MS-GAN: Text to Image Synthesis with Attention-Modulated Generators and Similarity-aware Discriminators

Fengling Mao[1,2]
fengling.mao@vipl.ict.ac.cn
Bingpeng Ma[3]
bpma@ucas.ac.cn
Hong Chang[2,3]
changhong@ict.ac.cn
Shiguang Shan[2,3,4]
sgshan@ict.ac.cn
Xilin Chen[2,3]
xlchen@ict.ac.cn

[1] School of Information Science and Technology, ShanghaiTech University, Shanghai, China.

[2] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, China.

[3] University of Chinese Academy of Sciences, Beijing, China.

[4] CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China.

Abstract

Existing approaches for text-to-image synthesis often produce images that either contain artifacts or do not well match the text, when the input text description is complex. In this paper, we propose a novel model named MS-GAN, composed of multi-stage *attention-Modulated generators* and *Similarity-aware discriminators*, to address these problems. Our proposed generator consists of multiple convolutional blocks that are modulated by both globally and locally attended features calculated between the output image and the text. With such an attention-modulation, our generator can better preserve the semantic information of the text during the text-to-image transformation. Moreover, we propose a similarity-aware discriminator to explicitly constrain the semantic consistency between the text and the synthesized image. Experimental results on Caltech-UCSD Birds and MS-COCO datasets demonstrate that our model can generate images that look more realistic and better match the given text description, compared to the state-of-the-art models.

1 Introduction

Text-to-image synthesis is one of the most important and challenging tasks in computer vision and natural language processing. Given a text description, this task aims to synthesize an image with contents semantically consistent with the text. Recently, text-to-image synthesis has drawn increasing attentions and significant progress has been made thanks to the

Figure 1: Samples generated by StackGAN [24], HDGAN [26], AttnGAN [23] and our model on CUB dataset. Please zoom in to see the details.

development of generative adversarial networks (GAN) [1, 2, 4, 10, 21, 22]. Most existing methods [17, 23, 24, 26] are based on conditional GANs [10].

However, the performance of conventional text-to-image synthesis methods are not satisfactory enough in terms of the following aspects. First, the content of the synthesized image may not well match the semantics of the input text. Important semantic information may be ignored or misunderstood by the generator. Second, many synthesized images are still not visually appealing enough. There can be many artifacts in the images.

One possible reason that causes these problems is that the text semantic information is only input at the beginning of the deep generator [17, 26]. In this situation, the semantic information of the text can be weak or even disappear at the last several layers of the generator. As a consequence, the synthesized images may contain many artifacts and not well match the text description. As shown in Fig. 1, when the input description is complex, the images generated by state-of-the-art models contain many artifacts. Several key regions such as the color of the belly, the shape of the bill and the color of the wings do not match the text. Another possible reason is that most existing models [17, 23, 25] typically match the image with the text by training a discriminator. This discriminator can distinguish between a matched text-image pair (a text and its ground-truth image) and a wrong text-image pair (a text and an image that does not match the text). This implicit constraint, however, is not strong enough to enforce the semantic meaning of the output image to be similar with the input text. As a result, the synthesized images can be inconsistent with the given text in many details, as shown in Fig. 1.

To tackle these problems, we propose a novel model named MS-GAN, which consists of multi-stage attention-modulated generators (AMG) and similarity-aware discriminators (SAD). Specifically, the AMGs are introduced to better preserve semantic information of the text during the transformation from the text embedding to the output image. Each generator is composed of multiple convolutional blocks. The intermediate blocks are modulated by globally and locally attended features. The global features are calculated between the whole image and the sentence embedding, while the local features are calculated between the image patches and the word embeddings [23]. By modulating the features, the intermediate convolutional blocks in the generator can adaptively perceive the semantic meaning of the input text. Therefore, the semantic information of the text is better leveraged in the generator.

