IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 5, 2019, accepted November 24, 2019, date of publication December 10, 2019,
date of current version December 27, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2958864

Dualattn-GAN: Text to Image Synthesis With Dual Attentional Generative Adversarial Network

YALI CAI[1], XIAORU WANG[1], ZHIHONG YU[2], FU LI[3], PEIRONG XU[1], YUELI LI[1], AND LIXIAN LI[1]

[1]Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Intel China Research Center, Beijing 100190, China
[3]Department of Electrical and Computer Engineering, Portland States University, Portland, OR 97207-0751, USA

Corresponding author: Xiaoru Wang (wxr@bupt.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61672108.

ABSTRACT Recent generative adversarial network based methods have shown promising results for the charming but challenging task of synthesizing images from text descriptions. These approaches can generate images with general shape and color but often produce distorted global structures with unnatural local semantic details. It is due to ineffectiveness of convolutional neural networks in capturing the high-level semantic information for pixel-level image synthesis. In this paper, we propose a Dual Attentional Generative Adversarial Network (DualAttn-GAN) in which the dual attention modules are introduced to enhance local details and global structures by attending to related features from relevant words and different visual regions. As one of the dual modules, the textual attention module is designed to explore the fine-grained interaction between vision and language. On the other hand, visual attention module models internal representations of vision from channel and spatial axes, which can better capture the global structures. Meanwhile, we apply an attention embedding module to merge multi-path features. Furthermore, we present an inverted residual structure to boost representation power of CNNs and apply spectral normalization to stabilize GAN training. With extensive experimental validation on two benchmark datasets, our method significantly improves state-of-the-art models over the evaluation metrics of inception score and Fréchet inception distance.

INDEX TERMS Generative adversarial network, textual attention, visual attention, inverted residual structure, spectral normalization.

I. INTRODUCTION

Synthesizing image from text description has been a hot topic crossing natural language processing and computer vision. It has significant impact on the applications of content production and advertisement design.

The core challenge of text-to-image synthesis lies in generating visually realistic and semantically sensible pixels associated with text descriptions. The task has two main issues. One is the semantic gap between the word-level textual semantic conception and the pixel-level visual information. Due to the sparse mapping between text space and image space, one word may change some sub-region details of generated images. On the other hand, the incomplete text description lacks a lot of conditional information, which limits the ability to express the visual characteristics of the network. The under-fitting expression of the generative network results in poor quality of the generated image that appears to have a distorted global structure with blurred edges. In this paper, we aim to boost representation power of generative network and generate text-related images with high authenticity and quality.

Methods [1]–[7] built upon Generative Adversarial Networks (GANs) [8] for text-to-image synthesis have shown good performance. And the attention mechanisms, one of the recent advances in neural networks, have been used in generative adversarial network for text-to-image synthesis. AttnGAN [4] is the first to introduce attention mechanisms into GAN frameworks, which guides to generate refined details. However, it still can not perfectly capture the global coherence structure, which may distort the shape of object.

To address the issues above, we propose the Dual Attentional Generative Adversarial Networks for synthesizing image from text description. In order to synthesize promising

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

local details with significant global structures, we introduce dual attention modules, which includes three main components: textual attention module, visual attention module and attenion embedding module. The textual attention module is designed to explore the fine-grained interaction between vision and language. While visual attention module models internal representations of vision from channel and spatial axes, which can better capture the global structures. Moreover, we apply an attention embedding module to merge multi-path features. When reconstructing feature maps from low-resolution to high-resolution, most iterative methods [4]–[7] for text-to-image synthesis merely consider that adjacent two-layers' feature maps are almost linear or polynomial. In this paper, we present an inverted residual structure to modeling the nonlinear relationship between hidden layers.

Our contributions are summarized as the following: First, we propose Dual Attention Modules, in which enhance local and global details by attending to related features from relevant words and different visual regions. Second, we introduce an inverted residual structure (residuals [9] enhancements) to boost representation power of CNNs. Third, we apply several techniques including using ReLU [10] instead of the gate linear unit (GLU) [11] as activation function to improve training speed, spectral normalization [12] to stabilize GAN training based on the stacked generative network [4], [6]. Comprehensive experiments and results analysis show that our proposed DualAttn-GAN has a significant improvement over the previous state-of-the-art models.

II. RELATED WORK

A. TEXT-TO-IMAGE SYNTHESIS

The aim of text-to-image synthesis is generating realistic images relevant to the language descriptions, which has attracted many researches to tackle the new task with different deep generative models. Mansimov *et al.* [13] used the AlignDRAW model for iteratively generating images with a soft attention mechanism based on variational autoencoders [14]. Reed *et al.* [15] introduced conditional PixelCNN [16] to synthesize images from captions with a multi-scale model structure.

Recently, more and more researchers focus on generative adversarial networks (GANs) [8]. Most methods have been proposed to stabilize training GANs and improve the quality of generated images [12], [17]–[25]. Specifically, Miyato *et al.* [12] proposed spectral normalization to stabilize the training of the discriminator. Mirza and Osindero [22] presented CGAN to control generating images with labeled condition. LAPGAN [20] is the first to use iterative approach to generate sharper image from coarse to fine based on GAN, which outperforms the original GAN. Meanwhile, many works on other tasks have shown remarkable performance by using GAN, such as domain transfer [26], [27], super-resolution [28], [29] and human face generation [30], [31]. As for text-to-image synthesis based on GAN, there have been many effective methods emerged. Reed *et al.* [3] proposed GAN-INT-CLS and firstly applied GAN to generate images from text descriptions. Their follow-up work GAWWN [32] generated impressive images with auxiliary location constraints. Nguyen *et al.* [2] presented PPGN by using activation maximization [33] to generate images. Dash *et al.* [1] proposed TAC-GAN, which combines GAN-INT-CLS [3] and ACGAN [23]. To generate high-resolution images from neural languages, [4]–[7] produced sharper images from coarse to fine with stacked GAN. StackGAN [5] was put forward with staged generative adversarial network that generate low-resolution in the stage-I GAN and synthesize high-resolution details in the stage-II GAN. Then, Zhang *et al.* followed work StackGAN++ [6] advanced the stacked generative network within a tree-like structure and color-consistency regularization. Based on StackGAN++, Xu *et al.* [4] apply attention mechanism on vision and language features to form refined details. Zhang *et al.* [7] adopted global and grid adversarial losses to render image details with a hierarchically-nested network.

B. ATTENTION MECHANISMS

As the attention mechanisms aim to focus on the necessary parts of the inputs, the attention-based approach shows good performance in a range of tasks.

Textual attention mechanisms are usually designed to find semantic alignment between the inputs and the outputs. In this task, it is obvious that we can build up the attention model to bridge language (as the inputs) and vision (as the outputs). AttnGAN [4] firstly synthesizes fine-grained details by applying attention mechanism, which learns the relationship between text and image. Different from AttnGAN, we consider the attention module as a residual learning network, which makes the synthesized image more natural and authentic.

