

Received January 18, 2021, accepted February 8, 2021, date of publication February 11, 2021, date of current version February 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3058674

DGattGAN: Cooperative Up-Sampling Based Dual Generator Attentional GAN on Text-to-Image Synthesis

HAN ZHANG¹, HONGQING ZHU¹ , (Member, IEEE), SUYI YANG², AND WENHAO LI¹

¹School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

²Department of Mathematics, Natural, Mathematical & Engineering Sciences, King's College London, London WC2R 2LS, U.K.

Corresponding author: Hongqing Zhu (hqzhu@ecust.edu.cn)

This work was supported by the National Nature Science Foundation of China under Grant 61872143.


ABSTRACT Text-to-image synthesis task aims at generating images consistent with input text descriptions and is well developed by the Generative Adversarial Network (GAN). Although GAN based image generation approaches have achieved promising results, synthesizing quality is sometimes unsatisfied due to discursive generation of background and object. In this article, we propose a cooperative up-sampling based Dual Generator attentional GAN (DGattGAN) to generate high-quality images from text description. To achieve this, two generators with individual generation purpose are established to decouple object and background generation. In particular, we introduce a cooperative up-sampling mechanism to build cooperation between object and background generators during training. This strategy is potentially very useful as any dual generator architecture in GAN models can benefit from this mechanism. Furthermore, we propose an asymmetric information feeding scheme to distinguish two synthesis tasks, such that each generator only synthesizes based on semantic information they accept. Taking advantage of effective dual generator, the attention mechanism we incorporated on object generator could devote to fine-grained details generation on actual targeted objects. Experiments on Caltech-UCSD Bird (CUB) and Oxford-102 datasets suggest that generated images by the proposed model are more realistic and consistent with input text, and DGattGAN is competent compared to state-of-the-art methods according to Inception Score (IS) and R-precision metrics. Our codes are available at: <https://github.com/ecfish/DGattGAN>.

INDEX TERMS Asymmetric information feeding, cooperative up-sampling, dual generator, generative adversarial networks, text-to-image synthesis.

NOMENCLATURE

s	Sentence-level text feature.
w	Word-level text feature.
G_B	Background generator.
G_O	Object generator.
c	Sampled vector from Gaussian distribution.
c'	Semantic vector with incomplete information.
θ_H	High resolution object feature.
θ_L	Low resolution object feature.
β_H	High resolution background feature.
β_L	Low resolution background feature.
O_H	High resolution object.

O_L	Low resolution object.
M_H	High resolution mask.
M_L	Low resolution mask.
B_H	High resolution background.
B_L	Low resolution background.
φ	Combination function.
I_H	High resolution image.
I_L	Low resolution image.
D_H	High resolution discriminator.
D_L	Low resolution discriminator.
D_B	Background discriminator.
z_B	Actual input of G_B .
z_O	Actual input of G_O .
\mathbb{E}	Expectation.
\mathcal{L}	Loss function.

The associate editor coordinating the review of this manuscript and approving it for publication was Emanuele Bellini .

f_G	Generation constrained loss function.
λ	Hyper-parameter of DAMSM loss term.
γ	Coefficient of real/fake loss term in \mathcal{L}_{D_B} .
HR	High resolution.
LR	Low resolution.
RF	Receptive field.
x_L	Sample from real low resolution images.
x_H	Sample from real high resolution images.
x_O	Sample from real object patches.
x_B	Sample from real background patches.
D_{KL}	KL-divergence.
D_{B_cls}	Background/object discriminant matrix.
D_{B_rf}	Real/fake discriminant matrix.
β_1, β_2	Hyper-parameters in Adam optimizers.

I. INTRODUCTION

In recent years, text-to-image synthesis has drawn much interest and rapidly expand the area of computer vision. This natural language visualization task is also a fundamental technique towards multiple applications such as computer aided design, text visualization and restoring face. One of the challenges is visual quality of generated images could hardly suppress real images in terms of resolution, object outline and vivid detail. Besides, semantically consistent image generation is of even higher difficulty compared to common image generation. Overall diversities on synthesized results may also be one of the difficult issues.

Aiming at generating photo-realistic images, Generative Adversarial Network (GAN) [3] is generally regarded as a feasible candidate. However, conventional GAN architecture generating based on input noises is less contributing to match the text information. Another image generation model Conditional GAN (cGAN) [4] is proposed, which almost all later text-to-image models are built based on this condition restrained architecture [1], [2], [5]. In these models, text descriptions are usually encoded into semantic vectors and fed to both generator and discriminator as conditions, which show impressive effect on controlling overall text-consisting image generation. Based on the structure of cGAN, a multi-stage text-to-image generation framework sketching outlines and fulfilling details by two degrees of resolution was introduced by StackGAN [6] and StackGAN++ [1]. However, lacking crucial fine-grained information is discovered as main problem hindering qualified image generation by StackGAN and StackGAN++. Another text-to-image synthesis model AttnGAN [2] is proposed consequently which aims at synthesizing more realistic and fine-grained images based on attentional word-level feature fusion. Although overall quality of generation is enhanced, some targets are in ambiguous outlines unable to differentiate from background areas as shown in Figure 1. Based on the considerations that most text data participated in image synthesis are object description texts, which background information is of less significant concern compared to

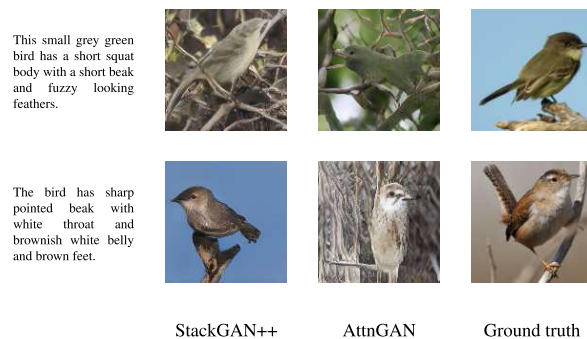


FIGURE 1. Generated images on CUB test set by StackGAN++ [1], AttnGAN [2], and their ground truth. A phenomenon that generated object has abnormal outline mixing with background appears.

targeted object, another branch of studies suggests that splitting foreground and background into different data spaces is helpful. InfoGAN [7] can learn disentangled representations of latent space by maximizing the mutual information between a subset of latent variables and the observation data. But data spaces aren't fully decoupled, which gains some extents of improvement by depicting object shape suggested by LR-GAN [8]. Inspired by these models, FineGAN [9] establishes a new unsupervised hierarchical image synthesis method with fine-grained details emphasized. Although two generators inserted still lead to some problems in synthesizing, the overall quality of realistic and detailed object generation had been promoted.

In this article, we propose a novel text-to-image generative framework named cooperative up-sampling based Dual Generator attentional GAN (DGattGAN). By this framework, original incompatible issue in data space decoupled could be solved and advanced structures from existing text-to-image models are emphasized at their most. Pursuing the strategy of enhancing object generation apart from background, we set up dual generator which object and background generation tasks are arranged to individual generator as show in Figure 2. To prevent unsynchronized and generator degradation issue unsolved in existing dual generator models [8], [9], a cooperative up-sampling scheme is designed to build feature interflow between generators. Being the first architecture that use dual generator GAN in text-to-image generation, two optimization methods are raised to coordinate dual generator in integral text-to-image synthesis model. In particular, an asymmetric information feeding scheme is raised apart from existing study [9] that adds generation constraint over output. By achieving separation of target and background, the attention mechanism that we inserted on object generator enable more contributing word-level feature fusion.

We summarize our main contributions as follows:

- Dual generator architecture is established to decouple object and background data distributions aiming at more realistic object generation.
- Two methodologies harmonizing synthesis behavior accompanying with two generators are discussed.

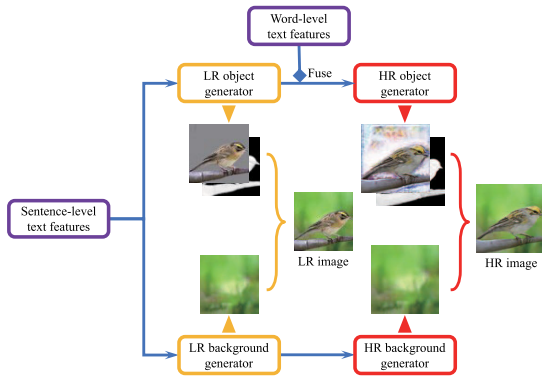


FIGURE 2. Text feature fusion on dual generator. LR and HR indicates low resolution and high resolution respectively. The proposed dual generator only delivers word-level features for object generation.

Asymmetric information feeding scheme is developed as a new training strategy for dual generator.

- We propose a cooperative up-sampling mechanism for feature interflow between generators, which prevents incompatible generation and model degradation problems. This design is considered valuable for any dual generator architecture and may trigger more proposals on generators coordination.
- Above contributions on dual generator and the attention mechanism incorporated to object generator enable object-wise word-level feature fusion, which contributes to more text-consistent generation and overall improvement in image quality.

The remainder of this article is organized as follows. Related works will be briefly discussed in Section II. In Section III, we present detail modules in building the DGattGAN architecture as well as the optimization strategies of dual generator. Experiments will be discussed in Section IV. Finally, we will conclude this article and future works in Section V.

II. RELATED WORKS

In this section, researches related to this study are presented in two subsections, dual generation architecture and GAN based text-to-image researches.

