Hybrid Attention Driven Text-to-Image Synthesis via Generative Adversarial Networks

Qingrong Cheng and Xiaodong Gu(✉)

Department of Electronic Engineering, Fudan University,
Shanghai 200433, China
xdgu@fudan.edu.cn

Abstract. With the development of generative models, image synthesis conditioned on the specific variable becomes an important research theme gradually. This paper presents a novel spectral normalization based *Hybrid Attentional Generative Adversarial Networks* (HAGAN) for text to image synthesis. The hybrid attentional mechanism is composed of text-image cross-modal attention and self-attention of image sub regions. Cross-modal attention mechanism contributes to synthesize more fine-grained and text-related image by introducing word-level semantic information in generative model. The self-attention solves the long distance reliance of image local-region features when generate image. With spectral normalization, the training of GANs become more stable than traditional GANs, which conduces to avoid model collapse and gradient vanishing or explosion. We conduct experiments on widely used Oxford-102 flower dataset and CUB bird dataset to validate our proposed method. During quantitative and non-quantitative experimental comparison, the results indicate that the proposed method achieves the best performance on Inception score (IS), Fréchet Inception Distance (FID) and visual effect.

Keywords: Text to image synthesis · Spectral normalization · Self-attention · Cross-modal attention · Generative Adversarial Networks

1 Introduction

Recent years have witnessed the great progress of Deep Neural Networks (DNNs), especially various kinds of generative tasks and discriminative tasks. Particularly, Convolutional Neural Networks (CNNs) have shown excellent performance on the challenging multi-category classification [1]. Besides, another branch of research focus is generative task, which is inverse mapping of discriminative task. In particular, generative tasks based on Generative Adversarial Networks (GANs) have achieved promising results [2] in image synthesis. Recently, photo-realistic image synthesis gradually becomes an important research direction with many potential applications, such as, computer graphics and photo retouching. To be specific, methods for text-to-image synthesis need generate image that are highly similar to meanings embedded in texts. However, image synthesis, conditioned on the given text descriptions, is also a knotty problem because of the great gap between text modality and image modality.

I. V. Tetko et al. (Eds.): ICANN 2019, LNCS 11731, pp. 483–495, 2019.
https://doi.org/10.1007/978-3-030-30493-5_47

Almost all existing text-to-image synthesis methods are based on GANs and some of them achieve remarkable performance. Generative Adversarial Networks (GANs) is proposed by Goodfellow in 2014 [3], which has made impressive performance in generative tasks. It is composed of two sub-networks, generator and discriminator, trained with a competing goal in an adversarial manner. From them on, GANs related work become a focused research direction. Meanwhile, adversarial learning mechanisms have shown great progress in many complex simulating problems [4].

Although excellent performance in many tasks, GANs are well known for difficulty in training and mode collapse. Many research works indicate that the instability in training is due to the disjoint of the generated data distribution and the real data distribution [5]. Besides, the mode collapse in GANs shows that the model will synthesize similar samples with uniform color and single texture. For addressing the knotty problem, many methods were proposed until now, such as WGAN [6], WGAN-GP [7] and SNGAN [8]. Some of those methods achieve excellent performance in stabilizing the training process and avoiding mode collapse.

Text-to-image synthesis is more challengeable than simply generate image from random noise or category condition. Text description contains more abundant and detailed image features, which should be drawn in synthesized image. Aiming at synthesizing photo-realistic image, there are two main branches of methods, VAE-based methods [9–11] and GAN-based methods [12–16]. Cai et al. [9] propose an image synthesis framework for fine-grained image in a multi-stage variational auto-encoder manner. Gulrajani et al. [10] present an improved PixelCNN-based model named PixelVAE, which introduces an autoregressive decoder for natural image synthesis. Deep Recurrent Attentive Writer (DRAW) [11] combines spatial attention mechanism with sequential VAE framework for constructing complex images.