While visual attention mechanisms allow the model to enhance the representation of visual interests. A number of methods have recently adopted visual attention to benefit image classification [34], [35], image detection [36], image generation [18], [25], [37], image captioning [38], visual question answering [39]–[41]. As for the visual attention, there are two main directions for modeling the attention map, channel-wise and spatial axes. In this study, we design the visual attention model to construct potential semantic relationships of internal visual features by the means of blending channel-wise and spatial information.

III. IMPROVED STACKED GENERATIVE ADVERSARIAL NETWORK

We first build our baseline model based on the most recent models for text-to-image synthesis [4]–[7], which can generate fine images from coarse images by using iterative methods with multi-generators.

A. COARSE-TO-FINE NETWORK ARCHITECTURE

As shown in Figure 1, the network architecture of our improved model consists of multi-generators and

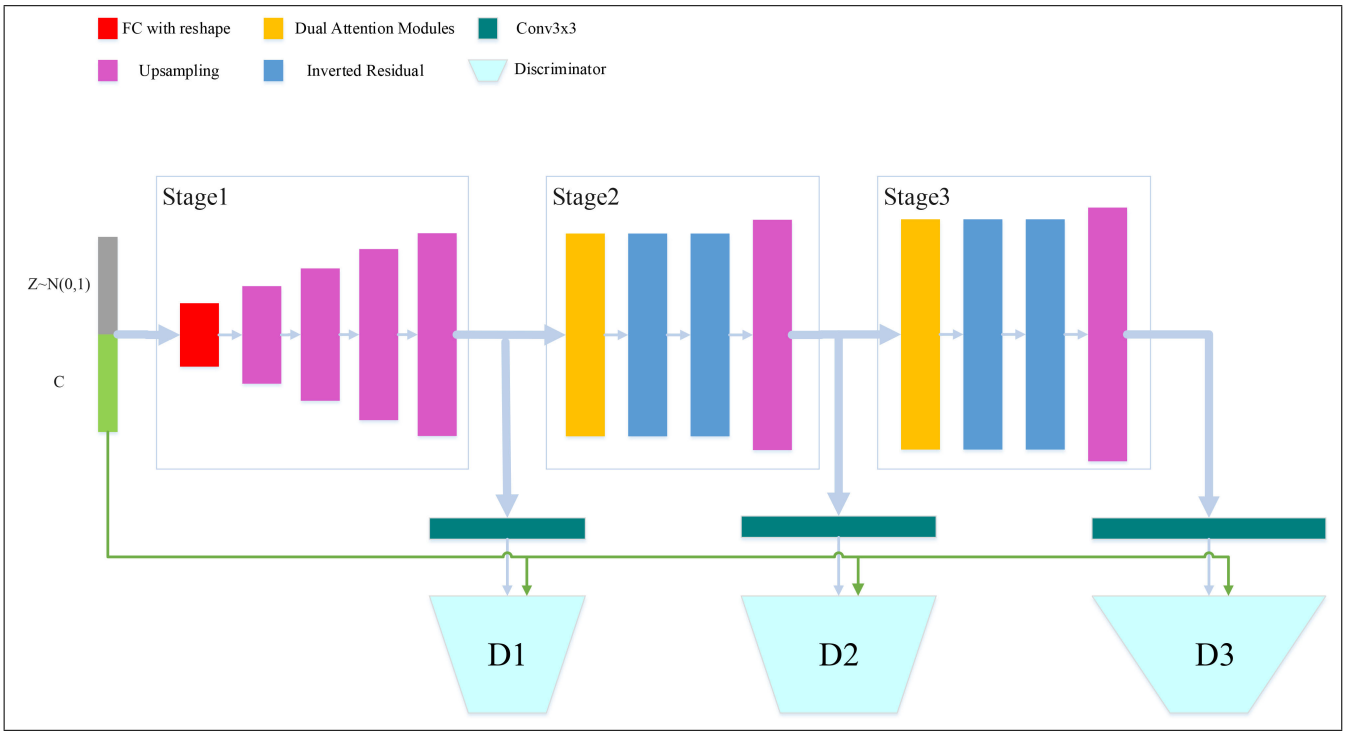


FIGURE 1. Overview of our improved stacked generative adversarial network, the coarse network of the first stage generates 64 ∗ 64 images, while the refined networks of the second stage and the third stage generate 128 ∗ 64128 and 256 ∗ 64256 images with dual attention modules.

multi-discriminators, and it follows a tree-like architecture [6] and uses a text encoder by BiLSTM [4] for training and testing. The first branch generates a low-resolution image which contains correct colors and rough structure for the object. The later branches generate high-resolution images that focuses on rich details. We introduce an inverted residual structure [42], in which the bottlenecks actually contain all the necessary information. In order to improve the quality of inter representations, we perform an expansion layer for increasing the width and use a squeeze layer for decreasing the width. Meanwhile, we use shortcuts directly between the two layers. Also, we apply ReLU as activation function instead of GLU in StackGAN++. To stabilize GAN training, we adopt spectral normalization to the generator networks and the discriminator networks, which shows great performance in recent models including SNGAN [12], SAGAN [25] and BigGAN [18].

B. CONDITIONAL-UNCONDITIONAL LOSSES AND DEEP MULTIMODAL SIMILARITY REGULARIZER

The objective of DualAttn-GAN is the joint conditional-unconditional losses [6] over each discriminator and generator, which is introduced to jointly approximates conditional-unconditional image distributions. The loss function for the i^{th} discriminator D_i is converted to:

$$\begin{aligned}\mathcal{L}_{D_i} = &-\frac{1}{2}\mathbb{E}_{x_i\sim P_{data_i}}[\log D_i(x_i)] \\ &-\frac{1}{2}\mathbb{E}_{\hat{x}_i\sim P_{G_i}}[\log(1-D_i(\hat{x}_i))] \\ &-\frac{1}{2}\mathbb{E}_{x_i\sim P_{data_i}}[\log D_i(x_i,\bar{e})] \\ &-\frac{1}{2}\mathbb{E}_{\hat{x}_i\sim P_{G_i}}[\log(1-D_i(\hat{x}_i,\bar{e}))] \end{aligned} \tag{1}$$

While the loss of the i^{th} generator G_i is computed as:

$$\begin{aligned}\mathcal{L}_{G_i} = &-\frac{1}{2}\mathbb{E}_{\hat{x}_i\sim P_{G_i}}[\log(D_i(\hat{x}_i))] \\ &-\frac{1}{2}\mathbb{E}_{\hat{x}_i\sim P_{G_i}}[\log(D_i(\hat{x}_i,\bar{e}))] \end{aligned} \tag{2}$$

where x_i is from the true data distribution P_{data_i} at the i^{th} scale, and $\hat{x}_i$ is from the model distribution P_{G_i} at the same scale.