A. DUAL GENERATION ARCHITECTURE

Dual generator architecture is currently applied to some synthesis tasks including image captioning [10], image generation [8], [9], video generation [11], etc. In specific, two generators proposed by Liu *et al.* [10] are responsible for caption generation and retrieval respectively in image captioning task. Dual-generator architecture has also been introduced into image generation by LR-GAN [8] that uses a learnable foreground mask to distinguish foreground and background region. Similarly, two-stream generative model is introduced by Vondrick [11] to improve quality in video generation. Then, FineGAN [9] has another background discriminator for image generation based on LR-GAN. However, the methodology of separate generation has not been

used in text-to-image synthesis, and degradation problem may exist using these mask untangling methods as discussed in Section III.C. Therefore, this study firstly adopts object and background generators with a cooperative up-sampling module to prevent degradation for text-to-image synthesis.

B. TEXT-TO-IMAGE SYNTHESIS METHOD

Recent studies on text-to-image synthesis tasks mainly aim at improvement on three aspects: visual quality, text consistency and scene synthesis. In this section, text-to-image synthesis models are introduced by these characteristics.

1) VISUAL QUALITY

The main branch of text-to-image synthesis models is exploited by Reed *et al.* [5] who first utilized GAN [3] in text-to-image generation. In this study, GAN-INT-CLS is proposed based on cGAN [4] to generate 64×64 images conditioned by text. However, generated images have poor resolution and the subsequently proposed StackGAN [6] brings an improvement. StackGAN generates low resolution 64×64 images conditioned on input text using Stage-I GAN. Stage-II GAN takes text and the result from Stage-I as input to synthesize 256×256 images. StackGAN is not an end-to-end model in which training and generation process of StackGAN are divided into two stages. StackGAN++ [1] is a follow-up work of StackGAN, which adopts a tree-like structure to generate 64×64 , 128×128 and 256×256 images. Compared with StackGAN, StackGAN++ shows more stable training behaviour. HDGAN [12] also provides a single-stream generator architecture with hierarchy-nested discriminators to synthesize high resolution images. Besides, some other studies [13], [14] also aim at improving definition for more photo-realistic generation with superior external modules, e.g., residual block feature pyramid attention module [13]. Although images of high resolution could be synthesized with these tree-like or hierarchy structures, objects generated are still in poor shape. Fundamentally, attributing learning ability of object shape in an unsupervised manner might be the reason that hinders further improvement of these studies. In this study, an object generator would take over the task of synthesizing accurate object mask to stitch up object and background.

2) TEXT CONSISTENCY

Text consistency is another key factor that draws much attention in text-to-image synthesis. Xu *et al.* [2] designed AttnGAN based on StackGAN++. AttnGAN encodes input text into sentence-level feature and word-level feature with a bidirectional Long Short-Term Memory (LSTM) model. Before up-sampling to higher resolution images, AttnGAN fuses image feature with word-level feature using attention mechanism. AttnGAN uses Deep Attentional Multimodal Similarity Model (DAMSM) model to calculate the match score between local image feature and word-level feature. Based on AttnGAN, Qiao *et al.* introduced MirrorGAN [15], which is a global-local attentive and semantic-preserving text-to-image-to-text framework containing three modules.

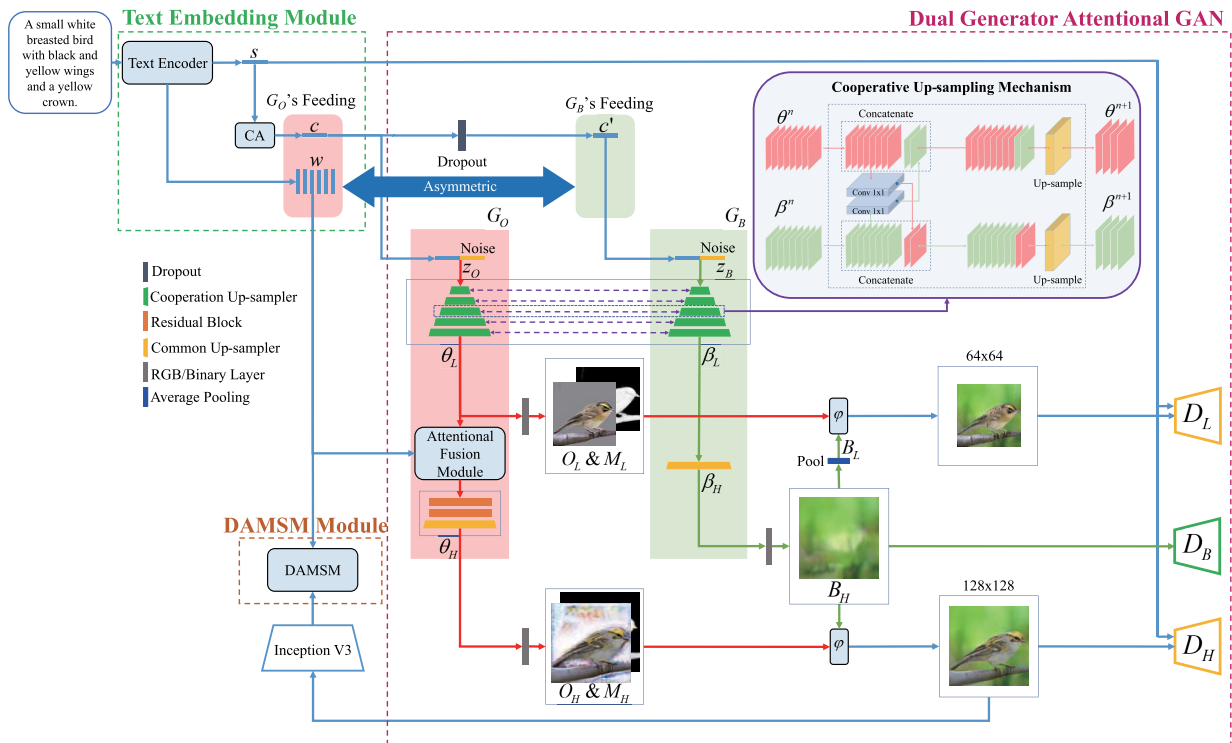


FIGURE 3. The architecture of the proposed DGattGAN architecture. Text embedding module generates c and w for generator input and DAMSM evaluation. G_O and G_B are used to generate object and background separately. The proposed asymmetric information feeding scheme places control over individual generator ahead of input. Features between generators are shared using the proposed cooperative up-sampling. ϕ is utilized to synthesize final 64×64 and 128×128 image with the object, mask and background.

Recently, a Dual Attn-GAN [16] adds another visual attention model together with the proposed mechanism by AttnGAN. This new Visual Attention Model (VAM) enhances local details and global structures by focusing on related features from relevant words and different visual regions. However, AttnGAN fuses word-level feature with the entire image, and object area of higher necessity towards detail generation is regarded equally than other area. Therefore, the proposed DGattGAN attributes attention mechanism proposed by AttnGAN more specifically to object area by completely decoupling object. Another model generates images by multi-pair generators and discriminators training is introduced in rdAttnGAN [17]. Other studies suggest that enhancing text consistency by single text information may hardly meet expectation, while building multiple text models might be helpful. RiFeGAN [18] designed by Cheng *et al.* takes several texts or captions as inputs and generates images based on all descriptions. Sharma *et al.* [19] uses dialogue as model input instead of text sentences. In addition, many other studies [20], [21] also achieve some extents of improvement on generation consistency.

3) SCENE SYNTHESIS

Further studies point out that real-world images contain large amount of information including object category, spatial configurations of objects, scene context, etc. However, some studies suggest that directly mapping from text to image is not

suitable in complex text description. Hong *et al.* [22] proposes a hierarchical approach, which constructs a semantic layout from input text and then synthesizes final image conditioned on generated semantic layout. Johnson *et al.* [23] constructs a semantic layout with Graph Convolutional Network (GCN). Li *et al.* [24] proposes an Obj-GAN based on the research of Hong *et al.* [22] which uses object-driven attentive image generator to synthesize salient objects. In addition, a Faster R-CNN [25] based object-wise discriminator is proposed by Obj-GAN to determine consistency of synthesized objects with text description and semantic layout.

III. METHODOLOGY

A. MODEL ARCHITECTURE

In this subsection, we would introduce the overall architecture as shown in Figure 3. This proposed framework consists of three modules, DAMSM, text embedding module and the proposed DGattGAN.

1) TEXT EMBEDDING MODULE

In text embedding module, we use a pretrained bi-directional LSTM text encoder provided by AttnGAN [2] to encode input text. The two hidden states of one word are concatenated and output as word-level text feature w . The last two hidden states of bi-directional LSTM are concatenated and output as sentence-level text feature s . As explained in StackGAN [6] that the latent space for text embedding is usually

high dimensional, such that discontinuity in latent data manifold would occur under limited data encoding, which is not desirable for training generator. Therefore, to prevent conditioning manifold and overfitting, the Conditioning Augmentation (CA) module proposed by StackGAN is used in our model to produce a Gaussian distribution over the possible values of c from which the datapoint s could have been generated. Sampled semantic vector c has an interpretation as a latent representation.