Apart from VAE-based methods, GAN-based approaches also show great effectiveness in text to image synthesis. Specifically, Reed et al. [12] firstly introduce the traditional GAN into text to image synthesis in 2016. Following on the previous work, they propose a Generative Adversarial What-Where Network (GAWWN) [17] by using position box as additional supervision, which achieves better performance. However, the images synthesized by the first model are blurry and unclear. Inspired by the drawing step of human beings, multi-stage strategy is introduced into image synthesis in recent years, such as StackGAN [13, 14], AttnGAN [16] and CWPGGAN [15]. To be specific, StackGAN has two versions, StackGAN-v1 [13] and StackGAN-v2 [14]. StackGAN-v1 is based on two-stage GANs, while the StackGAN-v2 is an advanced three-sage model. Therefore, the images synthesized by the second model are more realistic and richly-textured than the first method. Progressive growing mechanism [18] is adopted in CWPGGAN [15], which can gradually improve the resolution and quality.

Attention mechanism shows effectiveness in many applications, especially in natural language process and computer vision. More specifically, self-attention mechanism is introduced in image generation [19]. Besides, attention mechanism is also adopted in text to image generative task, such as alignDRAW [20] and AttnGAN [16]. The alignDRAW [20] based on the mentioned DRAW introduces soft attention mechanism for attending to the relevant words of image feature. Xu et al. [16] propose a multi-stage Attentional Generative Adversarial Network (AttnGAN) for fine-grained image synthesis from text. Their methods not only use generator to generate high-resolution realistic image but also add word-level feature into generator, while others' methods only adopt sentence feature.

Inspired by previous work, we propose a spectral normalization based *Hybrid Attentional Generative Adversarial Networks* (HAGAN) that combines the image self-attention and text-image cross-modal attention mechanism for fine-grained image synthesis in this paper. Firstly, the features are extracted by the pretrained model name DAMSM [16], which contains both text and image feature embedding. Then, we feed the encoded text feature into three-stage hybrid attentional generative adversarial networks for image synthesis. The self-attention mechanism is introduced in the first-stage network and cross-modal attention is adopted in second and third stage generators. We mainly use the publicly available Oxford-102 flowers dataset and the Caltech CUB-200 birds dataset to conduct the experimental analysis. During the evaluation metric and side-by-side comparison with the state-of-the-art methods, the results indicate that our proposed method can get better visual effect and competitive evaluation value.

Compared to existing works, the main contributions of our work are as follows.

(1) By developing a hybrid attention mechanism for text to image synthesis, self-attention of image generation can solve long distance reliance between local features and cross-modal attention can add word-level features in generator for fine-grained image details.

(2) Due to spectral normalization, the training of the model becomes more stable than traditional GANs. Therefore, the generator can synthesize more realistic image due to discriminator satisfied with K-Lipschitz constraint can provide useful and effective gradient information for model optimizing.

The rest of this paper is organized as follows. The second section presents our proposed HAGAN approach. The third section shows the experimental results and comparison, and the last section concludes this paper.

2 The Proposed Method

2.1 Background

A. Generative Adversarial Networks

The GANs consists of two sub-networks, a discriminator D and a generator G, that cooperate and compete in a minimax game until the game achieves zero-sum game. Such minimax game can be described as the following object function $V(G, D)$.

$$\begin{aligned}\min_G \max_D V(G, D) = & E_{x \sim p_{data}(x)}[\log(D(x))] \\ & + E_{x \sim p_z(z)}[\log(1 - D(G(z)))],\end{aligned} \tag{1}$$

where x is the real image and z is random noise. In the training process, the discriminator tries to maximize V, however, the generator wants to minimize the object function. In the last, the game of the two networks achieves the Nash Equilibrium that both can obtain the best performance.

B. Conditional Generative Adversarial Networks

Conditional GANs add conditional variable y to control the features of output image. The object function of conditional GAN can be described as follows.

$$\begin{aligned}\min_G \max_D V(G,D) = & \mathrm{E}_{x \sim p_{data}(x)}[\log(D(x|y))] \\ & + \mathrm{E}_{x \sim p_z(z)}[\log(1 - D(G(z|y)))],\end{aligned} \tag{2}$$

where y is a conditional variable. The function of generator $G(z|y)$ allows the generator G to generate images conditioned on the given conditioning variable. The discriminator $D(x|y)$ evaluates whether the generated image is matched with conditioning variable y or not.