In addition, we adopt the deep attentional multimodal similarity model (DAMSM) [4] to estimate how well the image matches the text. For given image-text pairs $\{I_i, T_i\}_{i=1}^{N}$, the DAMSM loss function can be defined as the negative log posterior probability that images and text descriptions match others:

$$\begin{aligned}\mathcal{L}_{DAMSM} = &-\sum_{i=1}^{N}\frac{\exp(\gamma R(I_i,T_i))}{\sum_{j=1}^{N}\exp(\gamma R(I_i,T_j))} \\ &-\sum_{i=1}^{N}\frac{\exp(\gamma R(I_i,T_i))}{\sum_{j=1}^{N}\exp(\gamma R(I_j,T_i))} \end{aligned} \tag{3}$$

where $R()$ donates the image-text matching score [4], and γ is a smoothing factor determined by experiments.

Finally, the DAMSM loss and the discriminant loss for the generative network are contributed to:

$$\mathcal{L}_G = \lambda \mathcal{L}_{DAMSM} + \sum_{i=1}^{M} \mathcal{L}_{G_i} \tag{4}$$

where M donates the total number of branches, and λ is a hyper-parameter to balance the two terms of loss.

IV. DUAL ATTENTION MODULES

Dual Attention Modules (DAM) are proposed to enhance details by the means of attending to related features from relevant words and different visual regions. As shown in Fig. 2, DAM consists of Textual Attention Module, Visual Attention Module and Attention Embedding Module. The Textual Attention Module (TAM) is designed to explore the fine-grained links between vision and language. While Visual Attention Module (VAM) models internal representations of vision from channel and spatial axes, which can better capture the global structures. Finally, Attention Embedding Module (AEM) is applied to merge multi-path features.

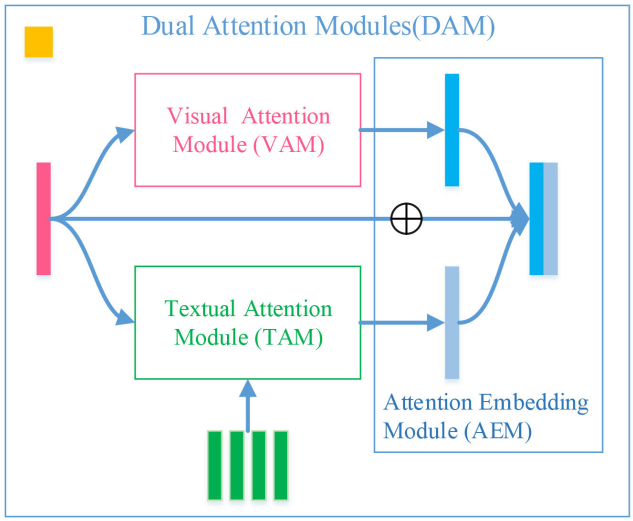


FIGURE 2. Dual Attention Modules (DAM). As illustrated, DAM contains Textual Attention Module (TAM), Visual Attention Module (VAM) and Attention Embedding Module (AEM). TAM model the links between vision and language, while VAM explore internal representations of vision. AEM is used to fuse multi-features.

A. TEXTUAL ATTENTION MODULE

Textual Attention Module [4] $F_t^{attn}(e, h)$ is presented to modeling the mapping relationship between the word features e and the visual features h from previous hidden layer, which can generate fine-grained visual details that are semantic connected to the text. First, we align the word and image features into the common semantic dimensions through a transform network. Then, the attention weights are obtained by calculating its relevancy of word and vision features within dot product and softmax normalization. Next, to acquired the word-context vector c_i for each sub-region of image, c_i is computed by weighted averaging on the word features e_i with the attention weights. Finally, the word-context matrix $\{c_0, c_1, \ldots, c_i, \ldots\}$ acquired by $F_t^{attn}(e, h)$, is passed to the Attention Embedding Module to fuse the features.

B. VISUAL ATTENTION MODULE

Visual Attention Module(VAM) $F_v^{attn}()$ is designed to improve the quality of representations. It can learn to use global information to selectively focus on important features and suppress unnecessary ones [36]. The proposed module utilizes information from dual perspectives, namely channel and spatial axes, to learn "what" and "where" are the informative parts.

As shown in Fig. 3, we define a Channel-Spatial (C-S) model that applies channel attention before spatial attention. At first, given the image features from the previous hidden layer h, we adopt Channel Attention Module $F_{v_c}^{attn}()$ to get channel-wise weighted attention map h'. Then we feed h' to the spatial attention module $F_{v_s}^{attn}()$ and obtain the spatial weighted attention map h''. $F_v^{attn}()$ can be defined as:

$$F_v^{attn}(h) = F_{v_s}^{attn}(F_{v_c}^{attn}(h)) \tag{5}$$

1) CHANNEL ATTENTION MODULE.

In order to gain "what" is important of image features, Channel Attention Module (CAM) is presented to mine the relationships between channel features of images. Taking into account the overall background and texture information, we summarize both average features and maximum features simultaneously. The work of CBAM [36] has verified the effectiveness of exploiting both features.

In Fig. 3 (A), we firstly assemble spatial feature vectors from two branches, one is applying global average pooling on the input features h, another is using global max pooling. Next, they are fed into a multi-layer perception to attain which feature maps should be attended. Finally, we obtain the channel-wise weighted attention map h' by leveraging an element-wise sum operation to combine the two output feature vectors and then performing an element-wise multiplication to the feature map h. All processes can be summarized as follows:

$$F_{v_c}^{attn}(h) = \sigma(MLP(AvgPool(h)) + MLP(MaxPool(h))) \otimes h \tag{6}$$

where σ denotes the sigmoid function, MLP donates the FC-ReLU-FC networks, and $\otimes$ denotes the element-wise multiplication.

2) SPATIAL ATTENTION MODULE.

Context relationship is critical to generating high-resolution details, with the goal of capturing global, long-range dependencies, not just local ones [25]. Therefore, we propose a spatial attention module that enhances local feature representations by encoding rich contextual information.

