2020 IEEE International Conference on Information Technology,Big Data and Artificial Intelligence (ICIBA 2020)

DCFGAN: Dynamic Convolutional Fusion Generative Adversarial Network for Text-to-Image Synthesis

Ming Tao, Songsong Wu, Xiaofeng Zhang, Cailing Wang
School of Automation, Nanjing University of Posts and Telecommunications, NanJing, China
Email: mingtao2000@126.com, sswuai@126.com, SemiZxf@163.com, wangcl@njupt.edu.cn
Corresponding Author: Cailing Wang Email: wangcl@njupt.edu.cn

***Abstract*—Text-to-image synthesis is the task of synthesizing realistic and text-matching images according to given text descriptions. Most text-to-image generative networks consist of two modules: a pre-trained text-image encoder and a text-to-image generative adversarial network. In this paper, we propose a stronger text encoder which employs a text Transformer to extract semantically meaningful parts from text descriptions. With the stronger text encoder, the generator can obtain more meaningful text information to synthesize realistic and text-matching images. In addition, we propose a Dynamic Convolutional text-image Fusion Generative Adversarial Network (DCFGAN) which employs the Dynamic Convolutional Fusion Block to fuse text and image features efficiently. The Dynamic Convolutional Fusion block adjusts the parameters in the convolution layer according to different text descriptions to synthesize text-matching images. It improves the efficiency of fusing text features and image features in generator network. We evaluate the proposed DCFGAN on two benchmark datasets, the CUB and the Oxford-102. The extensive experiments demonstrate that our proposed stronger text encoder and Dynamic Convolutional Fusion Layer can greatly promote the performance of text-to-image synthesis.**

***Index Terms*—Text-to-Image synthesis, generative adversarial network, cross-modal, image-text matching, deep representation learning.**

I. Introduction

Text-to-image synthesis is one of the most important generative tasks in recent years. It aims to synthesize text-matching images which bridges the gap between text descriptions and realistic images. Since its interesting and promising application, it has attracted many researchers and achieved many significant progresses [1], [2], [3], [4].

Current text-to-image models employ a simple BiLSTM [5] to encode text descriptions [2], [4], [6], [7]. Each hidden state in BiLSTM represents the semantic meaning of a word in a text description. Meanwhile, the last hidden state is employed as the global sentence vector [6]. During image generation process, the global sentence vector is concatenated on the input noise, and the word vectors are employed to compute word-context vectors through cross-modal attention mechanism, then the word-context vectors are concatenated on the middle visual feature maps to synthesize text-matching images [6], [2], [4].

However, there are two problems of current text encoders. First, each word in a text description contributes a different level of importance when pre-training the text-image encoder. Treating all words in a sentence equally reduces the efficiency of computing the word-context vector during image generation process. Second, the semantic meaning in the last hidden state of BiLSTM is more relevant to the last few words and cannot summarize the entire sentence information well.

For dealing this two problems, we introduce the Transformer architecture [8] to distill semantically meaningful parts from word features. The multi-head attention mechanism in Transformer can help the text-encoder distill important word-level information from all word vectors in a sentence. Moreover, we employ a word attention mechanism to summarize the entire sentence for a better global sentence vector.

In addition, for improving the efficiency of generator to learn text-specific features, we introduce the conditional convolution layer [9] and propose a Dynamic Convolutional text-image Fusion Blcok (DCFB) which adjusts the parameters in the convolution layer according to different text descriptions during image generation process. Compared with ordinary convolution layer, the Dynamic Convolutional Fusion Block is more efficient to synthesize text-matching images.

The contribution of our DCFGAN can be summarized as follows:

- We propose a stronger text encoder which employs the Transformer and word attention mechanism to extract semantically meaningful features from text descriptions.
- We propose the Dynamic Convolutional Fusion Block (DCFB) to help the generator learn text-specific features more efficiently.
- The experimental results on the CUB and Oxford-102 datasets prove that our proposed modules can promote the performance of text-to-image models effectively.