2.2 The Framework of Hybrid Attentional Generative Adversarial Networks

The HAGAN enables the generator to draw different sub-regions conditioned on related words and other long distance related image sub-regions. Meanwhile, the spectral normalization stabilize the training of the discriminator, which will contribute the optimization of the generator. The framework of the HAGAN, as shown in Fig. 1.

A. Hybrid Attentional Generative Adversarial Networks

Suppose the texts and images are stored in a N-pair document corpus (X^T, X^I). Here, X^T is text data and X^I is image data. The text feature and image feature are extracted by the well-trained embedding model DAMSM [16], which is based on the bi-directional Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN).

$$(\bar{\varphi}, \boldsymbol{\varphi}; \bar{\phi}, \boldsymbol{\Phi}) = F_{DAMSM}(X^T, X^I), \tag{3}$$

where $\boldsymbol{\varphi}$ indicates word feature matrix, $\bar{\varphi}$ denotes sentence feature, $\bar{\phi}$ is global image feature and $\boldsymbol{\Phi}$ presents the sub-region feature matrix.

The encoded sentence feature $\bar{\varphi}$ will be pretreated before input into the multi-stage generative networks. As following,

$$\tilde{\varphi} = F_{cat}(z, F^{ca}(\bar{\varphi})), \tag{4}$$

where z is random noise vector, F^{ca} denotes the Conditioning Augmentation [14] which converts the sentence feature vector $\bar{\varphi}$ to the conditioning vector, and F_{cat} is concatenate function. After several upsample operation, the hidden feature gradually denotes the image features. The self-attention mechanism acts on the hidden feature maps $(\hat{h}_0, \hat{h}_1)$. As following,

$$\hat{h}_i = \hat{F}_i(h_{i-1}, F_i^{self\_attn}(\hat{h}_{i-1})), \text{ where } i = 1, 2. \tag{5}$$

Here, $F_i^{self\_attn}$ is the self-attention mechanism. The first-stage generator synthesizes image conditioned on the output of self-attention block directly. The generative networks consists of three generators (G_0, G_1, G_2), which use the previous hidden feature (h_0, h_1, h_2) to generate different-scale images $(\hat{x}_0, \hat{x}_1, \hat{x}_2)$. To be specific, the process of multi-stage generator is defined as following.

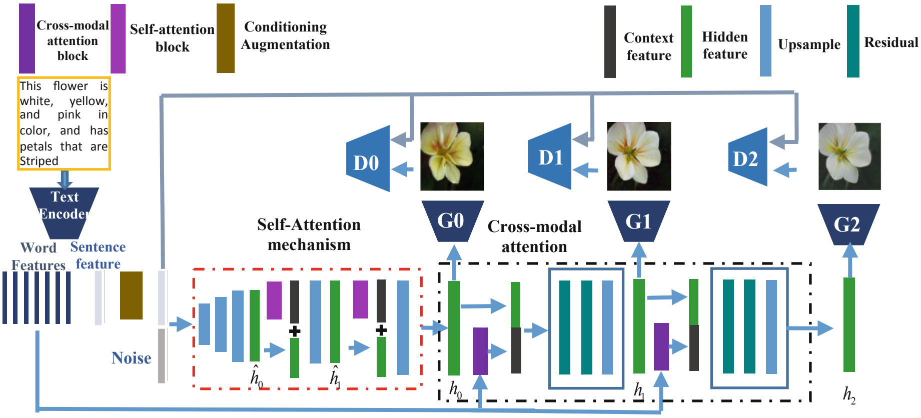

Fig. 1. The overall pipeline of the hybrid attentional generative adversarial networks.

$$\hat{x}_i = G_i(h_i), \text{ where } i = 1, 2. \tag{6}$$

The cross-modal mechanism is introduced in the second and the third networks, which can add more detailed attribute informations in the feature matrix. Specifically, the operation of cross-modal attention is defined as following.

$$h_i = F_i(h_{i-1}, F_i^{cro\_attn}(\boldsymbol{\varphi}, h_{i-1})), \text{ where } i = 1, 2. \tag{7}$$

Here, $\boldsymbol{\varphi}$ is the feature matrix of word features, and $F_i^{cro\_attn}$ is the cross-attention model of the i-th stage generator. All of these functions are modeled as neural networks.