The similarity-aware discriminator is introduced to explicitly and adversarially improve the quality of the generated images and the consistency between the images and the given text descriptions. On one hand, we adopt both conditional and unconditional losses to encourage realistic results, as the existing models [23, 24] do. On the other hand, to guarantee that the generated image matches the text, motivated by [11], we propose an adversarial similarity

loss to explicitly enforce better semantic consistency between the image and the input text. With such a constraint, the synthesized image can be further refined to match the text.

To evaluate our model, we conduct experiments on the challenging Caltech-UCSD Birds [20] and MS-COCO [9] datasets. Experiment results demonstrate that with the proposed generator and discriminator, our model is able to generate images that are more realistic and more semantically consistent with the input text description, significantly outperforming the state-of-the-art models.

2 Related Work

Reed *et al*. [17] make one of the earliest attempts to synthesize images from a text with DCGAN [15] architecture, which encodes the text description using a hybird character-level convolutional-recurrent neural network [16]. They generate 64×64 images given the text descriptions. To generate images of higher-resolution, HDGAN [26] employs a generator which hierarchically synthesizes multi-scale images up to 256×256 resolution from the given text. Nam *et al*. [12] input an image and a condition text description to the generator, coupled with a local discriminator to undertake the fine-grained attribute discrimination. Zhang *et al*. [24] adopt a coarse-to-fine scheme to synthesize photo-realistic 256×256 images with a two-stage model. The first stage generates a coarse image, which is then refined by the second stage. Their subsequent work [25] also adopts the effective stacked structure for compelling images generation. These models constrain the similarity between the generated images and the given text, by optimizing a matching-aware loss. However, it only implicitly encourages the image to match the text, while our model constrains the semantic consistency between image and text in an explicit way with our adversarial similarity loss. Based on the stacked architecture of StackGAN++, Xu *et al*. [23] propose an attentional generative network to assist the generation with attention between word-level text features and local image features. Although benefited from the attention, during layers of mapping, the semantic information of the input text can be weak or misunderstood. On the contrary, our model modulates the intermediate convolutional blocks of the generator with globally and locally attended text features, significantly enhancing the encoding of semantic information in each layer, so that the generator can generate images with more precise contents.

3 Our Approach

3.1 Overall Architecture

As shown in Fig. 2, our proposed MS-GAN contains multi-stages of attention-modulated generators G_0, G_1 and G_2, and corresponding similarity-aware discriminators D_0, D_1 and D_2. Given an input text description, we first use a text encoder to extract its sentence-level feature t and word-level features w. Then, G_0 takes a random noise $z \sim \mathcal{N}(0,1)$ and t as inputs, and outputs a hidden feature map h_0 and a 64×64 image I_0. G_1 takes w and h_0 as inputs, and outputs a hidden feature map h_1 and a 128×128 image I_1. G_2 shares a very similar structure with G_1. It inputs w and h_1, and outputs a 256×256 image I_2.

The detailed structures of the discriminators are shown in Fig. 3(b). I_0, I_1 and I_2 together with their corresponding sentence feature t are input to D_0, D_1 and D_2, respectively.