2) THE PROPOSED DGattGAN ARCHITECTURE

Next, sampled semantic vector c is conveyed respectively into dual generators that firstly adopt by this DGattGAN in text-to-image generation. Different from existing methods [9], we propose another asymmetric information feeding scheme to convey semantic information only available for individual generation task. Latent space samples input of object generator G_O and background generator G_B are labeled as c and c' respectively in which partial dimensions were randomly abandoned in c' during training by a dropout. To ensure generation diversity, random noises are concatenated with c and c' . In this case, G_B fed with incomplete information would only be available for target empty images synthesis. Then, a cooperative up-sampling mechanism for dual generator is raised by the proposed study such that features from G_O and G_B could be shared and up-sampled. Then, the 64×64 low resolution (LR) object feature θ_L would splits into two paths where one produces LR object O_L and mask M_L , and the other conveys features into an attentional fusion module proposed by AttnGAN [2] that combines θ_L with word-level feature w . Fused θ_L would pass through two residual blocks and up-sample layer to obtain 128×128 HR object feature θ_H and generate high resolution (HR) object O_H and mask M_H . In background generation path, HR background B_H is obtained by directly up-sampling background feature β_L without word-level feature fusion. Instead of using RGB layer, an average-pooling layer is utilized to down-sample high resolution background (B_H) to 64×64 low resolution background (B_L). The combination function $\varphi(G_B(z_B), G_O(z_O))$ generates final image using object, mask and background image. It is defined as:

$$\varphi(G_B(z_B), G_O(z_O)) = M \odot O + (1 - M) \odot B, \quad (1)$$

where \odot denotes element-wise product. M , O and B denote object mask, object, and background images, respectively. Then LR image I_L and HR image I_H can be obtained using φ :

$$\begin{aligned} I_L &= \varphi_L(G_B(z_B), G_O(z_O)) \\ &= M_L \odot O_L + (1 - M_L) \odot B_L, \end{aligned} \quad (2)$$

$$\begin{aligned} I_H &= \varphi_H(G_B(z_B), G_O(z_O)) \\ &= M_H \odot O_H + (1 - M_H) \odot B_H. \end{aligned} \quad (3)$$

DGattGAN uses discriminators D_L and D_H to determine whether I_L and I_H are real images that match input text, while discriminator D_B is utilized to discriminate if B_H is real background. However, D_L and D_H only take sentence-level

text feature s as criterion without detailed word-level text feature.

3) DAMSM MODULE

DAMSM module proposed by AttnGAN [2] takes word-level semantic vectors w and global sentence vectors s encoded by LSTM text encoder. Local and global image features of I_H extracted by Inception v3 [26] are also fed to DAMSM. Subsequently, image-text similarity to evaluate fine-grained loss is obtained by calculating multi-modal similarity losses on word-level features against local image features and sentence-level features against global image features.

B. OPTIMIZATION OF DGattGAN's DUAL GENERATOR

Generally, classical cGAN based text-to-image synthesis methods optimize single generator by loss functions of the generator and discriminator formulated as:

$$\mathcal{L}_G = -\mathbb{E}_{s \sim p_s, z \sim p_z} [\log D(G(s, z), s)], \quad (4)$$

$$\begin{aligned} \mathcal{L}_D &= -\mathbb{E}_{x \sim p_x} [\log D(x, s)] \\ &\quad - \mathbb{E}_{s \sim p_s, z \sim p_z} [\log(1 - D(G(s, z), s))], \end{aligned} \quad (5)$$

where \mathbb{E} denotes the expectation, x is from the true image distribution p_x , s is from the sentence-level semantic distribution p_s , z is random noise. However, (4) and (5) couldn't tackle with situation when two generators are synthesizing different parts of the image. Existing dual generator proposed by FineGAN [9] controls generation by adding generation constraint on output. We propose another asymmetric information feeding scheme to synchronize dual generator. The following part would introduce in detail.

1) ASYMMETRIC INFORMATION FEEDING SCHEME

The method of asymmetric information feeding is raised mainly due to concerns over distinguishing individual generation task. For object description text-to-image task specifically, only object generator is regarded as necessary towards complete semantic information in both sentence-level and word-level, while background generator isn't. Implementation detail might vary based on particular synthesis tasks, but this mechanism is considered an alternative providing for dual generator structures aside from constraining output.

(i) Sentence-level asymmetry scheme. In object description text-to-image task, the sentence-level semantic vector c could be regarded as samples from object's latent space. To distinguish background generation path, a dropout layer is inserted and subsequent samples from background's latent space is labelled as c' (see Figure 3). In this way, neural units and connections of c' are randomly dropped during each training iteration so that G_B is only available for object-free image synthesis. To further enhance generation diversity, c and c' are incorporated with noise to form actual input code z_O and z_B . Additionally, two conditional discriminators D_L and D_H accept sentence-level feature s as criterion. The loss functions of generators and discriminator can be formulated as:

$$\begin{aligned} \mathcal{L}_{G_B, G_O} &= -\mathbb{E}_{s \sim p_s} [\log(D_H(\varphi_H(G_B(z_B), G_O(z_O)), s))] \\ &\quad - \mathbb{E}_{s \sim p_s} [\log(D_L(\varphi_L(G_B(z_B), G_O(z_O)), s))], \end{aligned} \quad (6)$$

$$\begin{aligned} \mathcal{L}_{D_H} = & -\mathbb{E}_{x_H \sim p_{x_H}} [\log D_H(x_H, s)] \\ & - \mathbb{E}_{s \sim p_s} [\log(1 - D_H(\varphi_H(G_B(z_B), G_O(z_O)), s))], \end{aligned} \quad (7)$$

$$\begin{aligned} \mathcal{L}_{D_L} = & -\mathbb{E}_{x_L \sim p_{x_L}} [\log D_L(x_L, s)] \\ & - \mathbb{E}_{s \sim p_s} [\log(1 - D_L(\varphi_L(G_B(z_B), G_O(z_O)), s))]. \end{aligned} \quad (8)$$

Since D_L and D_H take complete s as criterion where partial information of it isn't accessed by G_B , text consistent object shouldn't appear in background generation. Therefore, discriminators should see a distinction in real-fake judgment.

(ii) Word-level asymmetry scheme. In DGattGAN, word-level features reflecting detailed object visual attributes are only fed to G_O . An attentional fusion module to fuse θ_L with w for G_O is established before θ_L is up-sampling to θ_H , while G_B directly up-samples β_L to β_H . Meanwhile, the total dual generator network should minimize \mathcal{L}_{DAMSM} [2] as follow:

$$\begin{aligned} \mathcal{L}_{G_B, G_O} = & -\mathbb{E}_{s \sim p_s} [\log(D_H(\varphi_H(G_B(z_B), G_O(z_O)), s))] \\ & - \mathbb{E}_{s \sim p_s} [\log(D_L(\varphi_L(G_B(z_B), G_O(z_O)), s))] \\ & + \lambda \mathcal{L}_{DAMSM}, \end{aligned} \quad (9)$$

where \mathcal{L}_{DAMSM} referring to (14) in AttnGAN [2] calculates image-text similarity at word-level and λ is a hyper-parameter. In this case, only G_O is available for detailed word-level information synthesis and thus DGattGAN benefits from word-level asymmetry in two aspects: improvement on synthesizing text consistent images by accurately fusing word-level feature with object feature; maintaining segregation on G_B from word-level feature.

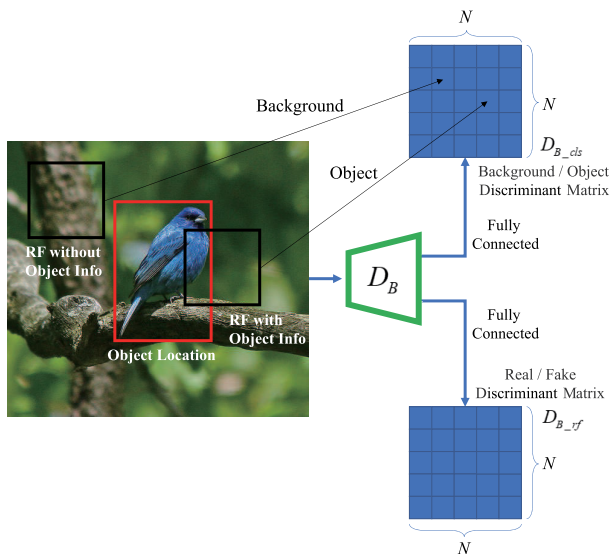


FIGURE 4. The architecture of DGattGAN background discriminator, black boxes represent the receptive field (RF) and red boxes represents the object location.