(1) *Self-Attention mechanism for the first stage generator*

The generator G and discriminator D of GAN models usually consist of convolutional neural networks. However, the convolutional filter only process the information in a local neighborhood, such as window size 3×3. Hence, long-range dependencies cannot be considered in the convolutional process. By introducing the self-attention mechanism into the GANs model, the generator can use the long-distance relationships between widely separated sub-regions.

In the deep model, the feature map $\hat{h} \in \mathbb{R}^{C \times N}$ of previous layer presents the hidden features of an image. We use two 1×1 convolutional layer to convert the feature map into two space $\bar{H}, \hat{H}$, and then calculate the attention of the two feature maps.

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N}\exp(s_{ij})}, \tag{8}$$

where $s_{ij} = \bar{H}(x_i)^T\hat{H}(x_j)$, and $\beta_{j,i}$ indicates how much attention from the i-th location when generating the j-th region. The attention map is obtained by weighted sum of all the output, as following.

$$\hat{C}_i = (\hat{c}_1, \cdots, \hat{c}_j, \cdots, \hat{c}_N) \in \mathbb{R}^{C\times N}, \tag{9}$$

where,

$$\hat{c}_j = \sum_{i=1}^{N}\beta_{j,i}h(x_i), h(x_i) = \mathbf{W}_h\hat{h}(x_i). \tag{10}$$

Then we apply a weight scale parameter γ on attention map. The final weighted output is given by,

$$\hat{h}_i = \gamma\hat{C}_i + \hat{h}_{i-1}, \tag{11}$$

where γ is initialized as 0.
In short, the self-attention mechanism can be denoted as

$$\hat{h}_i = \hat{F}_i(h_{i-1}, F_i^{self\_attn}(\boldsymbol{\varphi}, \hat{h}_{i-1})). \tag{12}$$

(2) *Cross-modal Attention mechanism for the second and third stage generators*
Cross-modal attention mechanism is adopted to add relevant word-level information to networks for producing fine-grained image. The input of the cross-modal attention mechanism is the previous hidden feature $h \in \mathbb{R}^{\hat{D}\times N}$ of image and the word-level features $\boldsymbol{\varphi} \in \mathbb{R}^{D\times T}$, which is encoded by the optimized model. Then, the word features are converted to a common space by adding a perceptron layer. Specifically, word feature $\widehat{\boldsymbol{\varphi}} \in \mathbb{R}^{D\times T}$ is converted by $\widehat{\boldsymbol{\varphi}} = \mathbf{U}\boldsymbol{\varphi}$, where $\mathbf{U} \in \mathbb{R}^{\hat{D}\times D}$. Then, we calculate the word-context vector of the j-th sub-region by attention mechanism. Hidden feature h denotes the query, and the converted word features are the value. In detail, the word-context of the j-th sub-region is calculated as follows.

$$c_j = \sum_{i=0}^{T-1}\beta_{j,i}\hat{\varphi}_i, \tag{13}$$

where

$$\beta_{j,i} = \frac{\exp(s'_{j,i})}{\sum_{k=0}^{T-1}\exp(s'_{j,k})}. \tag{14}$$

Here, the similarity is computed by dot-product similarity

$$s'_{j,i}{=}h_j^T\hat{\varphi}_i. \tag{15}$$

In short, the word-context can be denotes as

$$\mathbf{C} = F^{cro\_attn}(\boldsymbol{\varphi}, h) = (c_0, c_1, \ldots, c_{N-1}) \in \mathbb{R}^{\hat{D}\times N} \tag{16}$$

Then, the word-context and original image hidden feature is concatenated and feed in next layer.

(3) *Objective function of multi-stage GANs*

In our work, we adopt three generators and three discriminators in text-image translation. Each stage of generator $G_i(i = 0, 1, 2)$ has a corresponding discriminator D_i. The same with the conditional GANs, the objective function of the i-th generator is defined as follows.

$$L_{G_i} = -\frac{1}{2}E_{\hat{x}_i \sim p_{G_i}}[\log(D_i(\hat{x}_i))] - \frac{1}{2}E_{\hat{x}_i \sim p_{G_i}}[\log(D_i(\hat{x}_i, \bar{\varphi}))], \tag{17}$$

where the first part is unconditional loss and the second term is conditional loss. Meanwhile, in order to ensure the generated image is match with the text description, we introduce the DAMSM loss [16] into the objective function of the last-stage generator. As following,

$$L = L_{G_2} + \lambda_2 L_{DAMSM}, \tag{18}$$

where λ_2 is a balance factor.