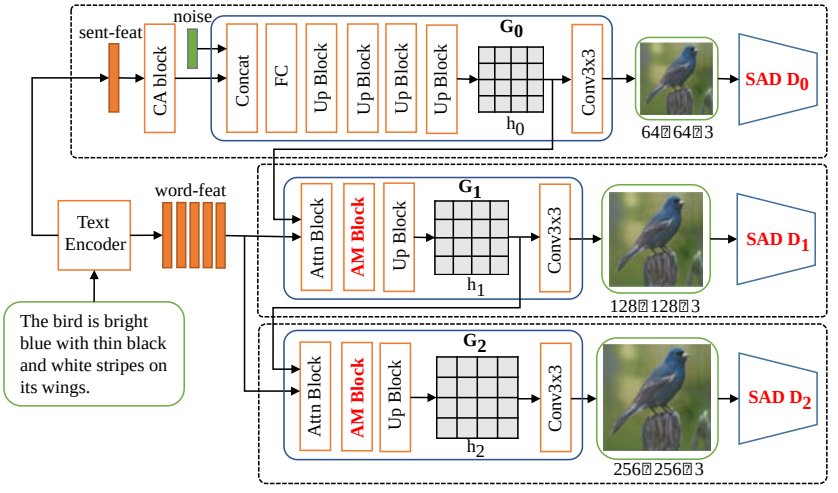


Figure 2: The overall architecture of our model. It contains the proposed attention-modulated generator (G_0, G_1 and G_2) and similarity-aware discriminator (SAD D_0, D_1 and $D2$). In the generator, some intermediate blocks (AM Block) are modulated by globally and locally attended features.

3.2 Attention-Modulated Generator

3.2.1 Feature Modulation

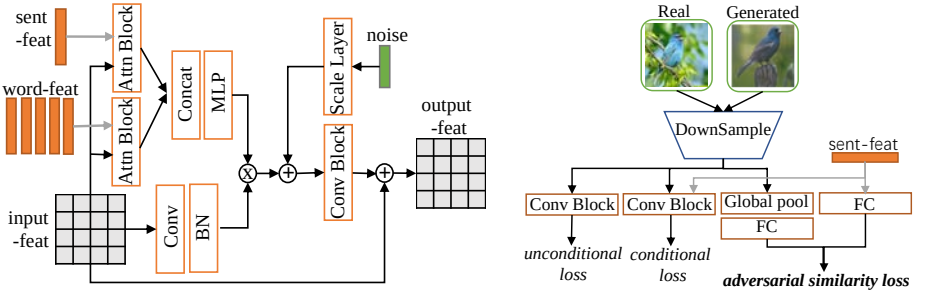
Previous feature modulation methods [3, 14] typically use conditional batch normalization (CBN) [3] to modulate convolutional feature maps using a text embedding, by predicting the affine parameters of batch norm layer with the text. Specifically, when performing Batch Normalization, the convolutional feature map F is first normalized with mean $\mu(F)$ and variance $\sigma(F)$ in a batch. Then, the normalized feature map is re-parameterized with two parameters γ and β, which are learned on condition of a given text embedding t. The CBN modulation process can be formulated as follows:

$$CBN(F_{i,c,h,w}\, t) = f_\gamma(t)\frac{F_{i,c,h,w} - \mu(F)}{\sqrt{\sigma(F)+\varepsilon}} + f_\beta(t), \tag{1}$$

where i,c,h,w are the sample index, number of channels, height and width of the feature map, respectively. $f_\gamma(t)$ and $f_\beta(t)$ refer to the learned parameters γ and β conditioned on t. In a typical conditional BN, the mapping from the text t to $f_\gamma(t)$ and $f_\beta(t)$ is usually realized with a multi-layer perceptron (MLP). With such a conditional BN, the original convolutional feature maps can be regularized or modulated by the text embedding.

3.2.2 Attention-Modulated Generator

Conventional generators usually transform a text embedding to an image with several layers of vanilla convolutional blocks or Resnet blocks [5]. However, during the text-to-image transformation, the semantic text information may be weakened or missing in the last several layers of the deep generator. Note that feature modulation allows the feature maps to be regularized with extra conditions. Motivated by this method, we propose an attention-modulated



(a) Attention-Modulation Block (b) Similarity-aware Discriminator

Figure 3: The structure of our proposed attention-modulation block (a) and similarity-aware discriminator (b).

generator to address this problem. Specifically, we design an attention-modulation block to replace the intermediate convolutional blocks of the generator. Feature maps in our attention-modulation blocks are modulated with globally and locally attended features. These features are calculated by using the text embedding to attend to the image regions.