As illustrated in Fig. 3 (B), to compute the spatial attention efficiently, we firstly utilize average and max pooling on the feature map $h' \in \mathbb{R}^{C \times H \times W}$ to obtain effective features of

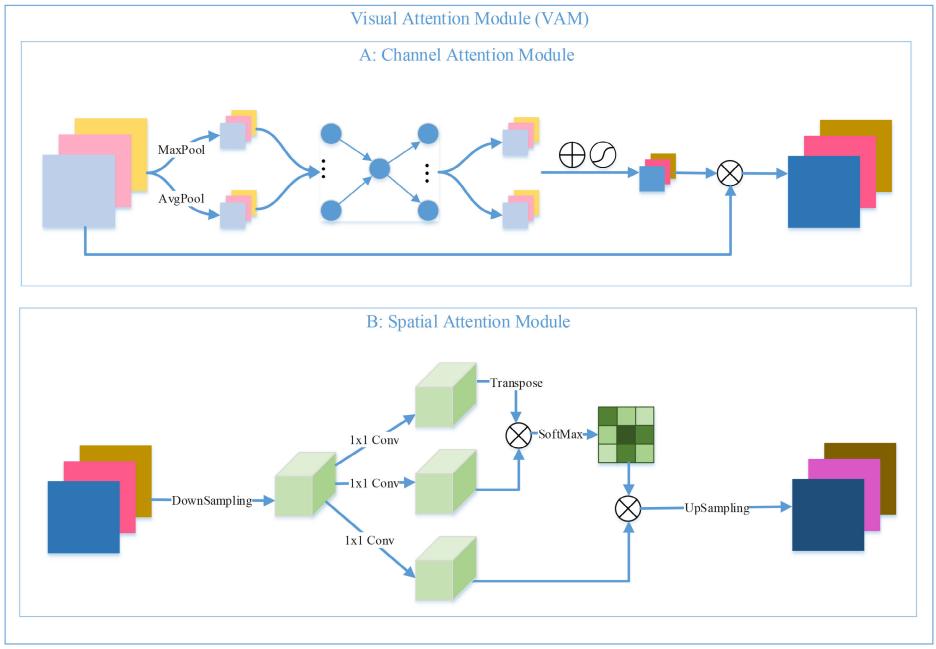


FIGURE 3. Visul Attention Module model the visual features from channel-wise and spatial dimensions. Channel Attention Module is proposed to focus on which layers' features are more important. Spatial Attention Module is applied to capturing global, long-range dependencies for each sub-region.

local regions f, where $f \in \mathbb{R}^{C\times\hat{H}\times\hat{W}}$. Next, we put it into a 1×1 convolution layer and a transform layer to acquire three feature maps θ, ϕ and g, where $\{\theta, \phi\} \in \mathbb{R}^{\hat{C}\times N}$, $g \in \mathbb{R}^{C\times N}$ and $N = \hat{H}\times\hat{W}$ is the number of features. Then the attention map $\beta \in \mathbb{R}^{N\times N}$ can be calculated as:

$$\beta_{j,i} = \frac{\exp(\theta_i \cdot \phi_j)}{\sum_{i=1}^{N} \exp(\theta_i \cdot \phi_j)} \tag{7}$$

$\beta_{j,i}$ indicates the extent to which the model attends to the i^{th} location when synthesizing the j^{th} region. Note that all sub-regions are contributed to generate details of each location.

To obtain the spatial weighted attention map $o \in \mathbb{R}^{C\times\hat{H}\times\hat{W}}$, we conduct a matrix multiplication between g and the transpose of β and reshape its result to $\mathbb{R}^{C\times\hat{H}\times\hat{W}}$. In addition, we multiply the output by a scale parameter η. All processes can be summarized as follows:

$$o_j = \eta \sum_{i=1}^{N} (\beta_{j,i} g_i) \tag{8}$$

Finally, we exploit up-sampling to get the image features h'', where $h'' \in \mathbb{R}^{C\times H\times W}$.

C. ATTENTION EMBEDDING MODULE

To enhance the representation of the image features, we summarize the functionality of these two attention modules. Specifically, as shown in Fig. 2, we respectively add the input features h to the outputs of the two attention modules through the skip-connection and perform concatenation to complete the feature fusion.

V. EXPERIMENTS

In order to thoroughly evaluate the proposed model, we conduct comprehensive experiments on two widely-used datasets: CUB [43] and Oxford-102 [44] datasets. Compared to several previous state-of-the-art GAN models for synthesizing image from text description, including GAN-INT-CLS [3], GAWWN [32], StackGAN [5], StackGAN++ [6], HDGAN [7] and AttnGAN [4], experimental results demonstrate that our model achieves better performance on the datasets. In addition, we conduct an ablation study to verify the effectiveness of important components in our proposed method.

A. DATASET

CUB [43] contains 11,788 images from 200 bird categories. **Oxford-102** [44] consists of 8,189 images of flowers from 102 categories. There are 10 captions for each image in the two datasets. We split CUB dataset into 150 training categories and 50 testing categories and split Oxford-102 dataset into 82 training categories and 20 testing categories by means of class-disjoint experimental settings. While training, each image is randomly cropped and flipped, and one caption is randomly selected from the 10 descriptions related to the image. During testing, all sentences in test set are chosen to generate images.

B. EVALUATION METRICS

We use three kinds of metrics to evaluate our method: Inception Score (IS), Fréchet Inception Distance (FID) and Human Rank (HR).