In the adversarial learning, the discriminators evaluate whether the synthesized image is realistic and matched with the text or not. The objective function of each stage discriminator is defined as follows.

$$\begin{aligned} L_{D_i} = & -\frac{1}{2}E_{x_i \sim p_{data}}[\log(D_i(x_i))] - \frac{1}{2}E_{\hat{x}_i \sim p_{G_i}}[\log(1 - D_i(\hat{x}_i))] + \\ & -\frac{1}{2}E_{x_i \sim p_{data_i}}[\log(D_i(x_i, \bar{e}))] - \frac{1}{2}E_{\hat{x}_i \sim p_{G_i}}[\log(1 - D_i(\hat{x}_i, \bar{\varphi}))], \end{aligned} \tag{19}$$

where x_i is from the real i-th scale image and $\hat{x}_i$ the generated image from the i-th stage generator. By optimizing the discriminator and generator alternately, the network will achieve zero-sum game that the generators and discriminators obtain the best performance.

B. Spectral Normalization for Stabilizing Training

Model Collapse, gradient vanishing and gradient explosion are very popular phenomena in the training of GANs. Besides, the balance of training between generator and discriminator is hard to control, which leads to converge difficultly. In order to solve the problem, many methods were proposed to improve the stability of model, such as WGAN [6] and WGAN-GP [7]. The original WGAN introduces Wasserstein distance to measure the distance between the real data and the generated data and minimize it. The Wasserstein distance is calculated as follows,

$$W(P_r, P_g) = \sup_{\|f\|_{Lip} \leq K} E_{x \sim P_r}[f(x)] - E_{x \sim P_g}[f(x)]. \tag{20}$$

Here, the formula $\|f\|_{Lip} \leq K$ indicates that the function $f(\bullet)$ is satisfied with K-Lipschitz constraint. The original WGAN presents a way of clipping the weights of discriminator in $[-c, c]$, which drops the fitting capacity of deep neural network. WGAN-GP adopts Gradient Penalty in discriminator to satisfy K-Lipschitz constraint, which increases computational effort. Therefore, those methods could not solve the problem absolutely. For stabilizing GAN-based model, the discriminator D should follow the Lipschitz continuity hypothesis. In other words, we need constrain the function of discriminator to satisfy the K-Lipschitz constraint.

$$\underset{\|f\|_{Lip} \leq K}{\arg\max}\, V(G, D), \tag{21}$$

where the $\|f\|_{Lip}$ is the smallest value of K such that $|f(x_1) - f(x_2)| \leq K|x_1 - x_2|$ for any x_1, x_2. Miyato et al. [8] propose a novel weight normalization named spectral normalization, which stabilize the training of discriminator by forcing the network to satisfy the Lipschitz constraint. Therefore, normalizing the weight parameters W of each layer can ensure the $\|f\|_{Lip}$ is bounded from above by 1. As following

$$\|\nabla_x(f(x))\|_2 = \|D_N \frac{W_N}{\sigma(W_N)} \cdots D_1 \frac{W_1}{\sigma(W_1)}\|_2 \leq \prod_{i=1}^{N} \frac{\sigma(W_i)}{\sigma(W_i)} = 1. \tag{22}$$

where $\sigma(W)$ is spectral normalization and D_N is nonlinear activation function of the N-th layer. With spectral normalization, the discriminator provides useful gradient to generator for optimization so that the network optimize better and generate images that are more realistic.

3 Experimental Results and Evaluation

3.1 Datasets and Evaluation Metric

We conduct experiments for text to image synthesis on the widely used CUB dataset [21] and Oxford-102 dataset [22]. The statistics of each datasets as shown in Table 1. In order to verify the effectiveness fairly, Inception Score (IS) [23] and Fréchet Inception Distance [24] are adopted for quantitative evaluation of generative model.

Table 1. Statistics of the datasets.