2) GENERATION CONSTRAINED LOSS

Another generally adopted method to control each generation task for existing dual generator GAN is adding constraint over the loss function of generator. In DGattGAN, only background generation is constrained due to less significant concern on object location compared to object generation. Constraining on individual generator would automatically achieve overall coordination of dual generator. The loss functions can be formulated as:

$$\begin{aligned} \mathcal{L}_{G_B, G_O} = & -\mathbb{E}_{s \sim p_s} [\log(D_H(\varphi_H(G_B(z_B), G_O(z_O)), s))] \\ & - \mathbb{E}_{s \sim p_s} [\log(D_L(\varphi_L(G_B(z_B), G_O(z_O)), s))] + f_G, \end{aligned} \quad (10)$$

where f_G is the generation constrained loss function, the expression of f_G will be presented in (12). We use a background discriminator D_B to determine whether the generated image is real background. Since there is no specific background or object dataset for training, the discriminator from PatchGAN [27] could well serve to distinguish object and background in single image. Then the output of our discriminator is no longer a scalar, but an $N \times N$ discriminant matrix as shown in Figure 4. Each element in the discriminant matrix is calculated from corresponding patch (called receptive field (RF)) in input image. D_{B_rf} and D_{B_cls} are two $N \times N$ discriminant matrixes fully connected after D_B . Each element in D_{B_rf} represents real/fake judgement the corresponding receptive field. Each element in D_{B_cls} whose receptive field includes pixels of object region is marked as object; each element whose receptive field excludes pixels of object region is marked as background. See Figure 4 for more details. In this way, we assume that object location (or segmentation) could be given by the dataset. D_{B_cls} is trained to accurately classify real image patches as background or object, while D_{B_rf} still competes with background generator. The loss function of D_B can be formulated as, (11), as shown at the bottom of the next page, where γ is the coefficient of real/fake term. This study uses D_B to add constraint on background output. f_G can be formulated as:

$$\begin{aligned} f_G = & \gamma (-\mathbb{E}_{s \sim p_s} [\log(D_{B_rf}(G_B(z_B)))] \\ & - \mathbb{E}_{s \sim p_s} [\log(D_{B_cls}(G_B(z_B)))]). \end{aligned} \quad (12)$$

The generated background is expected to be discriminated as real by D_{B_rf} and classified as background by D_{B_cls} . Using (2)(3)(6)(7)(8)(9)(10), the loss functions of DGattGAN's generators and discriminators can be rewritten as:

$$\begin{aligned} \mathcal{L}_{D_L} = & -\mathbb{E}_{x_L \sim p_{x_L}} [\log D_L(x_L, s)] \\ & - \mathbb{E}_{s \sim p_s} [\log(1 - D_L(I_L, s))], \end{aligned} \quad (13)$$

$$\begin{aligned} \mathcal{L}_{D_H} = & -\mathbb{E}_{x_H \sim p_{x_H}} [\log D_H(x_H, s)] \\ & - \mathbb{E}_{s \sim p_s} [\log(1 - D_H(I_H, s))], \end{aligned} \quad (14)$$

$$\begin{aligned} \mathcal{L}_{G_B, G_O} = & -\mathbb{E}_{s \sim p_s} [\log(D_L(I_L, s))] - \mathbb{E}_{s \sim p_s} [\log(D_H(I_H, s))] \\ & + f_G + \lambda \mathcal{L}_{DAMSM}. \end{aligned} \quad (15)$$

DGattGAN should optimize the background discriminator D_B at the same time according to (11).

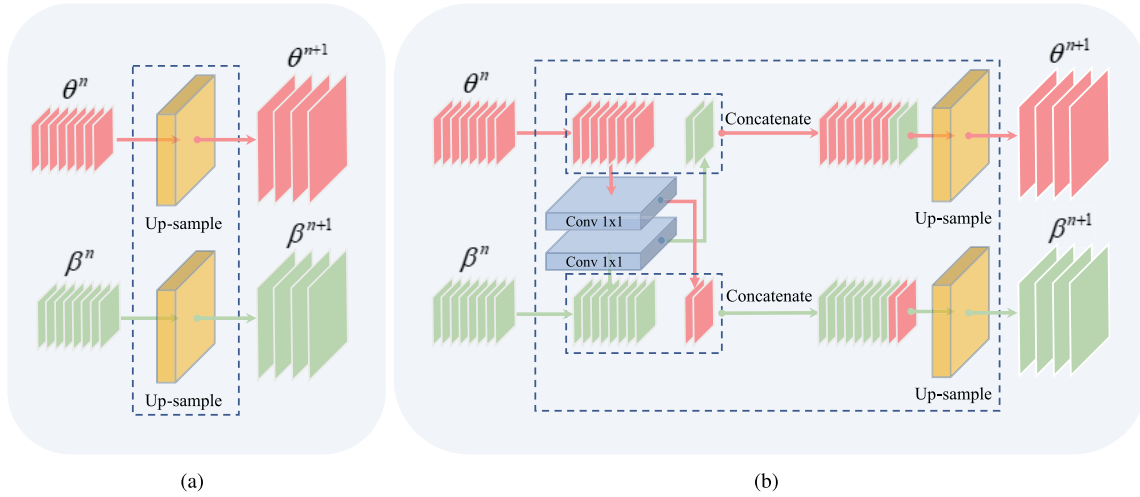


FIGURE 5. Comparing different up-sampling schematic, (a) common up-sample; (b) the proposed cooperative up-sampling scheme.

C. COOPERATIVE UP-SAMPLING MODULE

In [9], FineGAN uses two generators which are mutually independent and never share their features. Then two main problems seem to be: (i) unsynchronization between the generated background and object; (ii) degeneration from dual to single generator. In specific, degeneration might happen in situation that G_O learns faster than G_B and finally gains ability generating both object and background. In this case, mask M tends to have all elements equal to 1, and thus final image only reflects object generation. This degeneration issue can be seen from following formula:

$$\varphi(G_B(z_B), G_O(z_O)) = \underbrace{M}_{\approx 1} \odot O + \underbrace{(1 - M)}_{\approx 0} \odot B \approx O, \quad (16)$$

where dual generator degrades into single generator. According to the chain rule of derivatives and (16), we have

$$\frac{\partial \mathcal{L}_{G_B, G_O}}{\partial G_B} = \frac{\partial \mathcal{L}_{G_B, G_O}}{\partial \varphi} \cdot \frac{\partial \varphi}{\partial G_B} \approx \frac{\partial \mathcal{L}_{G_B, G_O}}{\partial O} \cdot \frac{\partial O}{\partial G_B} \equiv 0, \text{ where } \frac{\partial O}{\partial G_B} = 0. \quad (17)$$

Thus, G_B cannot have its weight updated. Then, G_B should always fail to synthesize background matching G_O , and degeneration is irreversible.

Based on the concern of synchronizing generation, this study proposes a cooperative up-sampling module to have two generators share their features. Figure 5 shows the architecture of this cooperative up-sampling module. By adopting interflow (Conv 1×1), each generator fuses the other's features before every up-sampling. Here, object features and background features of the n th up-sampling are denoted as θ^n

and β^n . θ^{n+1} is given by up-sampling the concatenation of θ^n , and convolved β^n with 1×1 convolution. Still assume that, G_B has lower learning speed than G_O . By adopting cooperative up-sampling, after m times up-sampling, generation can be formulated in terms of (16) as:

$$\varphi(G_B(z_B), G_O(z_O)) \approx O = F(\theta^m) = F(U(\theta^{m-1}, \beta^{m-1})), \quad (18)$$

where F is the mapping function (Conv 3×3) from θ^m to O , U denotes cooperative up-sampling. Considering that θ^{m-1} and β^{m-1} are all relevant to G_B , therefore, $\partial F(U(\theta^{m-1}, \beta^{m-1}))/\partial G_B \neq 0$, the updating formulation becomes:

$$\frac{\partial \mathcal{L}_{G_B, G_O}}{\partial G_B} = \frac{\partial \mathcal{L}_{G_B, G_O}}{\partial \varphi} \cdot \frac{\partial \varphi}{\partial G_B} \approx \frac{\partial \mathcal{L}_{G_B, G_O}}{\partial \varphi} \cdot \frac{\partial F(U(\theta^{m-1}, \beta^{m-1}))}{\partial G_B} \neq 0. \quad (19)$$

Such that even when M tends to have all elements equal to 1, G_B can be updated by interflow. Therefore, the proposed dual generator structure should benefit from adopting cooperative up-sampling in two aspects: (i) easy synchronization between G_O and G_B ; (ii) prevention of irreversible degeneration.

IV. EXPERIMENTAL RESULTS

A. DATASET

We use Caltech-UCSD Bird¹(CUB) [28] and Oxford-102 dataset² [29] to evaluate our DGattGAN. CUB and

¹<http://www.vision.caltech.edu/visipedia/CUB-200-2011.html>

²<http://www.robots.ox.ac.uk/vgg/data/flowers/102/>

$$\mathcal{L}_{D_B} = \underbrace{\gamma(-\mathbb{E}_{x_B \sim p_{x_B}} [\log(D_{B_rf}(x_B))] - \mathbb{E}_{s \sim p_s} [\log(1 - D_{B_rf}(G_B(z_B)))]]}_{\text{real/fake loss}} - \underbrace{\mathbb{E}_{x_B \sim p_{x_B}} [\log(D_{B_cls}(x_B))] - \mathbb{E}_{x_O \sim p_{x_O}} [\log(1 - D_{B_cls}(x_O))]}_{\text{background/object loss}}, \quad (11)$$

Oxford-102 provide object description texts that describe visual attributes of object with less background information. CUB dataset consists of 8855 training images from 150 species of birds and 2933 testing images from 50 other species. Oxford-102 dataset contains 7034 training images from 82 species of flowers and 1155 testing images from another 20 species. Both datasets provide 10 text descriptions for each image. CUB dataset marks all objects using bounding boxes and Oxford-102 gives objects' segmentation maps.

B. EVALUATION METRICS

The performance of DGattGAN is evaluated using two quantitative metrics Inception Score (IS) [30] and R-precision [2]. Instead of human annotators, these two metrics are regarded as alternatives for efficient human evaluation over a large number of synthesized data. IS is used to evaluate quality and diversity of generated images. Evaluating this score needs a pre-trained Inception v3 model [26] on ImageNet [31]. A higher value of IS indicates better visual diversity and quality. As reported in [30], IS calculates the KL-divergence D_{KL} between the conditional class distribution $p(y|x)$ and the marginal class distribution $p(y)$:

$$IS(G) = \exp(\mathbb{E}_{x \sim p_g} D_{KL}(p(y|x)||p(y))), \quad (20)$$

where p_g is generation distribution of the model, x is a generated image sampled from p_g . In this article, we use the fine-tuned Inception v3 model provided by Zhang et al. [6] that is more valuable for fine-grained dataset to predict class labels on testing images.