II. Related work

A. Generative Adversarial Networks

The Generative Adversarial Network (GAN) [10] is a successful model to synthesize realistic images. It is composed of the generator and discriminator networks. The generator attempts to synthesize plausible images to cheat the discriminator, and the discriminator distinguishes between real images and synthesized images. The ability of generator and

978-1-7281-5224-0/20/$31.00 ©2020 IEEE

the discriminator are all improved during adversarial training process. And the synthesized images become more and more realistic. Since its powerful generative ability, GAN has been widely applied in many tasks, such as image inpainting [11], image super resolution [12], image editing [13], etc. Text-to-image synthesis is one of the most important application of GANs [14], [15], [16], [17].

B. Text-to-Image Synthesis

Text-to-image synthesis aims to synthesize realistic and text-consistent images according to text descriptions. Reed *et al.* first employ the GAN architecture to synthesize low-resolution images from text descriptions [18], [14]. Zhang *et al.* proposed StackGAN which stacks multiple generators to synthesize high-resolution images [19]. Xu *et al.* proposed AttnGAN which employs the cross-modal attention to synthesize more fine-grained features based on StackeGAN, and proposed a Deep Attentional Multimodal Similarity Model (DAMSM) which encodes text and image features into a common semantic space to measure text-image similarity [6]. Most following text-to-image models [3], [4], [2], [7] are based on AttnGAN. They employ the pre-trained DAMSM model to encode text descriptions and image features. The DAMSM model employs a bi-directional Long Short-Term Memory (BiLSTM) as text encoder to extract word-level features in text descriptions. And the last hidden state of BiLSTM is adopted as the global sentence feature. The global sentence feature and word-level features will be concatenated on the input noise and middle visual feature maps in generator network to synthesize text-matching images. The text encoder in DAMSM treats all words in a text description equally. However, the importance of each word is different in a text description. And the last hidden state contains more semantic information about the last few words. It makes the extracted word-level features and global sentence feature cannot represent the entire sentence information well. Different from previous models, our proposed text Transformer is able to distill important word-level semantics from all word vectors in a sentence. And we employ a word attention mechanism to summarize the entire sentence information. Furthermore, we introduce the conditional convolution layer [9] and propose Dynamic Convolutional text-image Fusion Blcok (DCFB) to help the generator learn text-specific features.

III. Proposed Methods

In this section, we start to present our proposed DCFGAN in detail. We first introduce the text Transformer and word attention mechanism, then illustrate the proposed Dynamic Convolutional Fusion Blcok (DCFB).

A. A Stronger Text Encoder

Although current text-image encoder has shown its effectiveness in many works, it cannot distinguish the importance of different words in a text description. The text descriptions always contain some semantic-irrelevant words which reduce the efficiency of computing the semantic similarity between text descriptions and images during pre-training process.

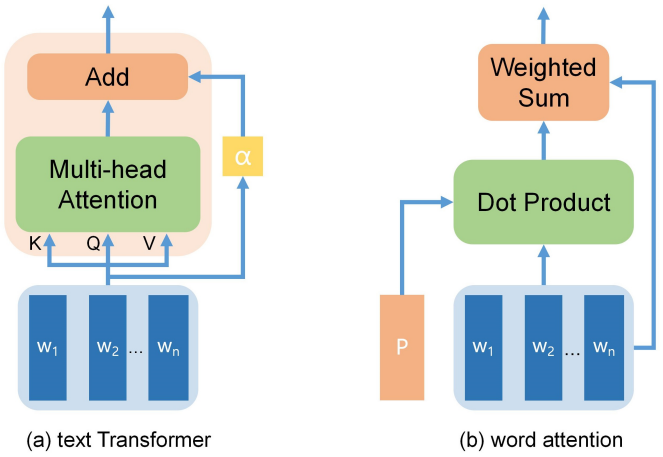

Fig. 1. Illustrations of the proposed text Transformer and word attention mechanism. The multi-head attention mechanism in text Transformer distills important word-level information from word vectors in a sentence. The word attention mechanism summarizes the entire sentence for a better global sentence vector.

To deal this problem, we introduce the Transformer architecture [8] to refine the word features extracted by BiLSTM [5]. Since the image encoder is an Inception V3 network [20] pre-trained on ImageNet [21]. And its parameters are fixed during the text-image encoder training process. So we adjusted the Transformer architecture to better adapt to the text-image similarity computation task. We remove the Feed Forward network and Layer Normalization layer. The architecture of our text Transformer is shown in Figure 1(a).