Datasets	CUB		Oxford-102	
	Train	Test	Train	Test
Number of samples	8,855	2,933	7,034	1,155
Captions/image	10	10	10	10
Categories	200		102	

Inception Score. The Inception Score (IS) is current well-known metric for evaluating the generative performance of GANs. The motivation of Inception Score is that excellent generative models should generate realistic, various and meaningful images. The calculation of IS score as follows.

$$IS = \exp(E_{X\sim P_G}[KL(P_{Y|X}(y|x))||P_Y(y)]), \tag{23}$$

where x denotes sample of generated image, and y is image label predicted by the inception model. The Eq. (22) indicates that classes of generated image should be as diverse as possible and the label prediction probability should be as accurate as possible. Therefore, the higher KL divergence shows excellent generative ability of model.

Fréchet Inception Distance. Assuming that both the real data and the generated data distribution following Gaussian distribution, so they have two major parameters, mean and covariance (m, C). The distance between the two data distribution is measured by Fréchet distance. The calculation is as following.

$$FID = ||m - m_r||_2^2 + Tr(C + C_r - 2(CC_r)^{\frac{1}{2}}), \tag{24}$$

where (m, C) are mean and covariance of generated data, and (m_r, C_r) are mean and covariance of real data. The lower distance of the mentioned two distributions presents that the synthesized image are more similar to the original data.

3.2 Experimental Results and Comparison

(1) Evaluation Metric Comparison

In experiment, we make quantitative and non-quantitative comparison with many state-of-art methods. Tables 2 and 3 show the quantitative comparison details of IS and FID score on Oxford-102 dataset and CUB dataset. For fair comparison, we choose some IS and FID value from the published paper [14, 15]. On the Oxford-102 dataset, the proposed method achieves 3.95 of inception score and 47.32 of Fréchet Inception Distance, which outperforms the previous methods. Likewise, the proposed method obtains the highest IS value (from 4.36 to 4.43) and competitive FID value (44.64). Comparing to the Oxford dataset, the CUB dataset is more difficult for text to image generation. The bird dataset can better reflect the performance of different methods. Significantly, the results show that the proposed method is able to achieve better performance than other state-of-art text to image synthesis methods.

Table 2. Fréchet Inception Distance and Inception Score for the Oxford-102 dataset.

Model	Resolution	FID	IS
GAN-INT-CLS	64 × 64	79.55	2.66 ± 0.03
WGAN-CLS	64 × 64	–	3.11 ± 0.02
WGAN-CLS with TTUR	64 × 64	–	3.20 ± 0.01
StackGAN-v1	256 × 256	55.28	3.71 ± 0.04
StackGAN-v2	256 × 256	48.68	3.82 ± 0.06
CLSPGGAN	256 × 256	–	3.76 ± 0.03
CWPGGAN	256 × 256	–	3.86 ± 0.02
AttnGAN	265 × 256	50.24	3.89 ± 0.02
HAGAN	256 × 256	47.32	3.95 ± 0.03

Table 3. Fréchet Inception Distance and Inception Score for the CUB dataset.

Model	Resolution	FID	IS
GAN-INT-CLS	64 × 64	68.79	2.88 ± 0.04
StackGAN-v1	256 × 256	51.89	3.70 ± 0.04
StackGAN-v2	256 × 256	–	3.82 ± 0.06
GAWWN	256 × 256	67.22	3.62 ± 0.07
CWPGGAN	256 × 256	–	4.09 ± 0.03
AttnGAN	256 × 256	46.43	4.36 ± 0.04
HAGAN	256 × 256	44.64	4.43 ± 0.03

(2) Visual Effect Comparison

The comparisons of state-of-art text-to-image generative methods by side-by-side comparison are shown in Fig. 2. Life part of Fig. 2 is various images generated by different methods, which are conditioned on the same text description of the Oxford-102 dataset. By scrutinizing the image details and text description roughly, the results show that all images generated by different methods matches with the text, and all those images are realistic and natural. However, the detailed comparison indicates that the image generated by our method are more realistic. On the challengeable CUB dataset, we can find that some previous methods have difficulty in generating highly real and clear image conditioned on the given text, such as GAN-CLS, GAWWN and

