 dataset are summarized in Table 6. From Table 6, we can see the quality of synthetic images generated by D-CGAN-A, RD-CGAN, CD-CGAN and CRD-CGAN are higher than those by AttnGAN.The synthetic images generated by CRD-CGAN exhibit the highest diversity than other methods. Our CRD-CGAN achieves the best score in the user study of similarity comparison, which also shows that the synthetic images generated by CRD-CGAN is the most similar to the real images.

We also visualize some synthetic images by our CRD-CGAN and other methods in Fig. 7. We can observe the synthetic images generated by AttnGAN have less diversity and have lower similarity to real images. While the sythetic images generated by our RD-CGAN, CD-CGAN and CRD-CGAN have better shape in details. For example, the input text is *"A table with a stack of plates and pizza to be served"*. The words *"table"*, *"plates"* and *"pizza"* are the main word, which are reflected in the synthetic images generated by our CRD-CGAN. In contrast, the results of AttnGAN only show a simple shape.

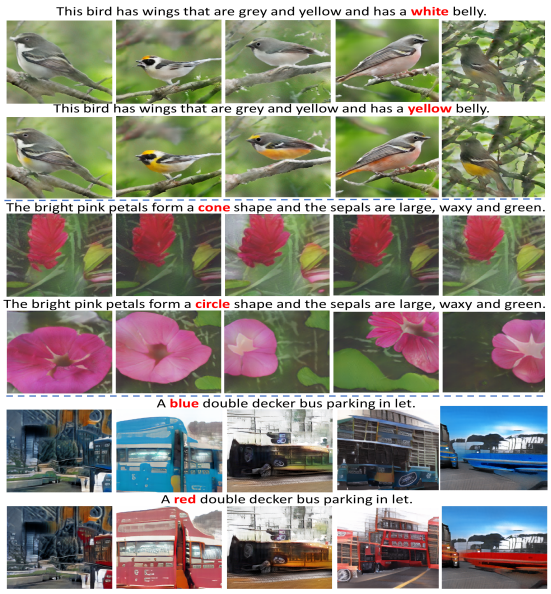

Fig. 8 Examples of CRD-CGAN on the ability of catching words changes (underline word in red) of the text description on the Birds-200-2011 dataset (top), on the Oxford-102 flower dataset (middle), and on MS COCO 2014 dataset(below).

4.5 Semantic Sensitivity Application

Furthermore, to evaluate the semantic sensitivity of the proposed CRD-CGAN, we change just one word in the input text. As shown in Fig. 8, the synthetic images are modified according to the changes of the input texts. For example, the color of bird "white" is changed to "yellow", the shape of flower "cone" is changed to "circle", and the color of bus "blue" is changed to "red". It demonstrates the proposed CRD-CGAN has the ability to retain the semantic diversity by catching the changes of the text description.

5 Conclusion

In this paper, we employed the category-consistent and relativistic diverse constraints to effectively exploit the rela-

tive real-or-fake relationship and main visual consistency between real image and K synthetic images. Our CRD-CGAN improves the estimating probability of more realism or more artifacts of K synthetic images relative to real image, and it uses the category consistency loss to ensure that the K synthetic images retain the main visual feature of corresponding category. Extensive experiments demonstrate the respected effectiveness and significance of proposed CRD-CGAN on the Birds-200-211 and Oxford-102 flower dataset. To evaluate the performance on lager-scale dataset, we also test our method on MS COCO 2014 datasets. The experiments results show that the proposed CRD-CGAN is also applicable to generate complex scenes with multiple categories.

Acknowledgements This work was co-supervised by Chengjiang Long and Chunxia Xiao, and supported by National Natural Science Foundation of China under Grants 61972298 and 61962019, and by the National cultural and tourism science and technology innovation project, under grant 2021064, and the Training program of high level scientific research achievements of Hubei Minzu University under Grant PY22011.