The detailed structure of the attention-modulation block is shown in Fig. 3(a). It takes a feature map h_{in}, the sentence feature t and the word feature w as inputs, and outputs the modulated feature map $h_{modulated}$. First, the attended sentence feature $\bar{t}$ is calculated with an attention model [23] that uses the sentence feature t to attend the whole output image. Then, the attended word feature $\bar{w}$ is calculated with an attention model [23] that uses the word feature w to attend the local regions of the image. The input feature map h_{in} is first mapped with a convolutional layer, then modulated with a Batch Normalization layer that is conditioned on the attended features $\bar{t}$ and $\bar{w}$. Specifically, we concatenate the attended features into one conditional feature $c = [\bar{w}, \bar{t}]$. Then, we input c to a MLP to obtain f_γ, to predict the parameter γ in batch normalization. As explained in StyleGAN [7], noise can influence the local details of the generated image and can enhance the diversity of the images. Therefore, we also input random noise $\bar{z}$ to predict the parameter β in batch normalization, with a linear layer f_β instead of a MLP. Finally, the modulated feature is passed to a non-linear activation function to obtain the final output of our attention-modulation block. We formulate our attention-modulation block as follows:

$$h_{modulated} = Act(f_\gamma(c)\frac{\phi(h_{in}) - \mu}{\overline{\sigma + \varepsilon}} + f_\beta(\bar{z})), \tag{2}$$

where $\phi(\cdot)$ is a convolutional layer, and $Act(\cdot)$ is a non-linear activation function. We replace all the Resblocks [5] (composed of conv, batch norm, and non-linear layers) in the generator with our attention-modulation block. Compared to the vanilla Conv blocks or Resblocks, our attention-modulation blocks can perceive the global and local semantic features at different resolutions, significantly benefiting the generator to preserve the text information.

3.3 Similarity-aware Discriminator

Our proposed similarity-aware discriminator can not only distinguish real images from fake images, but also discriminate whether the image matches the given text. The network structure of our discriminator is shown in Fig. 3(b). In the i-th stage, the discriminator D_i takes

three types of input: (I_i^r,t), (I_i^w,t) and (I_i^g,t), where I_i^r, I_i^w and I_i^g denote the real image, wrong image and the image generated by G_i respectively. D_i is constrained with three major objectives: an unconditional loss that judges the realness of the input image [25], a conditional loss that judges the realness of the joint image-text pair [25], and our proposed adversarial similarity loss that evaluates the semantic similarity between the image and the text. The unconditional losses for training G_i and D_i are as follows:

$$\begin{aligned} L_{D_i}^{UC} &= -\mathbb{E}_{I_i^r\sim p_{data_i}}\left[\log D_i\left(I_i^r\right)\right] - \mathbb{E}_{I_i^g\sim p_{G_i}}\left[\log\left(1-D_i\left(I_i^g\right)\right)\right] \\ L_{G_i}^{UC} &= -\mathbb{E}_{I_i^g\sim p_{G_i}}\left[\log D_i\left(I_i^g\right)\right], \end{aligned} \tag{3}$$

where p_{data_i} is the distribution of real images at the i-th scale and p_{G_i} is the distribution of the generated images. Similarly, the conditional losses are as follows:

$$\begin{aligned} L_{D_i}^{C} &= -\mathbb{E}_{I_i^r\sim p_{data_i}}\left[\log D_i\left(I_i^r,t\right)\right] - \mathbb{E}_{I_i^g\sim p_{G_i}}\left[\log\left(1-D_i\left(I_i^g,t\right)\right)\right], \\ L_{G_i}^{C} &= -\mathbb{E}_{I_i^g\sim p_{G_i}}\left[\log D_i\left(I_i^g,t\right)\right]. \end{aligned} \tag{4}$$

The conditional and unconditional losses can encourage the model to produce realistic results. However, they do not explicitly constrain the generated image to match the given text (see results in Fig. 1). To address this problem, we propose an adversarial similarity loss that provides a strong constraint on the text-image pairs for better semantic similarity.