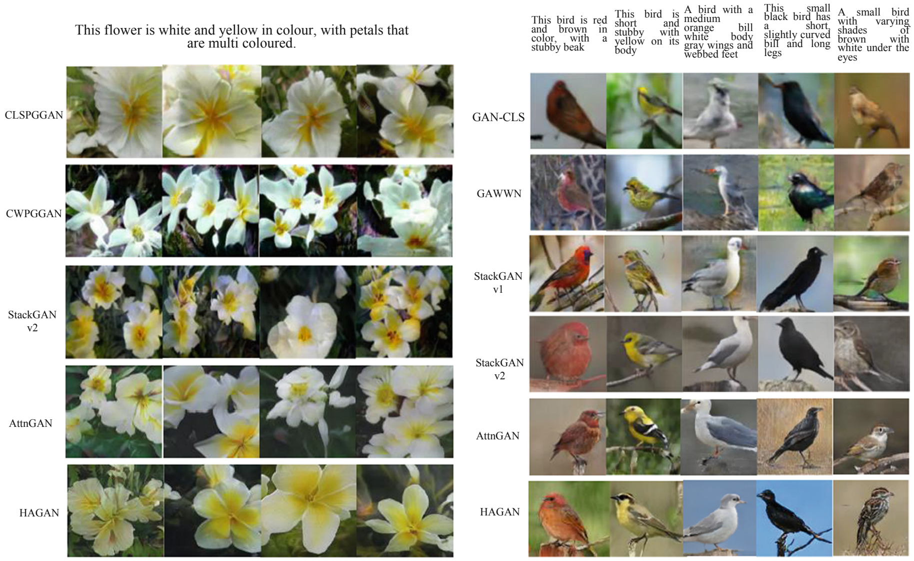
This flower is white and yellow in colour, with petals that are multi coloured.
CLSPGGAN
CWPGGAN
StackGAN v2
AttnGAN
HAGAN
This bird is red and brown in color, with a stubby beak
This bird is short and stubby with yellow on its body
A bird with a medium orange bill white body gray wings and webbed feet
This small black bird has a short, slightly curved bill and long legs
A small bird with varying shades of brown with white under the eyes
GAN-CLS
GAWWN
StackGAN v1
StackGAN v2
AttnGAN
HAGAN

Fig. 2. Side-by-side comparison on the Oxford-102 dataset and CUB bird dataset.

StackGAN_v1. On the contrary, our proposed method can generate photo-realistic and fine-grained image, especially the bird of the third column. Therefore, in conclusion, our proposed method generates more realistic, more fine-grained and more natural images than other methods in visual evaluation.

(3) Word-Level Attention Visualization

For better evaluating the performance of attention mechanism, we visualize the word-level attention results as shown in Fig. 3. The attention visualization are shown below the red box. The words belong to the paired text description, and the bright region is the corresponding attention area of the words. However, some words do not give attention to right area, such as articles and verbs, which make less contribution to image synthesis. The words describing object attributes, such as colours, shape, and parts of objects, can give attention to correct regions. With adding word-level semantic information in the latter two generators, the generators can redraw the word's information in the corresponding region, which can saliently enhance the significant details of generated image as well as make it be suitable for the human system (Fig. 4).

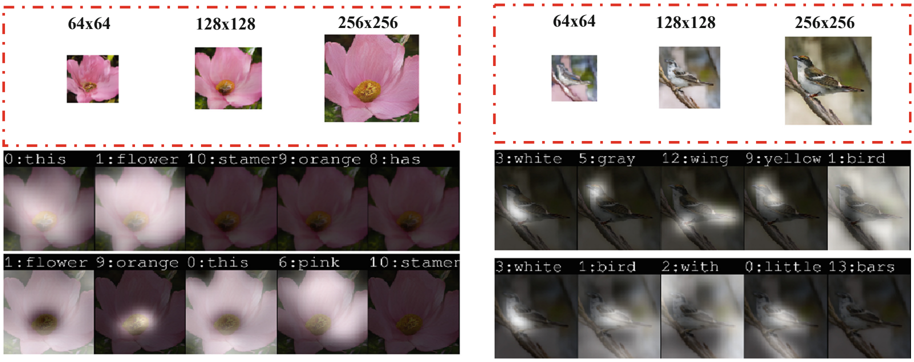

Fig. 3. Word-level attention visualization of the Oxford-102 flower dataset and CUB bird dataset.