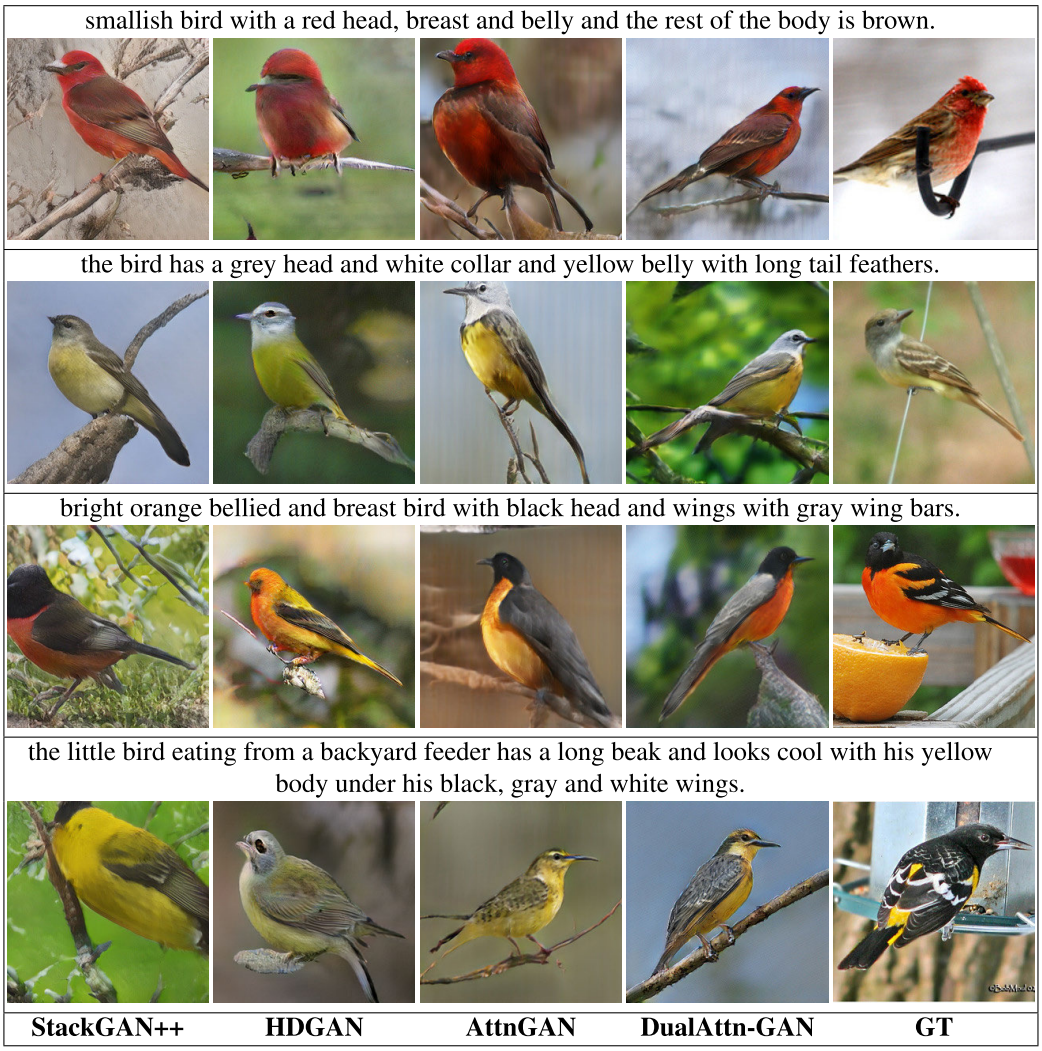


FIGURE 4. Generate Images on CUB test set by our proposed DualAttn-GAN, and compared models: AttnGAN, HDGAN and StackGAN++, which contain descriptions across different attributes.

1) INCEPTION SCORE [45]

It is a quantitative evaluation method that has been widely used for image generation. It can be donate as:

$$\mathrm{IS} = \exp(\mathbb{E}_x D_{KL}(p(y|x)\|p(y))) \tag{9}$$

where x denotes a generated sample, while y is the class label predicted by the Inception model [46]. High score indicates better model that can generates more diverse and meaningful images. We use the pre-trained Inception Model provided by [5].

2) FRÉCHET INCEPTION DISTANCE [47]

Obviously, the disadvantage of Inception Score is that the output samples are not compared to the ground truth images. It does not reflect whether the generated image is closer to the real image. Therefore, we introduce another evaluation metric, Fréchet Inception Distance, which is a more rule-based and comprehensive metric and has been shown to be more consistent with human assessments in assessing the authenticity and variability of generated samples. As is well-known, the top level of the pre-trained neural network can extract the advanced information of the image, which can reflect the essence of the image to a certain extent. In practice, images are encoded with visual features by the Inception model. FID can be calculated as:

$$\mathrm{FID} = \|\mu_r - \mu_g\|^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \tag{10}$$

where μ_r and μ_g are the means of real and synthetic image features respectively, and Σ_r and Σ_g are the covariance matrices of real and synthetic image features respectively. A lower FID value means that the distance is closer between the synthetic data distribution and the real data distribution.

To compute the IS and FID score for evaluating each model, we utilize all captions in the test set to generate samples.

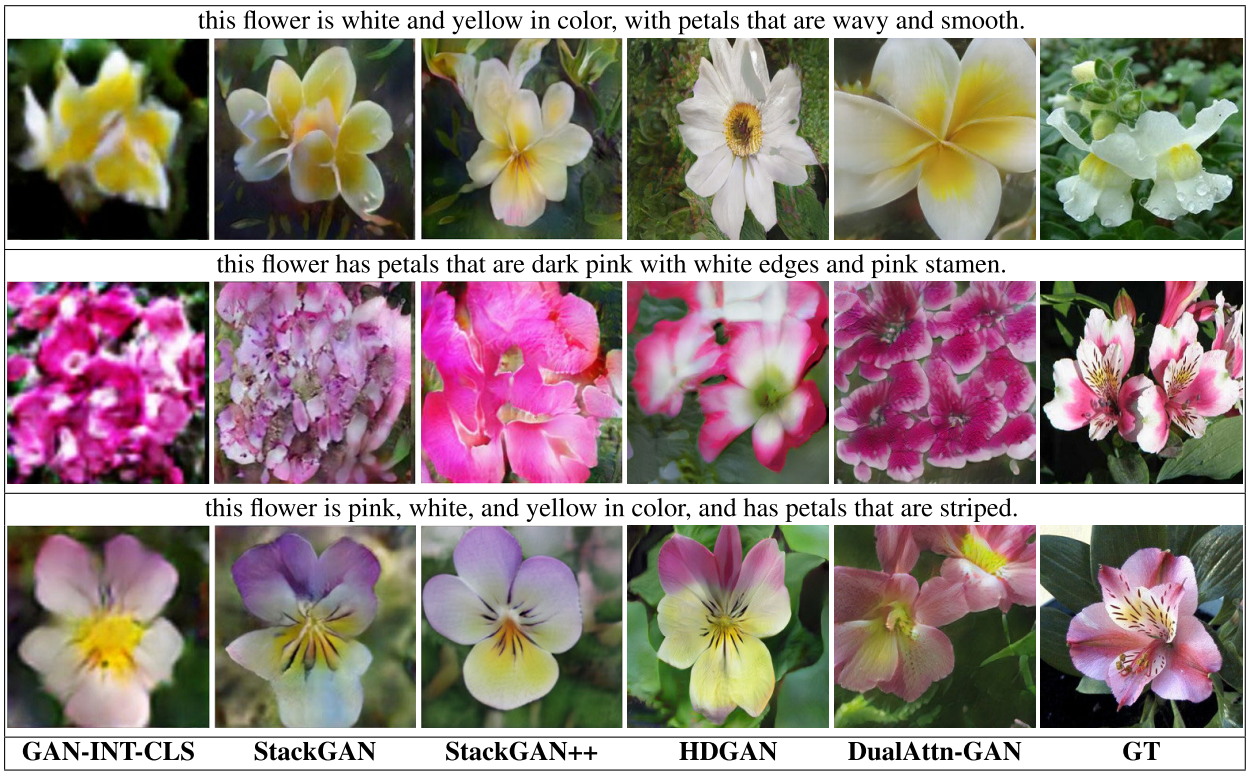


FIGURE 5. Example results on the Oxford-102 test set by proposed DualAttn-GAN, and previous methods: HDGAN, StackGAN++, StackGAN and GAN-INT-CLS.

3) HUMAN RANK

In addition, we conduct two human rank evaluations: **global quality rank (GQR)** and **local quality rank (LQR)** to evaluate generated images. The global quality rank indicates that users rank the generated images from the overall structures of object and the richness of background. The local quality rank means that visual reality and semantic consistency of local details. 3,000 sentences are randomly selected and used for synthesizing images by compared models. And 20 users (not including the authors) are asked to rank.