References

1. Hu T, Long C, Xiao C. A novel visual representation on text using diverse conditional gan for visual recognition. IEEE Transactions on Image Processing, 2021, 30: 3499-3512
2. Long C, Collins R, Swears E, et al. Deep neural networks in fully connected crf for image labeling with social network metadata. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision. 2019, 1607-1615
3. Long C, Hua G, Kapoor A. Active visual recognition with expertise estimation in crowdsourcing. In: Proceedings of the IEEE International Conference on Computer Vision. 2013, 3000-3007
4. Hua G, Long C, Yang M, et al. Collaborative active learning of a kernel machine ensemble for recognition. In: Proceedings of the IEEE International Conference on Computer Vision. 2013, 1209-1216
5. Long C, Hua G. Multi-class multi-annotator active learning with robust gaussian process for visual recognition. In: Proceedings of the IEEE international conference on computer vision. 2015, 2839-2847
6. Long C, Hua G, Kapoor A. A joint gaussian process model for active visual recognition with expertise estimation in crowdsourcing. International journal of computer vision, 2016, 116(2): 136-160
7. Long C, Hua G. Correlational gaussian processes for cross-domain visual recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, 118-126
8. Hua G, Long C, Yang M, et al. Collaborative active visual recognition from crowds: A distributed ensemble approach. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(3): 582-594
9. Wang Y, Wei Y, Qian X, et al. Sketch-guided scenery image outpainting. IEEE Transactions on Image Processing, 2021, 30: 2643-2655
10. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Communications of the ACM, 2020, 63(11): 139-144
11. Mirza M, Osindero S. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014
12. Reed S E, Akata Z, Mohan S, et al. Learning what and where to draw. Advances in neural information processing systems, 2016, 29
13. Reed S, Akata Z, Yan X, et al. Generative adversarial text to image synthesis. In: Proceedings of International conference on machine learning. 2016, 1060-1069
14. Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, 4681-4690
15. Zhang H, Xu T, Li H, et al. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. 2017, 5907-5915
16. Zhang H, Xu T, Li H, et al. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(8): 1947-1962
17. Zhang H, Goodfellow I, Metaxas D, et al. Self-attention generative adversarial networks. In: Proceedings of International conference on machine learning. PMLR. 2019, 7354-7363
18. Xu T, Zhang P, Huang Q, et al. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, 1316-1324
19. Mao Q, Lee H Y, Tseng H Y, et al. Mode seeking generative adversarial networks for diverse image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, 1429-1437
20. Yin G, Liu B, Sheng L, et al. Semantics disentangling for text-to-image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, 2327-2336
21. Cha M, Gwon Y L, Kung H T. Adversarial learning of semantic relevance in text to image synthesis. In: Proceedings of Proceedings of the AAAI conference on artificial intelligence. 2019, 3272-3279
22. Tan F, Feng S, Ordonez V. Text2scene: Generating compositional scenes from textual descriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, 6710-6719
23. Li Y, Gan Z, Shen Y, et al. Storygan: A sequential conditional gan for story visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, 6329-6338
24. Li W, Zhang P, Zhang L, et al. Object-driven text-to-image synthesis via adversarial training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, 12174-12182
25. Eghbal-zadeh H, Zellinger W, Widmer G. Mixture density generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, 5820-5829
26. Cheng J, Wu F, Tian Y, et al. RiFeGAN: Rich feature generation for text-to-image synthesis from prior knowledge. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 10911-10920