We first encode the word vectors into keys (K), queries (Q) and values (V) of dimension d. Next, we calculate the dot products between all queries and keys, divide the dot products by $\sqrt{d}$. We then apply the softmax function to get the attention weights on the values. The attended features are obtained as a weighted summation over all values V through the attention distribution learned before. The dot-product attention can be formally presented as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

Moreover, we found the Multi-head Attention (MA) can promote the representation capacity of the attended features. It consists of h paralleled attention heads which allow the text Transformer to jointly attend to word features from different representation subspace. Each attention head processes dot-product attention independently. With Multi-head Attention (MA), the calculation process of the attended features can be expressed as:

$$\begin{aligned} MA(Q, K, V) &= Concat(head_1, ..., head_h)P^O \\ head_i &= Attention(QP_i^Q, KP_i^K, VP_i^V) \end{aligned} \tag{2}$$

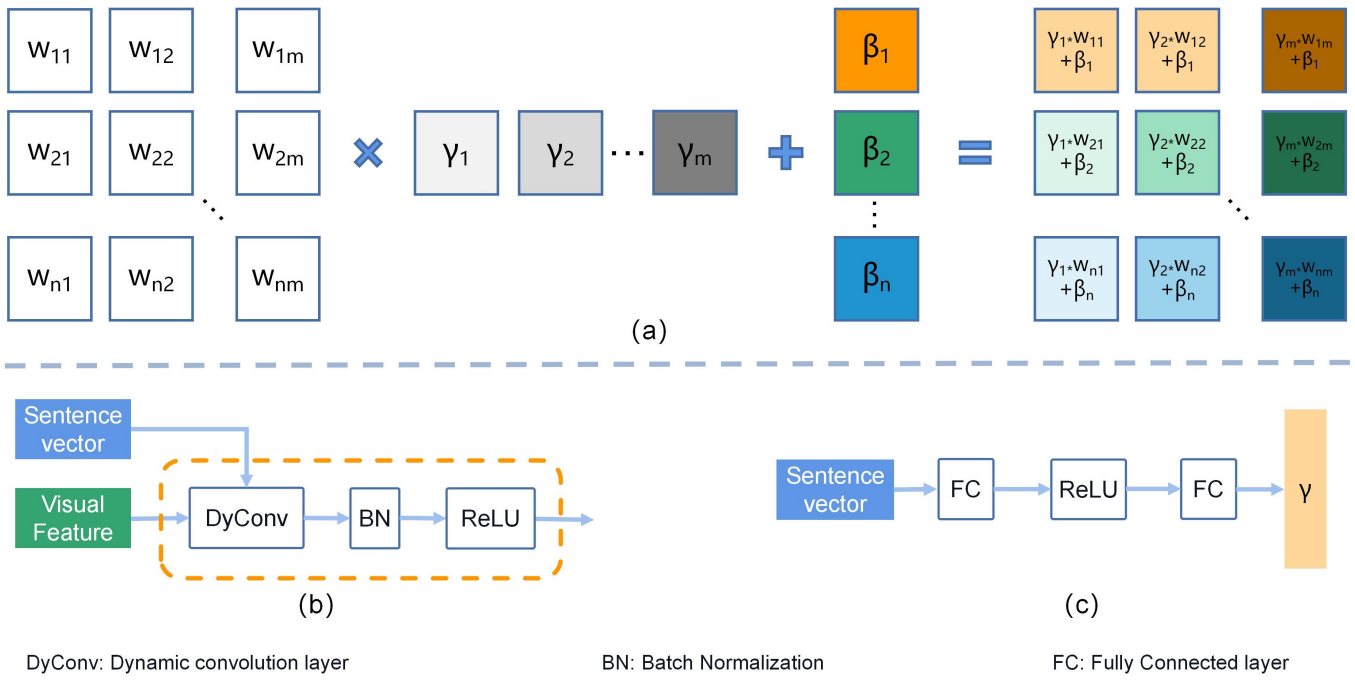

Fig. 2. Illustration of the proposed Dynamic Convolutional text-image Fusion Block (DCFB). (a) The filter-wise scaling and channel-wise shifting operations in conditional convolution layer. (b) The architecture of the proposed DCFB. (c) The sub-network which computes the filter-wise scaling parameter γ.