R-precision is proposed to evaluate text consistency of generated images in text-to-image synthesis task by Xu et al [2]. Given a single image, we pick R text descriptions consistent with it and $N - R$ descriptions inconsistent with it. After global image features and N sentence-level features are encoded respectively by image and text encoders, the cosine similarities between them are calculated. Finally, candidate text descriptions are ranked in descending similarity. If r text descriptions are found consistent in the top R ranked results, the R-precision is computed as r/R . We take $R = 1, N = 100$ in our experiments referring to many other text-to-image algorithms [2], [15], [32], etc.

C. IMPLEMENTING DETAILS

Original images with objects location labels are resized to 128×128 as real background samples for background discriminator training. For D_L and D_H , object areas mostly occupying only small region in CUB would be cropped and resized into 64×64 and 128×128 as real image samples. Sentence-level features are 256-dimension fed together with resized images. DGattGAN's generators and discriminators are trained using ADAM solver [33] with batch size 16 and an initial learning rate of 0.0002. According to [33], ADAM's parameters β_1 and β_2 are set to 0.5 and 0.999 in all experiments. The Conv 1×1 interflow used in

cooperative up-sampling quarter the number of feature channels. Parameter γ is set to 10 referring to FineGAN [9] and λ is set to 5 by experiments shown in section IV.G. All sentences in test set are utilized to generate test samples. Totally, 29330 samples on CUB and 11550 samples on Oxford-102 are obtained to evaluate DGattGAN. Our experiments are conducted on the platform where Python version is 3.7 and Pytorch version is 1.5.1. The network architecture is built on a server with Intel (R) Core (TM) i7-9700K CPU (4.9 GHz) with 32GB memory, NVIDIA GeForce RTX 2080Ti (GPU) with 11GB of memory.

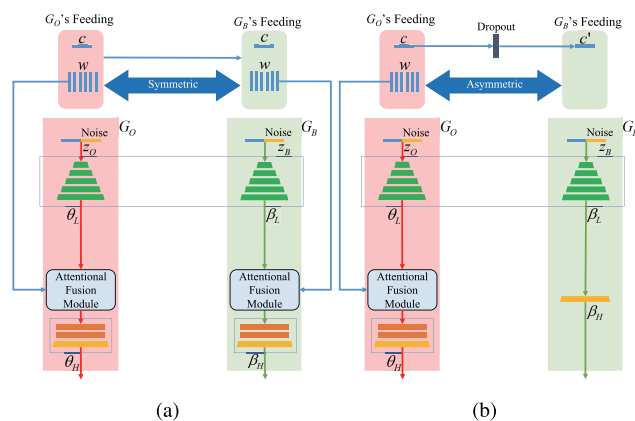


FIGURE 6. Schematic of the two information feeding schemes, (a) symmetric; (b) asymmetric.

D. EVALUATION ON ASYMMETRIC INFORMATION FEEDING SCHEME

In this study, asymmetric information feeding is a special scheme when training dual generator, which plays an important role in dominating each generator's synthesis behavior. To better show the effect of asymmetric information feeding, as shown in Figure 6(a), we conduct a symmetric information feeding training for comparison with following modifications: remove dropout layer before G_B , so that G_B can obtain all sentence-level information; fuse word-level features with β_L using the same attentional fusion module in G_O , so that G_B can obtain all word-level information. Symmetric information feeding indicates that both G_O and G_B obtain the same entire object information. Figure 6(b) shows the schematic of the asymmetric information feeding scheme. Figure 7 presents some examples of images generated under symmetric/asymmetric information feeding schemes. One can found from this figure, symmetric information feeding strategy leads to generators' confusing synthesis behaviour. More specially, the generated background may also depict object's visual attributes, so that the object mask has some 'hole' areas marked as background (see wings). This can be seen a failure in decoupling object and background, in that both G_O and G_B synthesize different parts of object. In contrast, asymmetric information feeding scheme helps both generators

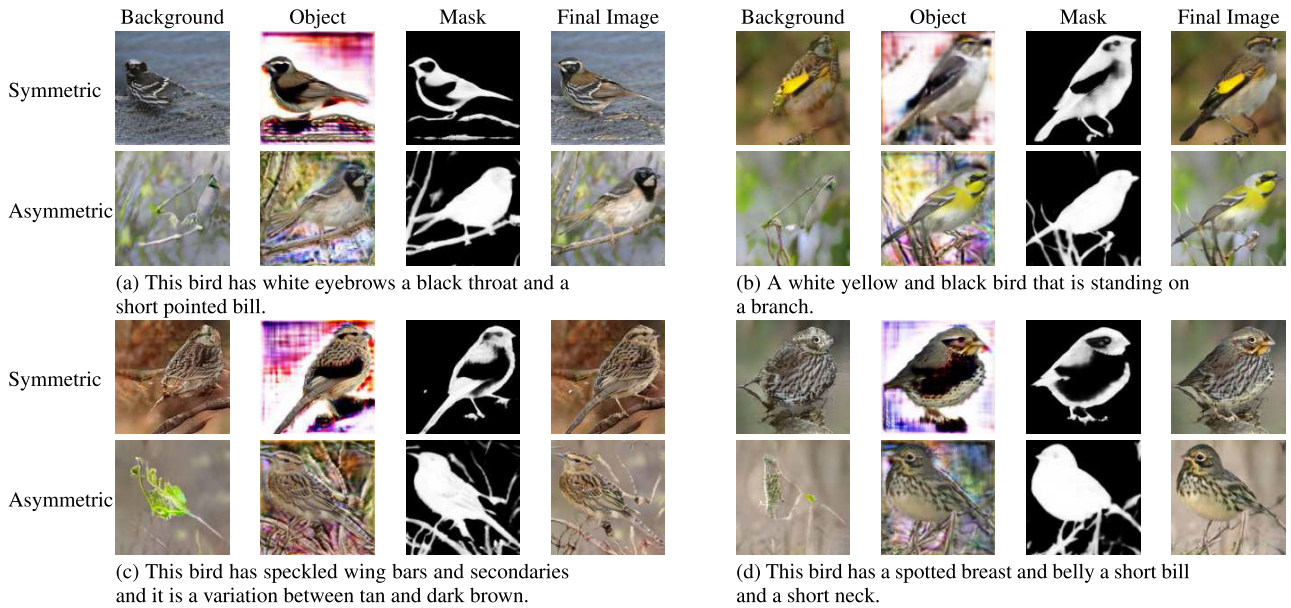


FIGURE 7. Comparisons of symmetric/asymmetric information feeding schemes. For each text description, the first and second rows show images generated by symmetric/asymmetric information feeding training methods. Background, objects, object masks and final results are listed from left to right.