Adversarial Similarity Loss. Given a text-image pair, we first extract the image feature f_{im} from the last convolutional layer (after global average pooling). The similarity distance between the sentence embedding t and the image feature vector f_{im} is defined as:

$$d\left(t,f_{im}\right) = \left(\left(W_t t\right)^T\left(W_{im}f_{im}\right)\right) / \left(\ W_t t\ \cdot\ W_{im}f_{im}\ \right), \tag{5}$$

where W_t and W_{im} are projection matrices that map the text embedding t and the image feature vector f_{im} into a common feature space. For the i-th stage, our proposed adversarial similarity loss can be computed as follows:

$$\begin{aligned} L_{D_i}^{ASL} &= \mathbb{E}_{t\sim p_{data}}[\max\left(0,1-d\left(t,f_{im}^r\right)\right) + \left(\max\left(0,1+d\left(t,f_{im}^w\right)\right) + \max\left(0,1+d\left(t,f_{im}^g\right)\right)\right)/2], \\ L_{G_i}^{ASL} &= -\mathbb{E}_{t\sim p_{data}}\left[d\left(t,f_{im}^g\right)\right], \end{aligned} \tag{6}$$

where f_{im}^r, f_{im}^w, f_{im}^g denote the image feature vector of I^r, I^w, I^g respectively. $L_{D_i}^{ASL}$ denotes the adversarial similarity loss for optimizing D_i and $L_{G_i}^{ASL}$ denotes that for G_i.

Intuitively, $L_{D_i}^{ASL}$ is optimized to minimize the semantic distance between the given text and the ground-truth image, and maximize the distance between the text and a wrong image (randomly sampled) as well as the generated image. In this way, the discriminator can judge the semantic similarity between the text and the synthesized image. On the contrary, $L_{G_i}^{ASL}$ is optimized to minimize the distance between given texts and generated images. This adversarial learning process forces the generator to synthesize images that are more semantically consistent with the given text descriptions.

Full loss of MS-GAN. Our MS-GAN model combines the unconditional loss, conditional loss and our proposed adversarial similarity loss. Finally, the full loss of the discriminator L_D and the full loss of the generator L_G can be formulated as follows:

$$\begin{aligned} L_D &= \sum_{i=0}^{2}\left(L_{D_i}^{UC} + L_{D_i}^{C} + \lambda L_{D_i}^{ASL}\right), \\ L_G &= \sum_{i=0}^{2}\left(L_{G_i}^{UC} + L_{G_i}^{C} + \lambda L_{G_i}^{ASL}\right) + \lambda_1 L_{DAMSM} + L_{CA}, \end{aligned} \tag{7}$$