Where P_i^Q, P_i^K, P_i^W are the projection matrices for the i-th attention head and P^O is a projection matrix which fuses all attended features calculated by different attention heads. To stabilize the training process, the final attended feature is multiplied by a trainable scale parameter α and added back the input word vectors (see Figure 1(a)). In our text encoder, we stack 6 text Transformers behind the BiLSTM.

After refining the word features, we employ a word attention mechanism to compute the global sentence vector. The word attention mechanism is shown in Figure 1(b). There is a trainable vector $\boldsymbol{p}$ of dimension d. We first calculate the dot products between $\boldsymbol{p}$ and each word vector in word features W refined from text Transformers. Next, we apply the softmax function to get the attention weight of each word vector. Then, the global sentence vector $\boldsymbol{s}$ can be obtained as a weighted summation over word vectors. It can be formally expressed as:

$$\boldsymbol{s} = softmax(\frac{\boldsymbol{p}W^T}{\sqrt{d}})W \tag{3}$$

The word attention mechanism calculates the global sentence vector dynamically according to the importance of each word. It improves the quality and representation capacity of the global sentence vector, which promotes generator to synthesize text-matching images.

B. Dynamic Convolutional text-image Fusion Block

Synthesizing text-matching images is very important for text-to-image models, it requires the model to be able to fuse the text information into the image features effectively. However, it is still challenging for current text-to-image methods to synthesize text-specific features [15], [1], [6]. Current text-to-image models typically concatenate the text vector or word-context features on the noise or intermediate visual features. They employ standard ordinary convolution layers which treat different features maps with same parameters regardless their different text conditions.

Inspired by conditional convolution layer proposed for class conditional generation task [9], we propose the Dynamic Convolutional text-image Fusion Block (DCFB) (see Figure 1(b)) for effectively fusing text and image features. We replaced the original ordinary convolutional layer with a text-conditioned convolutional layer. It modulates the parameters in convolution layer dynamically according to the linguistic cues from given text descriptions. Armed with DCFB, the generator can synthesize text-specific features more efficiently.

The proposed DCFB modulates the parameters in convolution layer through filter-wise scaling and channel-wise shifting operations conditioned on given text descriptions (see Figure 1(a)). If we ignore the bias parameters, the weights in the convolution kernel can be regarded as a 4-dimensional tensor $W \in \mathbb{R}^{m\times n\times k\times k}$, where m and n are the output and input channel size, k is the filter size. It means there are m filters in a convolution layer, and each filter contains n channels. As shown in Figure 2(c), we employ one-hidden-layer MLPs to predict the filter-wise scaling parameters $\boldsymbol{\gamma} \in \mathbb{R}^m$ and the channel-wise shifting parameters $\boldsymbol{\beta} \in \mathbb{R}^n$. Then, each filter in