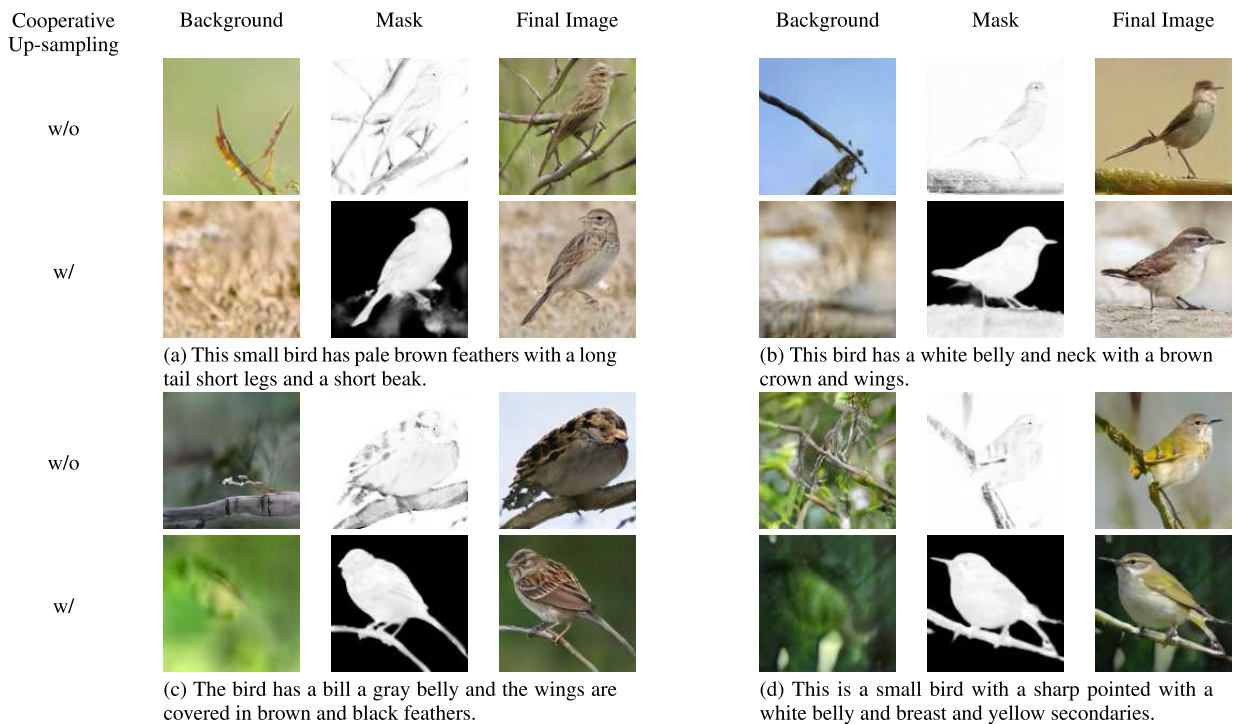


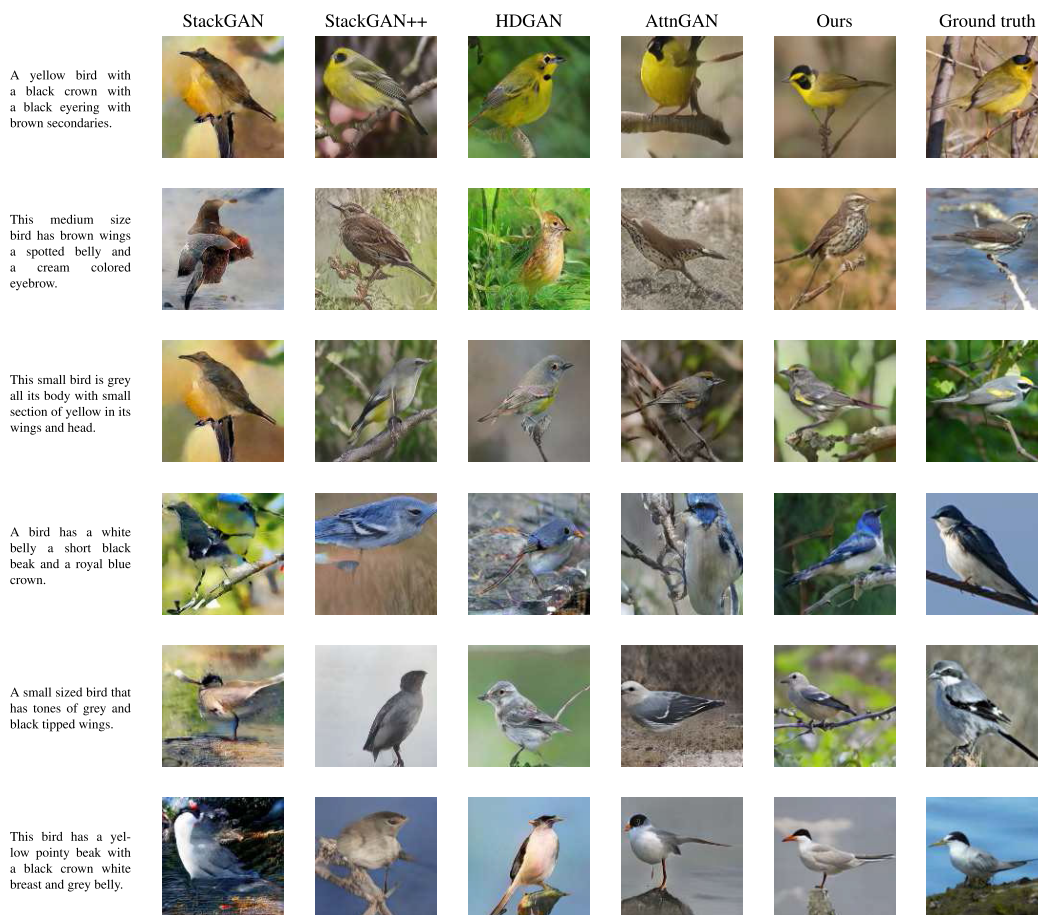
FIGURE 8. Ablation study of cooperative up-sampling mechanism. For each text description, the first and second row are generated images without (w/o) and with (w) cooperative up-sampling. Background, masks and final results are listed from left to right.

of DGattGAN successfully implement their different tasks. To validate the effectiveness of the proposed asymmetric scheme, we conducted image synthesizing experiment on all text images of CUB dataset. Table 1 lists both Inception Scores and R-precision. Symmetric information feeding leads to an evident decrease on R-precision by 4.52%, which may be caused by inappropriate fusion of background feature and

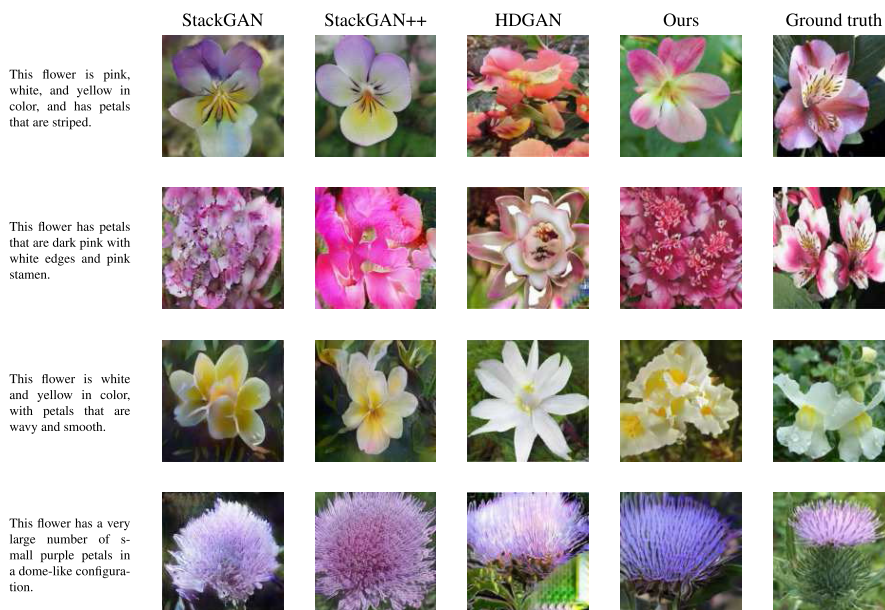
word-level feature. Using the proposed asymmetric scheme, the Inception Score is raised from 4.27 to 4.45.

E. ABLATION EXPERIMENT ON COOPERATIVE UP-SAMPLING MECHANISM

To illustrate how our proposed cooperative up-sampling mechanism works, we ablate all interflows before each



(a) Subjective visual comparisons on CUB dataset



(b) Subjective visual comparisons on Oxford-102 dataset

FIGURE 9. Generated images on CUB and Oxford-102 datasets. In (a), from left to right are StackGAN [6], StackGAN++ [1], HDGAN [12], AttnGAN [2], Ours and Ground truth. In (b), from left to right are StackGAN, StackGAN++, HDGAN, Ours and Ground truth. Zoom-in for better observation.

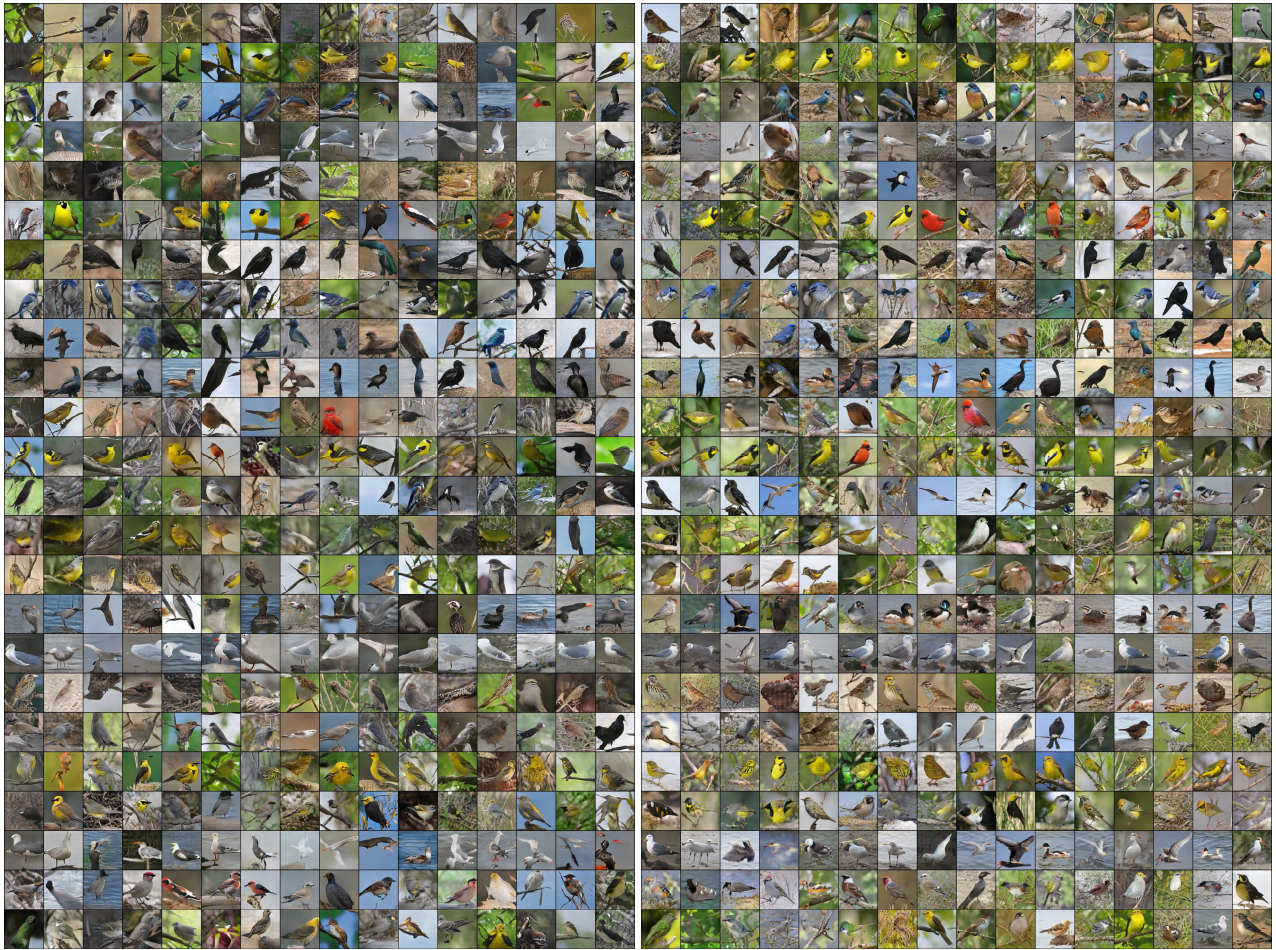


FIGURE 10. Comparisons of AttnGAN (left) and Ours (right). Same bird species are listed in same row for two methods ($\lambda = 5.0$ for both methods). Zoom-in for better observation.