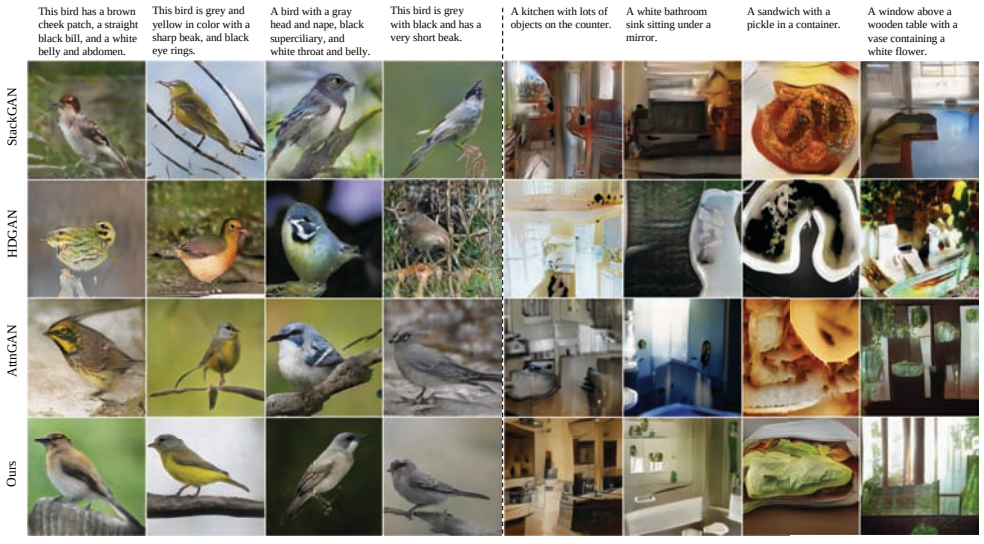


Figure 4: The visualization of generated images. The first row is the given text descriptions, and the second row to the fifth row are images generated by GAN_CLS_INT [17], HDGAN [26], AttnGAN [23] and our model, respectively. The first four columns are results on CUB dataset and the last four columns are results on MS-COCO dataset.

where L_{DAMSM} denotes the DAMSM loss [23] for the attention model, and L_{CA} denotes the KL loss for the conditional augmentation block [24]. By optimizing L_D in Eq. 7, the proposed similarity-aware discriminator can discriminate both the realness of an image and the similarity between the image and its corresponding text description.

4 Experiments

4.1 Experiment Settings

Datasets. We evaluate our model on the Caltech-UCSD Birds 200 (CUB) dataset [20] and MS-COCO [9] dataset. The CUB dataset contains 200 categories of birds with 11,788 images. Each image is annotated with 10 captions. Following the previous work [17, 23, 24, 26], we utilize 150 different categories for training, and employ the remaining 50 categories for testing. The MS-COCO dataset contains 80k images for training and 40k images for testing, with 5 captions per image.

Evaluation Metrics. We adopt the widely used Inception Score (IS) [19] and Fréchet Inception Distance (FID) [6] to evaluate the quality of generated images. To precisely evaluate whether the generated images match the given text description, we conduct a User Study. Specifically, we randomly select 50 text descriptions from the CUB test dataset. For each text, we generate only one image per method. Then, we ask 18 users to select the image that most matches the text. Finally, we collect 900 votes and use the percentage of votes for each method as the Perceptual Similarity Score (**PSS**).

Implementation Details. We adopt AttnGAN [23] as the baseline model since it has achieved state-of-the-art performance. Our multi-stage model can generate images with resolutions of

Method	CUB		MS COCO	
	IS	FID	IS	FID
GAN_CLS_INT [17]	2.88 ± .04	68.79	7.88 ± .07	60.62
GANWN [18]	3.62 ± .07	67.22	-	-
StackGAN [24]	3.70 ± .04	51.89	8.45 ± .03	74.05
StackGAN++ [25]	4.04 ± .05	15.30	8.30 ± .10	81.59
PPGN [13]	-	-	9.58 ± .21	-
HDGAN [26]	4.15 ± .05	18.23	11.86 ± .18	75.34
AttnGAN [23]	4.36 ± .03	10.65	23.88 ± .27	29.43
Ours	**4.56 ± .05**	**10.41**	**25.98 ± .04**	**29.29**

Table 1: Comparative results on CUB and MS-COCO datasets.

Method	PSS
StackGAN[24]	15.11%
HDGAN[26]	9.00%
AttnGAN[23]	29.78%
Ours	**46.11%**

(a) User Study.

Method	IS
Baseline	4.36 ± .03
λ=0.01	4.43 ± .05
λ=0.1	4.52 ± .05
λ=1	4.49 ± .04

(b) Different λ.

Method	IS
Baseline	4.36 ± .03
Baseline+SAD	4.52 ± .05
Baseline+AMG	4.48 ± .07
Baseline+SAD+AMG	4.56 ± .05

(c) Effectiveness of components.

Table 2: (a) The perceptual similarity score (PSS) of each model evaluated by users. (b) The results under different weight λ of the adversarial similarity loss. (c) The results of our model with different components.