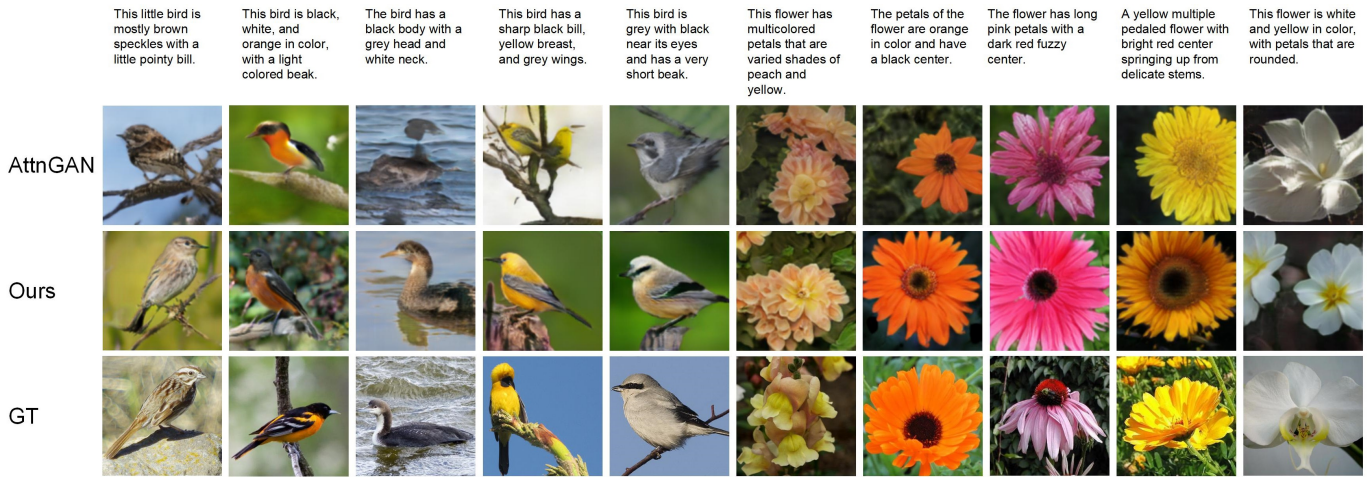

Fig. 3. Ground truth (GT) and samples synthesized by AttnGAN [6] and our proposed DCFGAN conditioned on text descriptions from the test set of CUB and Oxford-102 datasets.

convolution layer will be multiplied by $\boldsymbol{\gamma}$, and each channel in all filters will be added with $\boldsymbol{\beta}$ (see Figure 2(a)). This Conditional Modulation (CM) can be expressed as follows:

$$CM(\boldsymbol{w}_{i,j}) = \boldsymbol{w}_{i,j} \cdot \gamma_i + \beta_j \tag{4}$$

where $\boldsymbol{w}_{i,j}$ is the j-th channel of the i-th filter. Scaled and shifted by γ and β, the weights in convolution layer will be text-adaptive. Compared with previous ordinary convolutional Block, the Dynamic Convolutional text-image Fusion Block is more effective to synthesize text-specific features.

IV. Experiments

A. Datasets

We evaluate our proposed DCFGAN on CUB [22] and Oxford-102 datasets [23]. The CUB dataset contains 11788 images of 200 bird species. We divided 150 species (8841 images) for training and 50 species (2947 images) for testing as previous work. Oxford-102 is another dataset which contains 8189 images of 102 flower species. We divided 82 species (7034 images) for training and 20 species (1155 images) for testing. Each image in CUB and Oxford-102 datasets has 10 text descriptions.

B. Evaluation Metric

As previous methods, we employ the Inception Score (IS) [24] to evaluate the performance of our DCFGAN. The Inception Score (IS) is computed by a pre-trained Inception V3 [20] network. Since the Inception V3 network was only fine-tuned on the test set of CUB and Oxford-102, respectively. It makes the Inception Score (IS) can evaluate the synthesized image quality and text-image semantic consistency during testing.

TABLE I
THE IS OF STACKGAN, STACKGAN++, HDGAN, ATTNGAN AND DCFGAN ON THE TEST SET OF CUB AND OXFORD-102.

Methods	CUB↑	Oxford-102↑
StackGAN [19]	3.70±0.04	3.20±0.01
StackGAN++ [15]	3.84±0.06	-
HDGAN [1]	4.15±0.05	3.45±0.04
AttnGAN [6]	4.36±0.04	3.75±0.04
DCFGAN (Ours)	**4.58±0.04**	**3.80±0.04**

C. Comparative Results

We computed the Inception Score (IS) of our model and compare it with previous models. The results are listed in Table I. Our DCFGAN achieves the highest IS on both CUB and Oxford-102 datasets. On the CUB dataset, our DCFGAN achieves the IS 4.58±0.04, which significantly outperforms the previous text-to-image models. Compared with the baseline model AttnGAN, our DCFGAN also improves the IS from 4.36 to 4.58 on CUB dataset, and improves the IS from 3.75 to 3.80 on Oxford-102 dataset. The quantitative results demonstrate the superiority of DCFGAN over previous text-to-image methods.

In Figure 3, we show some samples synthesized by baseline model AttnGAN [6] and our DCFGAN on the test set of CUB and Oxford-102. Compared with AttnGAN, the images synthesized by our DCFGAN are more realistic and have more fine-grained details. Since we employ the Dynamic Convolutional text-image Fusion Block (DCFB), our model can synthesize more text-specific features like "sharp black bill", "grey with black near its eyes", "dark red fuzzy center" (4^{th}, 5^{th} and 8^{th} column). The qualitative results also demonstrate the effectiveness of our DCFGAN.