C. RESULTS AND ANALYSIS

In this section, we compare our results with the previous state-of-the-art models including GAN-INT-CLS [3], GAWWN [32], StackGAN [5], StackGAN++ [6], HDGAN [7] and AttnGAN [4] for text-to-image synthesis on the CUB and Oxford-102 datasets. In particular, we make detailed comparison among them by means of obtaining results from their provied models. The contrast results of IS, FID and HR are shown in Table 1 and Table 2. We sample 30,000 $\sim 256^2$ images from all text descriptions in test sets for computing the scores. IS is computed on the generated images, and FID measures the distribution distance between the synthetic images and the test set images. We also present the results of HR designed to test global and local quality. Compared with other methods, our DualAttn-GAN achieves significant improvements in the major evaluation metrics.

TABLE 1. Performances compared with the state-of-art on CUB and Oxford-102 datasets by inception score (IS) and fréchet inception distance (FID).

Method	CUB		Oxford-102	
	IS ↑	FID ↓	IS ↑	FID ↓
GAN-INT-CLS [3]	2.88 ± 0.04	-	2.66 ± 0.03	79.55
GAWWN [32]	3.62 ± 0.07	-	-	-
StackGAN [5]	3.70 ± 0.04	51.89	3.20 ± 0.01	55.28
StackGAN++ [6]	4.04 ± 0.05	18.35	3.26 ± 0.01	48.68
HDGAN [7]	4.15 ± 0.05	24.00	3.45 ± 0.07	43.17
AttnGAN [4]	4.36 ± 0.03	16.48	-	-
DualAttn-GAN	$\mathbf{4.59 \pm 0.07}$	**14.06**	$\mathbf{4.06 \pm 0.05}$	**40.31**

TABLE 2. Human rank of our DualAttn-GAN and compared methods on CUB and Oxford-102 datasets: global quality rank (GQR) and local quality rank (LQR).

Method	CUB		Oxford-102	
	GQR ↓	LQR ↓	GQR ↓	LQR ↓
StackGAN [5]	1.82	1.95	2.54	2.56
StackGAN++ [6]	1.67	1.63	2.29	2.43
HDGAN [7]	1.73	1.66	2.21	2.25
AttnGAN [4]	1.55	1.58	-	-
DualAttn-GAN	**1.21**	**1.43**	**1.77**	**1.90**

As shown in Table 1, our proposed model DualAttn-GAN significantly boosts in terms of IS StackGAN++ by 0.64, HDGAN by 0.44 and AttnGAN by 0.23 on CUB dataset, and StackGAN by 0.86, StachGAN++ by 0.80 and HDGAN by 0.61 on Oxford-102 dataset. It demonstrates that our model can synthesize more unambiguous and various images than previous models. Furthermore, DualAttn-GAN outperforms

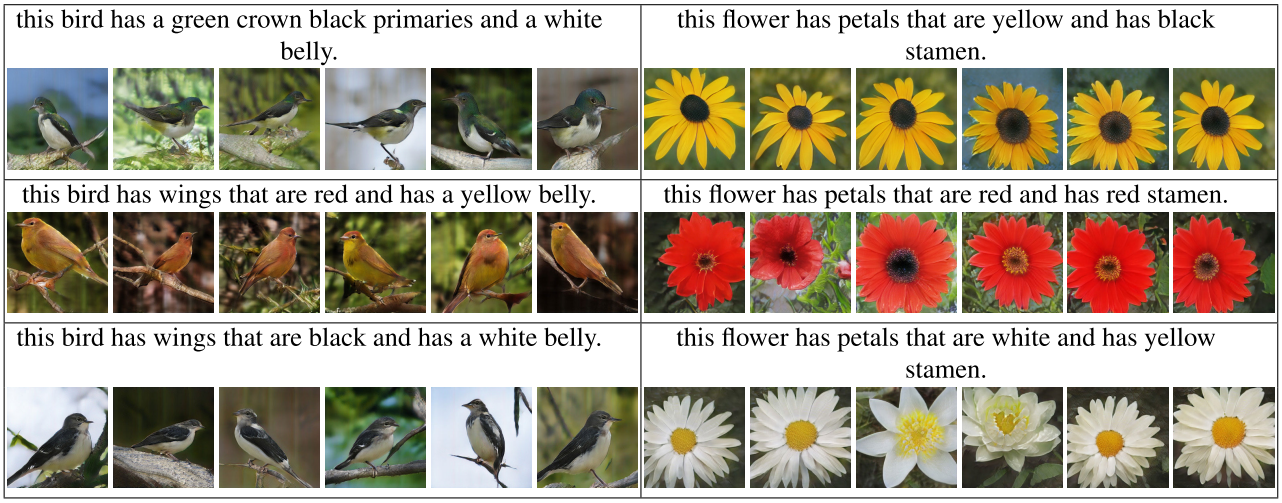


FIGURE 6. The images of each column are synthesized by DualAttn-GAN with the same noise vector. Each sentence is used to generate multiple images with different noise vectors.