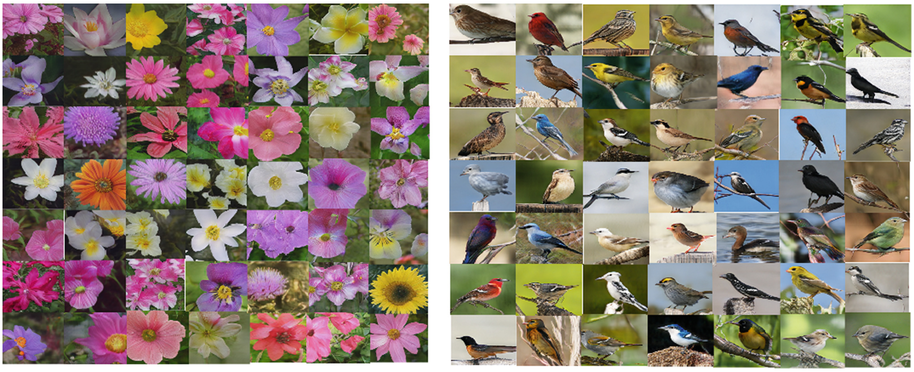

Fig. 4. More examples synthesized by our proposed method.

4 Conclusion

This paper presents a hybrid attentional model to fulfill text-to-image synthesis. The hybrid attentional mechanism contributes to improve performance of generating fine-grained and realistic image. Meanwhile, the training of network become more stable by introducing spectral normalization in discriminator network. The conducted experiments show that our proposed method synthesizes realistic images in visual comparison, and outperforms the state-of-the-art approaches in FID and IS metric.

Acknowledgments. This work is supported by National Natural Science Foundation of China under grant 61771145 and grant 61371148.

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
2. Huang, H., Yu, P.S., Wang, C.: An introduction to image synthesis with generative adversarial nets. arXiv preprint arXiv:1803.04469 (2018)
3. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
4. Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T.: Adversarial cross-modal retrieval. In: Proceedings of the 2017 on Multimedia Conference, pp. 154–162. ACM Press, California (2017)
5. Arjovsky, M., Bottou, L.: Towards principled methods for training generative adversarial networks. arXiv preprint arXiv:1701.04862 (2017)
6. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN. arXiv preprint arXiv:1701.07875 (2017)
7. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of Wasserstein GANs. In: Advances in Neural Information Processing Systems, pp. 5767–5777 (2017)
8. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957 (2018)
9. Cai, L., Gao, H., Ji, S.: Multi-stage variational auto-encoders for coarse-to-fine image generation. arXiv preprint arXiv:1705.07202 (2017)
10. Gulrajani, I., et al.: PixeLVAE: a latent variable model for natural images. arXiv preprint arXiv:1611.05013 (2016)
11. Gregor, K., Danihelka, I., Graves, A., Rezende, D.J., Wierstra, D.: Draw: a recurrent neural network for image generation. In: International Conference on Machine Learning, Lille, pp. 1462–1471 (2015)
12. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396 (2016)
13. Zhang, H., et al.: StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5907–5915. IEEE Press, Venice (2017)
14. Han, Z., Tao, X., Hongsheng, L., Shaoting, Z., Xiaogang, W., Xiaolei, H.: StackGAN++: realistic image synthesis with stacked generative adversarial networks. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 1947–1962 (2018)

15. Bodnar, C.: Text to image synthesis using generative adversarial networks. arXiv preprint arXiv:1805.00676 (2018)
16. Xu, T., et al.: AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. arXiv preprint arXiv:1711.10485 (2017)
17. Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. In: Advances in Neural Information Processing Systems, pp. 217–225 (2016)
18. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
19. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318 (2018)
20. Mansimov, E., Parisotto, E., Ba, J.L., Salakhutdinov, R.: Generating images from captions with attention. In: International Conference on Learning Representations, San Juan (2016)
21. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 dataset (2011)
22. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: IEEE Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pp. 722–729. IEEE Press, Bhubaneswar (2008)
23. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems, pp. 2234–2242 (2016)
24. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a nash equilibrium. arXiv preprint arXiv:1706.08500 (2017)