64×64, 128×128 and 256×256. We optimize our model using ADAM [8] optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate for the generators and the discriminators is set to 0.0002. The weight λ of our proposed adversarial similarity loss is set to 0.1 as default in all experiments on both CUB and MS-COCO datasets. The weights λ_1 is set to be 5 for CUB dataset, and 50 for MS-COCO dataset. The maximal number of training epochs is set to be 700 for CUB dataset, and 140 for MS-COCO dataset.

4.2 Quantitative Results

We compare our model with state-of-the-art models, GAN_CLS_INT [17], GANWN [18], StackGAN [24], StackGAN++ [25], PPGN [13], HDGAN [26] and AttnGAN [23]. The results on CUB dataset and MS-COCO dataset are shown in Table 1. Note that all the scores of our model are computed using the output 256×256 images for the two datasets. The Inception Scores of most conventional models are given by their own papers. The Inception Score of AttnGAN is calculated using the official evaluation code and the images generated by the official pre-trained models provided by the authors. The FID scores of GAN_CLS_INT, GANWN, StackGAN, StackGAN++ are provided by [25]. The FID scores of HDGAN, AttnGAN and our model are computed by adopting the same strategy with StackGAN++ [25] for equal comparison. Specifically, we compute the FID score by randomly choosing 30,000 256×256 images generated by each model.

As shown in Table 1, on the CUB dataset, our model achieves the highest Inception Score (higher is better) and the lowest FID score (lower is better), which indicates better quality and diversity of our generated images. Compared with AttnGAN, we improve the Inception Score from 4.36 to 4.56, and the FID from 10.65 to 10.41. On the MS-COCO

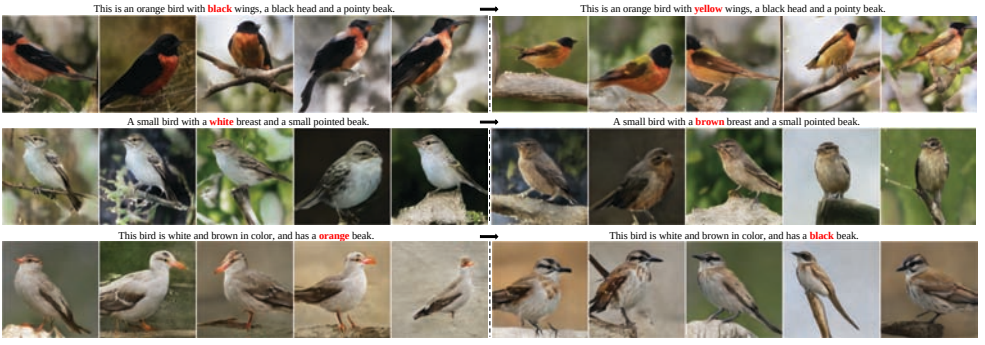


Figure 5: The visual results of text modification. The left part of the dashed line shows the result given the original text, while the right part shows the results given the modified text.

dataset, we improve the Inception Score from 23.88 to 25.55, and the FID from 29.43 to 27.10. The significant improvements demonstrate the superiority of our model on generating high-quality images.

User Study. The goal of text-to-image synthesis is not only to generate high-quality images, but also to create images well matching the given text descriptions. However, to the best of our knowledge, there is still no optimal solution for evaluating the degree of matching between the image and the text. Therefore, we conduct User Study to precisely evaluate the semantic similarity between the generated images and the text descriptions. We compare the perceptual similarity score with StackGAN [24], HDGAN [26] and AttnGAN [23] on CUB dataset. As shown in Table 2(a), our model gains much higher score compared with all the other models, showing better ability to generate well-matched images.