27. Liang J, Pei W, Lu F. Cpgan: Content-parsing generative adversarial networks for text-to-image synthesis. In: Proceedings of European Conference on Computer Vision. 2020, 491-508
28. Koh J Y, Baldridge J, Lee H, et al. Text-to-image generation grounded by fine-grained user attention. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021, 237-246
29. Gao L, Chen D, Zhao Z, et al. Lightweight dynamic conditional GAN with pyramid attention for text-to-image synthesis. Pattern Recognition, 2021, 110: 107384
30. Yang Y, Wang L, Xie D, et al. Multi-Sentence Auxiliary Adversarial Networks for Fine-Grained Text-to-Image Synthesis. IEEE Transactions on Image Processing, 2021, 30: 2798-2809
31. Arroyo D M, Postels J, Tombari F. Variational transformer networks for layout generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, 13642-13652
32. Fang F, Li Z, Luo F, et al. Discriminator Modification in GAN for Text-to-Image Generation. In: Proceedings of IEEE International Conference on Multimedia and Expo. 2022, 1-6
33. Fang F, Li Z, Luo F, et al. PhraseGAN: Phrase-Boost Generative Adversarial Network for Text-to-Image Generation. In: Proceedings of IEEE International Conference on Multimedia and Expo. 2022, 1-6
34. Park T, Liu M Y, Wang T C, et al. Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, 2337-2346
35. Hu M, Li J, Hu M, et al. Hierarchical Modes Exploring in Generative Adversarial Networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 10981-10988
36. Liu Z, Wang J, Liang Z. Catgan: Category-aware generative adversarial networks with hierarchical evolutionary learning for category text generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 8425-8432
37. Huang X, Li Y, Poursaeed O, et al. Stacked generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, 5077-5086
38. Wah C, Branson S, Welinder P, et al. The caltech-ucsd birds-200-2011 dataset, 2011
39. Nilsback M E, Zisserman A. Automated flower classification over a large number of classes. In: Proceedings of Sixth Indian Conference on Computer Vision, Graphics & Image Processing. 2008, 722-729
40. Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context. In: Proceedings of European conference on computer vision. 2014, 740-755
41. Ding B, Long C, Zhang L, et al. Argan: Attentive recurrent generative adversarial network for shadow detection and removal. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, 10213-10222
42. Zhang L, Long C, Zhang X, et al. Ris-gan: Explore residual and illumination with generative adversarial networks for shadow removal. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 12829-12836
43. Liu D, Long C, Zhang H, et al. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 8139-8148
44. Islam A, Long C, Basharat A, et al. DOA-GAN: Dual-order attentive generative adversarial network for image copy-move forgery detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 4676-4685
45. Zhang L, Long C, Yan Q, et al. CLA-GAN: A Context and Lightness Aware Generative Adversarial Network for Shadow Removal. In: Proceedings of Computer Graphics Forum. 2020, 483-494
46. Zhang J, Long C, Wang Y, et al. Multi-context and enhanced reconstruction network for single image super resolution. In: Proceedings of IEEE International Conference on Multimedia and Expo. 2020, 1-6
47. Vasu B, Long C. Iterative and adaptive sampling with spatial attention for black-box model explanations. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020, 2960-2969
48. Zhang J, Long C, Wang Y, et al. A two-stage attentive network for single image super-resolution. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32 (3): 1020-1033
49. Islam A, Long C, Radke R. A hybrid attention mechanism for weakly-supervised temporal action localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 1637-1645
50. Wei J, Long C, Zou H, et al. Shadow inpainting and removal using generative adversarial networks with slice convolutions. In: Proceedings of Computer Graphics Forum. 2019, 381-392
51. Yang Z, Dong J, Liu P, et al. Very long natural scenery image prediction by outpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, 10561-10570
52. Zheng Z, Zheng L, Yang Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of the IEEE international conference on computer vision. 2017, 3754-3762
53. Zheng Z, Yang X, Yu Z, et al. Joint discriminative and generative learning for person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, 2138-2147
54. Wang X, Zhu L, Zheng Z, et al. Align and Tell: Boosting Text-Video Retrieval With Local Alignment and Fine-Grained Supervision. IEEE Transactions on Multimedia, 2022
55. Shrivastava A, Pfister T, Tuzel O, et al. Learning from simulated and unsupervised images through adversarial training. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, 2107-2116
56. Shi J, Zhong Y, Xu N, et al. A simple baseline for weakly-supervised scene graph generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, 16393-16402
57. Zhang H, Koh J Y, Baldridge J, et al. Cross-modal contrastive learning for text-to-image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, 833-842
58. Arjovsky M, Bottou L. Towards principled methods for training gener-

ative adversarial networks, arXiv preprint arXiv:1701.04862, 2017.
59. Jolicoeur-Martineau A. The relativistic discriminator: a key element missing from standard GAN, arXiv preprint arXiv:1807.00734, 2018
60. Jolicoeur-Martineau A. On relativistic f-divergences. In: Proceedings of International Conference on Machine Learning. PMLR, 2020, 4931-4939
61. Mao X, Li Q, Xie H, et al. Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. 2017, 2794-2802
62. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks, Communications of the ACM, 2017, 60(6): 84-90
63. Pattnaik S, Nayak A K. Summarization of odia text document using cosine similarity and clustering. In: Proceedings of International Conference on Applied Machine Learning. 2019, 143-146
64. Heusel M, Ramsauer H, Unterthiner T, et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017, 6629-6640
65. Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training GANs. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016, 2234-2242
66. Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, 586-595
67. Zhang Z, Xie Y, Yang L. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, 6199-6208
68. Souza D M, Wehrmann J, Ruiz D D. Efficient neural architecture for text-to-image synthesis. In: Proceedings of International Joint Conference on Neural Networks. 2020, 1-8
69. Liang J, Pei W, Lu F. Cpgan: Content-parsing generative adversarial networks for text-to-image synthesis. In: Proceedings of European Conference on Computer Vision. 2020, 491-508
70. Nguyen A, Clune J, Bengio Y, et al. Plug & play generative networks: Conditional iterative generation of images in latent space. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, 4467-4477

Dr. Tao Hu received his PhD degree in computer science from School of Computer science, Wuhan University, Wuhan, in 2020. His current research interests include deep learning, and image processing.

Dr. Chengjiang Long received his Ph.D. degree in Computer Science from Stevens Institute of Technology in 2015. His research interests involve various areas of Computer Vision, Computer Graphics, Machine Learning, and Artificial Intelligence.

Chunxia Xiao is currently a professor at the School of Computer Science, Wuhan University, China. He received his Ph.D. from the State Key Lab of CAD & CG of Zhejiang University in 2006. His research areas include Computer graphics, Computer vision, Image processing, Virtual reality and Augmented reality. He has published more than 120 papers in journals and conferences.