up-sampling to make two generators mutually independent as shown in Figure 5(a). In this case, the cooperative up-sampling mechanism is degraded onto common up-sampling. Figure 8 shows synthesis results by models with (w/) and without (w/o) cooperative up-sampling mechanism. As we can see, masks generated without cooperative up-sampling tend to have all region marked as object and the backgrounds are incompatible with final images. This indicates that the generative model has degraded into a single generator structure. With cooperative up-sampling, however, synthesized masks with clear shape of the object are gained. Besides, dual generator becomes synchronized and irreversible degeneration could be well prevented. We also provide a quantity evaluation in our ablation study, as listed in Table 2. It could be seen that Inception Scores seen a significant increase from 4.32 ± 0.03 to 4.45 ± 0.05 , while R-precision also increase by about 1%. This reflects an overall increase on image quality, generation diversity and text consistency. Therefore, the proposed cooperative up-sampling scheme seem to well tackle previous dual generator difficulties and achieve improved performance in image generation.

TABLE 1. Quantity Evaluation With Different Information Feeding Schemes on Test Images of CUB Dataset, Highlighted Values Represent the Best Results.

	Inception Score	R-precision (%)
Symmetric	4.27(± 0.05)	57.93
Asymmetric	4.45(± 0.05)	62.45

F. SUBJECTIVE VISUAL COMPARISONS

Visual quality and text consistency are two key factors when evaluating generation quality in text-to-image synthesis task. As shown in Figure 9, it compares the generated images of StackGAN³ [6], StackGAN++⁴ [1], HDGAN⁵ [12], AttnGAN⁶ [2] and DGattGAN on CUB and Oxford-102 datasets. It could be seen from the experiments that StackGAN obtain the most basic quality of image generation where some artifacts appear among most results with rough object shapes generated. StackGAN++, HDGAN achieve relatively

³<https://github.com/hanzhanggit/StackGAN>

⁴<https://github.com/hanzhanggit/StackGAN-v2>

⁵<https://github.com/ypxie/HDGAN>

⁶<https://github.com/taoxugit/AttnGAN>

better smoothness and boundary coherence with few generated images lacking vivid object appearance. Besides, AttnGAN seems to outperform StackGAN, StackGAN++, HDGAN in generating vivid detail and clear object texture information such as “a spotted belly” and “a cream colored eyebrow” in the second row of Figure 9. However, AttnGAN shows relatively poor generation in realistic object outline. In general, our method shows advantages in terms of object shape and detailed information which is most likely to competently meet appearance-focused tasks. In terms of text consistency, StackGAN seems to perform less satisfied since similar image appears for different text description. AttnGAN and our method show better reflection on text meaning such as “yellow in its wings and head” and “black tipped wings” in Figure 9(a), and “dark pink with white edges” in Figure 9(b).

TABLE 2. Quantity Evaluation of DGattGAN With Cooperative Up-Sampling Mechanism or Not, Highlighted Values Represent the Best Results.

Cooperative Up-sampling	Inception Score	R-precision (%)
w/o	4.32(\pm 0.03)	61.20
w/	4.45(\pm0.05)	62.45

G. QUANTITY EVALUATION ON GENERATORS

We next quantitatively compare DGattGAN with several state-of-art algorithms: GAN-INT-CLS [5], GAWWN [34], StackGAN [6], StackGAN++ [1], HDGAN [12], AttnGAN [2], MirrorGAN [15] and SegAttnGAN [32], etc. Table 3 reports the Inception Score and R-precision on CUB and Oxford-102, and all figures are from their respective papers. Except for MirrorGAN (4.56 \pm 0.05), our DGattGAN achieves higher score than previous methods both on CUB and Oxford-102 datasets at 4.45 \pm 0.05 and 3.48 \pm 0.06 respectively. This indicates that our DGattGAN can generate more realistic and diversity images conditioned on text descriptions than most previous models. Although training models of some networks are not available publicly, we could still observe the preformation of AttnGAN, MirrorGAN and SegAttnGAN from these R-precision values mentioned in their papers shown in Table 4. From these R-precision values demonstrated in Table 4, we can observe that DGattGAN achieves the highest R-precision score at 62.45% compared with 53.31% and 57.67% of AttnGAN and MirrorGAN. This indicates consistency with text descriptions on synthesizing images of this proposed DGattGAN. According to Tables 3 and 4, MirrorGAN achieves the highest Inception Score, which is approximately 0.1 higher than ours. However, they show a slightly lower R-precision at 57.67%, while our approach arrives at 62.45 on R-precision. Therefore, this validates the effectiveness of our framework in terms of text-to-image consistency.

H. USER EVALUATION

In this section, a human perceptual evaluation is arranged on CUB test dataset. Totally, 30 volunteers with different

TABLE 3. The Inception Score Comparisons on CUB and Oxford-102 Datasets, Highlighted Values Represent the Best Results.

Models	Datasets	
	CUB	Oxford-102
GAN-INT-CLS [5]	2.88(\pm 0.04)	2.66(\pm 0.03)
GAWWN [34]	3.62(\pm 0.07)	-
StackGAN [6]	3.70(\pm 0.04)	3.20(\pm 0.01)
StackGAN++ [1]	4.04(\pm 0.05)	3.26(\pm 0.01)
HDGAN [12]	4.15(\pm 0.05)	3.45(\pm 0.07)
AttnGAN [2]	4.36(\pm 0.03)	-
MirrorGAN [15]	4.56(\pm0.05)	-
SegAttnGAN [32]	4.44(\pm 0.06)	3.36(\pm 0.08)
Ours	4.45(\pm 0.05)	3.48(\pm0.06)

TABLE 4. The R-Precision Comparisons on CUB Dataset, Highlighted Values Represent the Best Results.

Models	R-precision (%)
AttnGAN [2]	53.31
MirrorGAN [15]	57.67
SegAttnGAN [32]	52.29
Ours	62.45

professional backgrounds are recruited to conduct two tests: recognition test and text consistency test. In specific, recognition test is presented to compare the recognizability and reality of synthesized object using different methods, while text consistency test is applied to compare matching degree between text and image. Each participant is presented with 100 groups of images, in which 4 images from StackGAN++, HDGAN, AttnGAN and DGattGAN are arranged in random order given by the same text description. In recognition test, participants are asked to select the most realistic and recognizable image among 4 results. In text consistency test, participants are asked to select the image that best matches given text. The statistical results shown in Table 5 demonstrate that images of DGattGAN gain 42.07% and 43.36% of preference for recognition and text consistency respectively. In addition, the proposed method seems to gain more satisfied visual results from human perception as compared to quantitative analysis.

TABLE 5. The Vote Rate of Each Model in all Recognition Tests and Text Consistency Tests. In That There are 30 Volunteers and 100 Groups for Each Volunteer, We Have Totally 3000 Samples in Each Test.

	Recognition Test (%)	Text Consistency Test (%)
StackGAN++ [1]	15.43	12.10
HDGAN [12]	13.27	13.17
AttnGAN [2]	29.23	31.37
DGattGAN	42.07	43.36

I. QUANTITATIVE COMPARISON ON MODEL COMPLEXITY

Since the proposed DGattGAN mainly aims at improving generation quality based on existing methods, model complexity also raised to some extent. In this experiment, number

of network parameters and floating-point operations (FLOPs) for several competitive methods that synthesize images of similar level of resolution are illustrated in Table 6. It could be seen that StackGAN, StackGAN++ and HDGAN have noticeably higher model complexity compared to other methods. HDGAN has the most complicated structure, where a high parameter number (41.23M) and FLOPs (28.97G) are shown. This is mainly because of its deepest network architecture with residual blocks stacked after each up-sampling layer. Also, StackGAN and StackGAN++ have slightly larger amount of computation due to their wide structure resulted from high dimensional text embedding. In addition, MirrorGAN and AttnGAN have the least computational cost as these approaches take a similar structure of less complexity. In general, our model has a medium number of parameters and FLOPs among compared approaches at 15.19M and 3.23G, respectively. In specific, the dual generator structure with feature interflow by cooperative up-sampling almost doubles the number of parameters compared to AttnGAN. Nevertheless, background generator in DGattGAN of less significant concern on generation quality only leads to limited increase in terms of computational cost.

TABLE 6. Quantitative Comparisons of Parameters and FLOPs. All Generative Models Synthesize 128×128 Images.