TABLE II
THE QUANTITATIVE ASSESSMENT OF DIFFERENT VARIANTS BY REMOVING OUR PROPOSED MODULES FROM DCFGAN.

Architecture	Inception Score ↑
DCFGAN	4.58±0.04
- text T	4.52±0.04
- text T - word Attn	4.47±0.04
- text T - word Attn - DCFB	4.37±0.04

D. Ablation Study

To verify the effectiveness of our proposed modules, we conduct ablation studies on CUB datasets. By removing our proposed modules from DCFGAN, we can verify the effectiveness of different modules. The experimental results are listed in Table II. We can find that the Dynamic Convolutional text-image Fusion Block (DCFB) improves the IS from 4.37 to 4.47, the word attention mechanism (word Attn) improves the IS from 4.47 to 4.52 and the text Transformer (text T) improves the IS from 4.52 to 4.58. The ablation studies demonstrate the effectiveness of every proposed modules in DCFGAN.

V. CONCLUSION

In this paper, we propose a novel architecture called Dynamic Convolutional Fusion Generative Adversarial Network (DCFGAN) for text-to-image synthesis task. We propose a stronger text encoder which employs the Transformer architecture to extract semantically meaningful parts from text descriptions, and applies the word attention mechanism to summarize the entire sentence for a better global sentence vector. Moreover, we propose the Dynamic Convolutional text-image Fusion Block which adjusts the parameters in convolution layer to synthesize text-specific visual features. Experimental results on two benchmark datasets demonstrate the effectiveness of our DCFGAN by qualitative and quantitative measures.

ACKNOWLEDGMENT

The work described in this paper was supported by Nanjing University of Posts and Telecommunications General School Project (NY220057).

REFERENCES

[1] Z. Zhang, Y. Xie, and L. Yang, "Photographic text-to-image synthesis with a hierarchically-nested adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6199–6208. 1, 3, 4

[2] M. Zhu, P. Pan, W. Chen, and Y. Yang, "Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5802–5810. 1, 2

[3] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang, and J. Shao, "Semantics disentangling for text-to-image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2327–2336. 1, 2

[4] W. Li, P. Zhang, L. Zhang, Q. Huang, X. He, S. Lyu, and J. Gao, "Object-driven text-to-image synthesis via adversarial training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 174–12 182. 1, 2

[5] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997. 1, 2

[6] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1316–1324. 1, 2, 3, 4

[7] Y. Gou, Q. Wu, M. Li, B. Gong, and M. Han, "Segattngan: Text to image generation with segmentation attention," *arXiv preprint arXiv:2005.12444*, 2020. 1, 2

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008. 1, 2

[9] M.-C. Sagong, Y.-G. Shin, Y.-J. Yeo, S. Park, and S.-J. Ko, "cgans with conditional convolution layer," *arXiv preprint arXiv:1906.00709*, 2019. 1, 2, 3

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680. 1

[11] R. Yeh, C. Chen, T. Lim, M. Hasegawa-Johnson, and M. Do, "Semantic image inpainting with perceptual and contextual losses. arxiv 2016," *arXiv preprint arXiv:1607.07539*, 2016. 2

[12] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690. 2

[13] T. R. Shaham, T. Dekel, and T. Michaeli, "Singan: Learning a generative model from a single natural image," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4570–4580. 2

[14] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Advances in neural information processing systems*, 2016, pp. 217–225. 2

[15] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE TPAMI*, vol. 41, no. 8, pp. 1947–1962, 2018. 2, 3, 4

[16] B. Li, X. Qi, T. Lukasiewicz, and P. Torr, "Controllable text-to-image generation," in *Advances in Neural Information Processing Systems*, 2019, pp. 2065–2075. 2

[17] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Learn, imagine and create: Text-to-image generation from prior knowledge," in *Advances in Neural Information Processing Systems*, 2019, pp. 887–897. 2

[18] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proceedings of the International Conference on Machine Learning*, 2016, pp. 1060–1069. 2

[19] 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 *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5907–5915. 2, 4

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826. 2, 4

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105. 2

[22] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011. 4

[23] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. IEEE, 2008, pp. 722–729. 4

[24] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in neural information processing systems*, 2016, pp. 2234–2242. 4