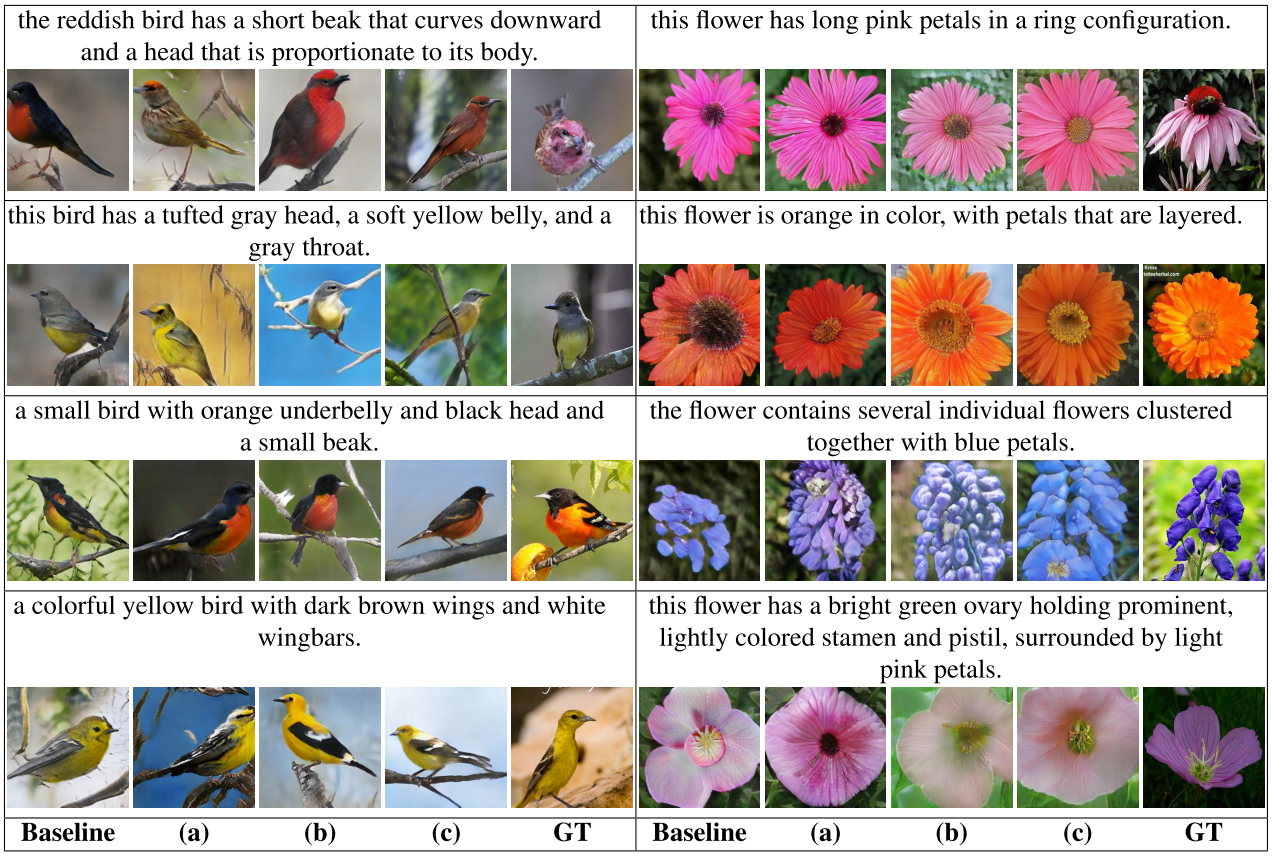


FIGURE 7. Comparisons of different components for text-to-image synthesis. (a): +IR, (b): +IR+TAM, (c): +IR+TAM+VAM.

other models as for the metric of FID, which achieves FID of 14.06 on CUB dataset and 40.31 on Oxford-102 dataset. The result indicates that synthetic samples based on our model are much closer to the real images. Reported in Table 2, DualAttn-GAN also achieves the best result on average Human Rank, which indicates that it is effective to enhance global structures and local detals.

Fig. 4 and Fig. 5 compare the qualitative results with other models on CUB and Oxford-102 datasets, by demonstrating more semantic details, natural color and complex structures. GAN-INT-CLS just synthesizes low-resolution ($64 * 64$) images, which lacks of many details. The samples generated by StackGAN and HDGAN can reflect general shape and color of the birds, but lack vivid objects and is not much

TABLE 3. Ablation study over IS and FID on CUB and Oxford-102 datasets.

Method	CUB		Oxford-102	
	IS ↑	FID ↓	IS ↑	FID ↓
Baseline	4.09 ± 0.05	24.73	3.64 ± 0.06	57.89
+ IR	4.33 ± 0.04	22.34	3.78 ± 0.06	51.92
+ IR + TAM	4.46 ± 0.06	16.32	3.85 ± 0.09	44.22
+ IR + TAM + VAM	$\mathbf{4.59 \pm 0.07}$	**14.06**	$\mathbf{4.06 \pm 0.05}$	**40.31**

TABLE 4. Component analysis of our DualAttn-GAN over Human Rank on CUB and Oxford-102 datasets.

Method	CUB		Oxford-102	
	GQR ↓	LQR ↓	GQR ↓	LQR ↓
Baseline	1.81	1.86	2.41	2.60
+ IR	1.70	1.78	2.30	2.52
+ IR + TAM	1.53	1.52	2.09	2.14
+ IR + TAM + VAM	**1.21**	**1.43**	**1.77**	**1.90**

consistent with the semantic information of the text decriptions. StackGAN++ improves in color based on StackGAN. AttnGAN obtains better scores, which is still slightly lower than ours. Although it can generate images with more details relevant to the captions, it lacks of the ability to capture global coherent structures, which makes images not realistic. The images generated by our model not only do outperform on structure, color and semantic details of the object (such as ''bright orange bellied and breast bird'' and ''white edge''), but also have fascinating background.

Comprehensive comparison, the image generated by our model is better than others. Further experiments prove that our model can accurately grasp the semantic information of the text and synthesize the related images, as well as the richness of the images synthesized by our model. Some generated images are shown in Fig. 6. For each column, our model can synthesize images corresponding to it with the same noise vector by changing several words in the sentence that contain important semantics. For each caption, rich images are generated according to different noise vectors.

D. ABLATION STUDY

To investigate the impact of each component in the proposed DualAttn-GAN, we gradually modify the baseline model and compare their difference. The results are reported in Table 3 and Table 4, and the generated images are shown in Fig. 7. IR donates the inverted residual structure. TAM represents the textual attention module, while VAM means the visual attention module.

First, we construct our baseline model by using spectral normalization, replacing GLU with ReLU based on StackGAN++ architecture, and the same configurations of the input text as AttnGAN. Compared with StackGAN++, our baseline model achieves great improvement in terms of IS but worse scores of FID and HR. As StackGAN++ has improved in color with extra color-consistency regularization, it makes sense for improving the realism of synthetic image. Then, by replacing the residuals with inverted residual structures, the scores of IS, FID and HR are improved. This indicates that inverted residual structures can improve the quality of the generated images by adding a few parameters. Next, we add the textual attention module, which boosts slight improvement of IS and GQR from the baseline model and has promising improvement in FID and LQR. As a residual learning network is applied to the textual attention module, the result is better than AttnGAN. It can enhances local details of images based on textual information and makes the synthesized image more natural and authentic. Finally, by further adding the visual attention module, it achieves 4.59 of IS, 14.06 of FID on CUB dataset and 4.06 of IS, 40.31 of FID on Oxford-102 dataset, and specially get great improvement in GQR. It demonstrates that the VAM plays an significant role in improving the structural quality of image generation.

VI. CONCLUSION

In this paper, we propose a Dual Attentional Generative Adversarial Network (DualAttn-GAN) for synthesizing image from text description, which adaptively integrates features with attention mechanisms. Specifically, we introduce a textual attention module and a visual attention module to enhance local and global attributive details by attending on related features from relevant words and different visual regions respectively. In addition, we present an inverted residual structure to boost representation power of CNNs. Moreover, useful techniques including spectral normalization and sample activation function are used to facilitate the training of the proposed model.

REFERENCES

[1] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Liwicki, and M. Z. Afzal, ''Tac-gan-text conditioned auxiliary classifier generative adversarial network,'' 2017, *arXiv:1703.06412*. [Online]. Available: https://arxiv.org/abs/1703.06412

[2] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, ''Plug & play generative networks: Conditional iterative generation of images in latent space,'' in *Proc. CVPR*, Honolulu, HI, USA Jul. 2017, A4477.

[3] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, ''Generative adversarial text to image synthesis,'' in *Proc. ICML*, New York, NY, USA, 2016.

[4] 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. CVPR*, Salt Lake City, Utah, USA, Jun. 2018, pp. 1316–1324.

[5] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, ''Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks,'' in *Proc. ICCV*, Venice, Italy, Oct. 2017, pp. A5907–5915.

[6] 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,'' in *Proc. TPAMI*, 2018.