4.3 Qualitative Results

To fully evaluate our performance, we visualize the results of our model, StackGAN [24], HDGAN [26] and AttnGAN [23] in Fig. 4. When the input text describes complex scenes, the existing methods tend to fail and generate obvious artifacts. Moreover, the contents of images generated by existing models do not well match the text description. For example, in the second column of Fig. 4, the color of the bird is supposed to be grey, while bird generated by HDGAN is deep yellow. In the third column of Fig. 4, the head of the bird should be gray. However, the bird generated by AttnGAN is dark blue. On the contrary, even under these challenging situations, our model can still generate high-quality images that are semantically consistent with the input text.

Furthermore, we evaluate the generalization capacity of our model. Specifically, we first randomly select text descriptions from the CUB test dataset. Then, we change several important words in each text and obtain the new text. We utilize the original text descriptions and man-made new text descriptions to generate images using our model. As shown in Fig. 5, our model are sensitive to the changes of text and can synthesize images matching the unseen text descriptions.

Figure 6: Visualization results of different components with the same input text description. For the first and the last four columns, from left to right are: the input text, generated image of the baseline model and the generated image of the baseline with our SAD/AMG component.

4.4 Ablation Study

To explore the influence of the weight λ of our adversarial similarity loss, we conduct experiments on CUB dataset only using our proposed similarity-aware discriminator with the hyper-parameter λ set to be 1, 0.1 and 0.01. Results are shown in Table 2(b). The performance of our model varies with the weight of our adversarial similarity loss. However, under all λ settings, our model with adversarial similarity loss performs much better than the model without the loss.

To evaluate the contributions of each component, we compare different variants of our model. We visualize the images generated by different components of our method. Some examples are shown in Fig 6. Compared with the baseline model, SAD tends to keep global semantic consistence and image quality and AMG fine-tunes local semantic and quality details. both our SAD and AMG components significantly improve the visual quality. Moreover, Table 2(c) shows the quantitative results of four versions of our model on CUB dataset, *i. e.*, the baseline, baseline with only SAD (meaning that we modify the discriminator of the baseline model to our similarity-aware discriminator), baseline with only AMG (meaning that we modify the generator of the baseline model to our attention-modulated generator), and our full model (baseline + SAD + AMG). Results show that our proposed attention-modulation generator and similarity-aware discriminator both bring significant improvements, and our MS-GAN combining both components achieves the best performance.

5 Conclusion

In this paper, we propose a novel model named MS-GAN, which consists of attention-modulated generators and similarity-aware discriminators. The attention-modulated generator contains several attention-modulation blocks, which adaptively modulate the feature maps of the generator with attended global sentence feature and local word feature. The similarity-aware discriminator can evaluate the realness of the images as well as judge whether the generated image matches the semantics of the given text description. Our model significantly improves the quality of the generated images and the semantic similarity between the images and the given text descriptions, demonstrating its effectiveness.

6 Acknowledgement

This work is partially supported by National Key R&D Program of China (No.2017YFA0700800), Natural Science Foundation of China (NSFC): 61876171 and 61572465, and Beijing Municipal Science and Technology Program: Z181100003918012.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *International Conference on Learning Representations*, 2019.

[2] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

[3] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2017.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018.

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2015.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. European Conference on Computer Vision*, pages 740–755, 2014.

[10] Mehdi Mirza and Simon Osindero. Conditional generative adversarial networks. *Manuscript: https://arxiv. org/abs/1709.02023*, 9:24, 2014.

[11] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *International Conference on Learning Representations*, 2018.

[12] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: Manipulating images with natural language. In *Advances in Neural Information Processing Systems*, pages 42–51, 2018.

[13] Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4467–4477, 2017.

[14] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI Conference on Arti cial Intelligence*, 2018.

[15] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[16] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–58, 2016.

[17] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. 2016.

[18] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *Advances in Neural Information Processing Systems*, pages 217–225, 2016.

[19] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[20] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[21] Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2018.

[22] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 5840–5848, 2019.

[23] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2018.

[24] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proc. IEEE International Conference on Computer Vision*, pages 5907–5915, 2017.

[25] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[26] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 6199–6208, 2018.