	Parameters (M)	FLOPs (G)
StackGAN [6]	27.13	10.37
StackGAN++ [1]	16.33	4.73
HDGAN [12]	41.23	28.97
MirrorGAN [15]	6.82	2.17
AttnGAN [2]	6.81	2.14
DGattGAN	15.19	3.23

J. MORE COMPARISONS WITH BASELINE

Next, we show more experimental results by comparing DGattGAN with baseline AttnGAN [2] as used by many start-of-art models [15], [16], [18], [32]. AttnGAN mentioned that the coefficient of DAMSM loss term λ greatly affects both generation quality and text consistency. Here, we set λ as 0.1, 1.0 and 5.0 to train DGattGAN respectively, while AttnGAN trained using the same λ is illustrated as well for comparisons. To avoid impact of different resolutions, we set synthesizing size at 128×128 for AttnGAN the same as DGattGAN. An example ($\lambda = 5.0$) of visual comparison of AttnGAN and DGattGAN on CUB dataset can be found in Figure 10. We found from this figure, although there are still some images generated by DGattGAN shows unsatisfied object shapes, general quality verifies that DGattGAN has achieved great improvement in synthesizing more vivid objects. Figure 11 shows the quantitative evaluation on the CUB dataset. It could be observed that our DGattGAN achieves higher Inception Score and R-precision scores than AttnGAN at all values of λ . In this figure, all results for AttnGAN are referenced from [2]. All these validate that

DGattGAN can generate more realistic and diversity images that consistent with text descriptions.

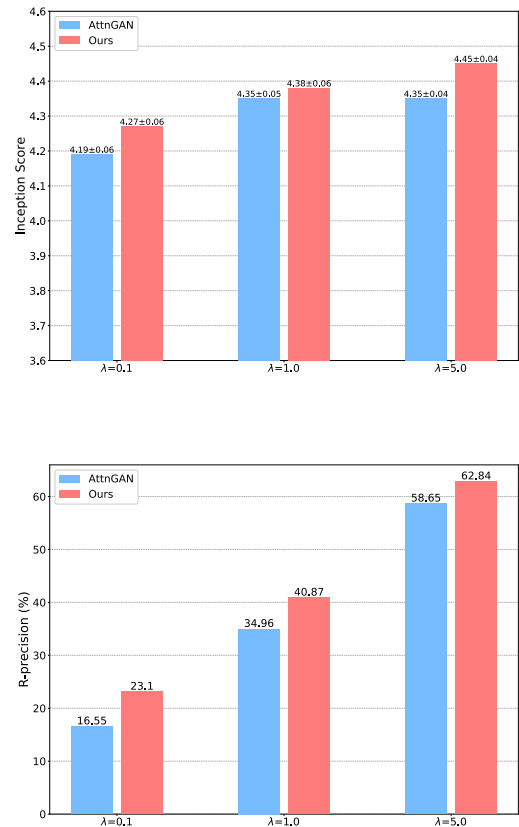


FIGURE 11. Inception Score and R-precision by AttnGAN and DGattGAN with different λ on CUB test set.



FIGURE 12. Failure cases of the proposed model on CUB dataset, text images with size of 128×128 .

K. FAILURE CASE ANALYSIS

Although our model shows an improvement on existing approaches, there are still some less satisfied generating images. In the last section, we examine the generated images and find some bad cases on CUB dataset as shown in Figure 12. These less satisfied generations are mostly related to aquatic birds. From model aspect, DGattGAN's object generator synthesizes accurate object masks with the tendency to learn objects' shape. Aquatic and arboreal birds differ considerably in terms of morphology (aquatic birds are usually duck-like). Initially, distinguishable latent representation is accomplished by CA module in DGattGAN that maps related text embeddings to Gaussian distributions with a large difference in mean and variance. This architecture

would be effective if different-shape objects are uniformly distributed in dataset, such as in Oxford-102, since their latent Gaussian distributions have similar prior probabilities. However, a relative proportion of arboreal and aquatic birds at approximately 9:1 exists in CUB dataset, which possibly results in an insufficient training of the latent distribution of aquatic birds. Future studies might draw attention on improving generation quality of species occupying of relatively fewer numbers in training dataset.

V. CONCLUSION

We have proposed a novel dual generator attentional GAN based on cooperative up-sampling scheme for text-to-image synthesis. The proposed generator architecture provides a more thorough alternative for decoupling object and background distribution space. Unsynchronized issue in existing dual generator models could also be solved by the proposed cooperative up-sampling mechanism, which is considered valuable for any dual generator design. Two optimization strategies to harmonize synthesis behaviour accompanying with dual generator is explored. In particular, the asymmetric information feeding scheme is introduced as a novel training scheme for dual generator. Additionally, word-level feature fusion on targeted object would be improved by these designs and contributes to better text-to-image generation quality. Experimental results on standard dataset showed that DGattGAN achieves better performance in synthesizing diverse photo-realistic and text-consistent images. For future improvement, modifications could be made towards more satisfied generation of a specific object category among the whole dataset and targeted adjustment for application tasks.

REFERENCES

- [1] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1947–1962, Aug. 2019.
- [2] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.
- [3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [4] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: <http://arxiv.org/abs/1411.1784>
- [5] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text-to-image synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1–11.
- [6] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5907–5915.
- [7] X. Chen, Y. Duan, R. Houthoofd, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2172–2180.
- [8] J. Yang, A. Kannan, D. Batra, and D. Parikh, "LR-GAN: Layered recursive generative adversarial networks for image generation," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–21.
- [9] K. K. Singh, U. Ojha, and Y. J. Lee, "FineGAN: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6490–6499.
- [10] J. Liu, K. Wang, C. Xu, Z. Zhao, R. Xu, Y. Shen, and M. Yang, "Interactive dual generative adversarial networks for image captioning," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11588–11595.
- [11] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 613–621.
- [12] Z. Zhang, Y. Xie, and L. Yang, "Photographic text-to-image synthesis with a hierarchically-nested adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6199–6208.
- [13] J. Sun, Y. Zhou, and B. Zhang, "ResFPA-GAN: Text-to-image synthesis with generative adversarial network based on residual block feature pyramid attention," in *Proc. IEEE Int. Conf. Adv. Robot. Social Impacts (ARSO)*, Oct. 2019, pp. 317–322.
- [14] M. Yuan and Y. Peng, "Bridge-GAN: Interpretable representation learning for text-to-image synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4258–4268, Nov. 2020.
- [15] T. Qiao, J. Zhang, D. Xu, and D. Tao, "MirrorGAN: Learning text-to-image generation by redescription," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1505–1514.
- [16] Y. Cai, X. Wang, Z. Yu, F. Li, P. Xu, Y. Li, and L. Li, "Dualattn-GAN: Text to image synthesis with dual attentional generative adversarial network," *IEEE Access*, vol. 7, pp. 183706–183716, 2019.
- [17] A. Tian and L. Lu, "Attentional generative adversarial networks with representativeness and diversity for generating text to realistic image," *IEEE Access*, vol. 8, pp. 9587–9596, 2020.
- [18] J. Cheng, F. Wu, Y. Tian, L. Wang, and D. Tao, "RiFeGAN: Rich feature generation for text-to-image synthesis from prior knowledge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10911–10920.
- [19] S. Sharma, D. Suhubdy, V. Michalski, S. E. Kahou, and Y. Bengio, "ChatPainter: Improving text to image generation using dialogue," 2018, *arXiv:1802.08216*. [Online]. Available: <http://arxiv.org/abs/1802.08216>
- [20] H. Dong, J. Zhang, D. McIlwraith, and Y. Guo, "I2T2I: Learning text to image synthesis with textual data augmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2015–2019.
- [21] B. Zhu and C.-W. Ngo, "CookGAN: Causality based text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5519–5527.
- [22] S. Hong, D. Yang, J. Choi, and H. Lee, "Inferring semantic layout for hierarchical text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7986–7994.
- [23] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1219–1228.
- [24] W. Li, P. Zhang, L. Zhang, Q. Huang, X. He, S. Lyu, and J. Gao, "Object-driven text-to-image synthesis via adversarial training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12174–12182.
- [25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
- [26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
- [27] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.
- [28] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.
- [29] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 722–729.
- [30] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 2234–2242.
- [31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

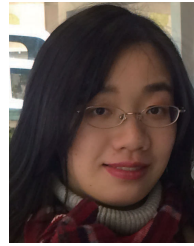
- [32] Y. Gou, Q. Wu, M. Li, B. Gong, and M. Han, "SegAttnGAN: Text to image generation with segmentation attention," 2020, *arXiv:2005.12444*. [Online]. Available: <http://arxiv.org/abs/2005.12444>
- [33] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, San Diego, CA, USA, May 2015, pp. 1–15.
- [34] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 217–225.



HAN ZHANG received the B.S. degree from the School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China, in 2019, where he is currently pursuing the Ph.D. degree with the Department of Electronics and Communication Engineering. His research interests include text-to-image generation and machine reading comprehension in natural language processing.



HONGQING ZHU (Member, IEEE) received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2000. From 2003 to 2005, she was a Postdoctoral Fellow with the Department of Biology and Medical Engineering, Southeast University, Nanjing, China. She is currently a Professor with the East China University of Science and Technology, Shanghai. Her current research interests include medical image processing, deep learning, computer vision, and pattern recognition. She is a member of IEICE.



SUYI YANG is currently pursuing the B.Sc. degree with the Department of Mathematics, Natural, Mathematical & Engineering Sciences, King's College London. Her interests include mathematical modeling and mathematical problems in image processing, especially partial differential equations.



WENHAO LI received the B.S. degree from the School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China, in 2019, where he is currently pursuing the Ph.D. degree with the Department of Electronics and Communication Engineering. His current research interests include text-image matching, image-text retrieval, machine learning, and computer vision.

...