[7] Z. Zhang, Y. Xie, and L. Yang, ''Photographic text-to-image synthesis with a hierarchically-nested adversarial network,'' in *Proc. CVPR*, Salt Lake City, Utah, USA, Jun. 2018, pp. 6199–6208.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, ''Generative adversarial nets,'' in *Proc. NIPS*, Montreal, QC, Canda, 2014, pp. 2672–2680.

[9] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
[11] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. ICML*, Sydney, VIC, Australia, 2017.
[12] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. ICLR*, Vancouver, Canada, 2018.
[13] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov, "Generating images from captions with attention," in *Proc. ICLR*, San Juan, Puerto Rico, 2016, pp. 1–12.
[14] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. ICLR*, 2014, pp. 1–14.
[15] S. Reed, A. van den Oord, N. Kalchbrenner, S. G. Colmenarejo, Z. Wang, Y. Chen, D. Belov, and N. de Freitas, "Parallel multiscale autoregressive density estimation," in *Proc. ICML*, Sydney, NSW, Australia, 2017, pp. 2912–2921.
[16] A. van den Oord, N. Kalchbrenner, L. Espeholt, K. Kavukcuoglu, O. Vinyals, and A. Graves, "Conditional image generation with pixelcnn decoders," in *Proc. NIPS*, Barcelona, Spain, 2016, pp. 4790–4798.
[17] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. ICML*, Sydney, NSW, Australia, 2017, pp. 214–223.
[18] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," in *Proc. ICLR*, New Orleans, LA, USA, 2019, pp. 1–35.
[19] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. NIPS*, Barcelona, Spain, 2016, pp. 2172–2180.
[20] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a laplacian pyramid of adversarial networks," in *Proc. NIPS*, Montreal, QC, Canda, 2015, pp. 1486–1494.
[21] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Proc. NIPS*, Long Beach, CA, USA, 2017, pp. 5767–5777.
[22] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: https://arxiv.org/abs/1411.1784
[23] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *Proc. ICML*, Sydney, NSW, Australia, 2017, pp. 2642–2651.
[24] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. ICLR*, San Juan, Puerto Rico, 2016, pp. 1–16.
[25] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2018, *arXiv:1805.08318*. [Online]. Available: https://arxiv.org/abs/1805.08318
[26] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, Honolulu, HI, USA, Jul. 2017, pp. 1125–1134.
[27] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, Venice, Italy, Oct. 2017, pp. 2223–2232.
[28] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. CVPR*, Honolulu, HI, USA, Jul. 2017, pp. 4681–4690.
[29] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proc. ECCVW*, Munich, Germany, 2018. p. 0.
[30] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *Proc. ICLR*, 2018, pp. 1–26.
[31] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. CVPR*, Long Beach, CA, USA, Jun. 2019, pp. 4401–4410.
[32] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Proc. NIPS*, Barcelona, Spain, 2016, pp. 217–225.
[33] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *Proc. NIPS*, Barcelona, Spain, 2016, pp. 3387–3395.
[34] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. NIPS*, Montreal, QC, Canda, 2014, pp. 2204–2212.
[35] M. F. Stollenga, J. Masci, F. Gomez, and J. Schmidhuber, "Deep networks with internal selective attention through feedback connections," in *Proc. NIPS*, Montreal, QC, Canda, 2014, pp. 3545–3553.
[36] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. ECCV*, Munich, Germany, Sep. 2018, pp. 3–19.
[37] K. Gregor, I. Danihelka, A. Graves, and D. Wierstra, "Draw: A recurrent neural network for image generation," in *Proc. ICML*, Lille, France, 2015, pp. 1462–1471.
[38] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, Lille, France, 2015, pp. 2048–2057.
[39] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proc. CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 4613–4621.
[40] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *Proc. ICML*, New York City, NY, USA, 2016, pp. 2397–2406.
[41] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 21–29.
[42] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. CVPR*, Salt Lake City, Utah, USA, Jun. 2018, pp. 4510–4520.
[43] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," Comput. Neural Syst., California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2010-001, 2011. [Online]. Available: https://resolver.caltech.edu/CaltechAUTHORS:20111026-120541847
[44] M. E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. ICCVGIP*, Dec. 2009, pp. 722–729.
[45] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Proc. NIPS*, Barcelona, Spain, 2016, pp. 2234–2242.
[46] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 2818–2826.
[47] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. NIPS*, Long Beach, CA, USA, 2017, pp. 6626–6637.

YALI CAI received the B.S. degree in computer science and technology from the Beijing University of Posts and Telecommunications, Beijing, China, in 2017, where she is currently pursuing the M.S. degree.

Her current research interests include deep learning, computer vision, and image generation.

XIAORU WANG received the M.S. and Ph.D. degrees in computer science and technology from the Beijing University of Posts and Telecommunications, in 2001 and 2015, respectively.

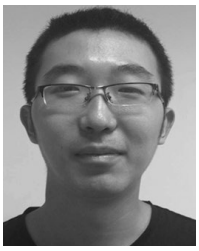
She is currently an Associate Professor and a Ph.D. Tutor with the School of Computer Science and Technology, Beijing University of Posts and Telecommunications, where she is also the Director of the Big Data Center. Her research interests include image processing and understanding, computer vision, and pattern recognition.

ZHIHONG YU received the B.S. and Ph.D. degrees in information engineering from the Beijing University of Posts and Telecommunications.

He is currently a Solution Architect with Intel Corporation. His research interests are on visual cloud, heterogeneous computing, and accelerators.

FU LI received the B.S. and M.S. degrees in physics from Sichuan University, China, in 1982 and 1985, respectively, and the Ph.D. degree in electrical engineering from the University of Rhode Island, in 1990.

Since 1990, he has been with Portland State University, where he is currently a Full Professor of electrical and computer engineering. His research interests include signal, image, and video processing, as well as wireless networks and multimedia communications.



PEIRONG XU was born in Shenyang, Liaoning, China, in 1996. He received the B.S. degree in information security from the Beijing University of Posts and Telecommunications, Beijing, China, in 2019, where he is currently pursuing the M.S. degree.

His research interests include semantic segmentation and image object extraction. He is also involved in research on image caption and has published an article named *Feedback LSTM Network Based on Attention for Image Description Generator.*

YUELI LI was born in Chongqing, China, in 1995. She received the B.S. degree in computer science and technology from the Beijing University of Posts and Telecommunications, Beijing, China, in 2017, where she is currently pursuing the M.S. degree.

Her research interests include multimedia data mining and image semantic annotation.

LIXIAN LI was born in Hegang, Heilongjiang, China, in 1993. She received the B.Sc. degree in computer science and technology from North Minzu University, Yinchuan, China, in 2016, and the master's degree in computer science from The University of Texas at Arlington, TX, USA, in 2019. She is currently pursuing the master's degree with the Beijing University of Posts and Telecommunications.

She has strong and pure enthusiasm in research, especially in computer vision and machine learning.