Stacking VAE and GAN for Context-aware Text-to-Image Generation

Chenrui Zhang and Yuxin Peng*
Institute of Computer Science and Technology, Peking University
Beijing 100871, China
pengyuxin@pku.edu.cn

***Abstract*—Generating high-quality images based on text descriptions is an appealing research topic, which has widespread applications in various fields. However, it is quite challenging since that images and language descriptions in real world are noisy with great variability. Most existing text-to-image methods aim to generate images in a holistic manner, which ignore the difference between images' foreground and background, resulting in that objects in images are easily disturbed by the background. Moreover, they commonly ignore the complementarity of different kinds of generative models. In this paper, we propose a context-aware approach to perform text-to-image generation, which separates background and foreground for generating high-quality images, as well as utilizes complementarity between Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) for robust text-to-image generation. First, context-aware conditional VAE is proposed to capture images' basic layout and color based on text, which pays different attention on the background and foreground of images for effective text-image alignment. Then, conditional GAN is adopted for refining the generation of VAE, which recovers lost details and corrects the defects for realistic image generation. Attributed to such stacked VAE-GAN structure, two kinds of generative models can boost each other for more effective and stable text-to-image generation. Experimental results on 2 widely-used datasets empirically verify the effectiveness of our proposed approach.**

***Index Terms*—text-to-image, context-aware learning, deep generative model**

I. Introduction

As an attractive research topic, text-to-image generation has been receiving extensive attention in computer vision and natural language processing communities [3], [6]. This task aims to generate realistic images conditioned on text description, which has widespread applications in various fields, including photo editing and data augmentation, *etc*. Meanwhile, learning to generate is a promising paradigm for semi-supervised and unsupervised learning, which can provide meaningful hints for deep models' interpretability [1].

Due to the noisy nature of both images and language "in the wild", generating high-quality images that match abstract text description is challenging. Attributed to the advance of representation learning in both natural language processing [2] and computer vision [5], text-to-image generation has made significant progress recently. Deep generative models, such as Variational Autoencoder (VAE) and Generative Adversarial Network (GAN), are used to capture the complicated text-image matching relations for text-conditioned image generation. GAN-INT-CLS [3] method adopts conditional GAN with deep convolutional structure, which generates plausible images of flowers and birds. StackGAN [6] stacks two GANs for photo-realistic images generation given text description. It performs text-to-image generation via a sketch-refinement process, which is inspired by the drawing process of humans.

Despite the promising results, most existing text-to-image methods generate images in a holistic manner, while ignoring the difference between foreground and background of images. Furthermore, single kind of generative models have limited generation ability with their own shortcomings, resulting in suboptimal performance on realistic image generation. For example, images generated by VAE are easily blurred, while GAN is difficult to train. For addressing these problems, in this paper, we propose a context-aware approach to perform text-to-image generation, aiming to generate high-quality images in a decoupled and stacked manner. For the *decoupled* aspect, we decouple the background and foreground of images via a context-aware conditional VAE, forcing the model to learn more refined text-image alignments. For the *stacked* aspect, we decompose the text-to-image generation problem into two manageable sub-problems, i.e., the conditional VAE mentioned above first capture the key layout and color of images, and a conditional GAN is adopted for refining the generation of VAE, which boosts the quality of generated images significantly, and relieves instability of GAN for robust generation.

Our proposed approach can fully utilize the complementarity of VAE and GAN for more robust text-to-image generation solution, which mainly lies in the following aspects: (1) *Why need VAE:* GAN is hard to train, whose training process is unstable and sensitive to hyper-parameters [6], [7]. Moreover, it suffers from *mode collapse issue* [7], causing the generator collapses into certain parameter settings where a low diversity of samples are generated. Many existing works [25], [26] believe that such training difficulties are due to the disjoint supports of the data distribution and the generated distribution. VAE is known to be more stable than GAN without mode collapse issues, which can be utilized for fitting real distribution robustly and capturing the diversity of natural image. (2) *Why need GAN:* Images generated by VAE are easily blurred, which cannot meet the requirements of generating high-quality images. GAN is good at generating realistic images attributed

*Corresponding author.

978-1-5386-5321-0/18/$31.00 ©2018 IEEE

to adversarial learning, so it can be adopted for refining the output of VAE. To sum up, in our proposed framework, VAE captures the main factor of real data distribution robustly, and GAN refines images for realistic text-to-image generation.

To verify the effectiveness of the proposed approach, we conduct experiments on two widely-used image datasets, and the experimental results demonstrate that our approach can improve generation performance significantly.

II. Related Work

A. Deep Generative Models

1) Variational Autoencoder (VAE): VAE [15], [16] is a recently emerged deep generative model, which captures various data distributions such as image, audio or video by variational inference. The prior distribution in VAE is explicitly defined for back propagation through Monte-Carlo sampling, and its object can be divided into two parts, namely the log-likelihood and the KL divergence, forcing VAE to learn a latent space for good reconstruction performance.

2) Generative Adversarial Networks (GAN): As one of the most potential generative models, GAN [8] is studied extensively in recent years. GAN contains two submodels which work in an adversarial fashion, where the generator G fits the data distribution by synthesizing fake samples, while the discriminator D scores realism of the generated samples by comparing them with real samples. Mathematically, the goal of GAN is to learn a generator distribution $P_g(x)$ that matches real distribution $P_{real}(x)$ over data x via a minimax game:

$$\min_G \max_D V(G,D) = \mathbb{E}_{x\sim P_{real}}[\log D(x)] + \mathbb{E}_{z\sim P_{noise}}[\log(1-D(G(z)))]$$

B. Text-to-Image Generation

Generating images by text descriptions emerges as a promising research topic, which combines representation learning techniques in both computer vision and natural language process. It is difficult to generate high-quality images conditioned on high-level text description, since images and language in the real world are noisy with great variability. Recently, various methods are proposed for improving the quality of generated images. Reed et al. [3] adopt a conditional GAN based on convolutional architecture for text-to-image generation, which provides both generator and discriminator with text embedding, showing plausible generation performance on flowers and birds. Dong et al. [11] augment text data with synthesized captions, which boosts conditional generation from the text embedding perspective. Moreover, Mansimov et al. [12] propose an AlignDRAW model for aligning text and image by recurrent models. Besides the text embedding, Dash et al. [13] add conditions of class labels to further imporve the quality of generated images.

In addition to single generative model, there are also many works which cascade multiple models for better image generation. For example, Wang et al. [14] treat image as a combination of structure and style, and infer a surface normal map as an intermediate structure for indoor scene images generation. Zhang et al. [6] propose to generate photo-realistic image in a two-stage manner, which stacks two GANs for coarse-to-fine sketch-refinement process. Bao et al. [27] combine VAE and GAN with asymmetric loss functions, which generates images without text description. In contrast, our approach exploits CVAE-GAN architecture for text-to-image generation.

III. Our Proposed Approach

As illustrated in Fig. 1, our stacked framework contains two components, namely a context-aware conditional VAE (CVAE)[1] and a conditional GAN. The CVAE component captures basic layout and color via decoupling the background and foreground of images. The GAN component refines the output of CVAE with adversarial learning, which recovers lost details and corrects the defects for realistic image generation.

A. Preliminaries

A VAE contains two sub-models, namely an encoder (recognition model) and a decoder (generative model). The encoder $p(h|x)$ encodes the data instance x into a latent representation space, for learning the distribution of hidden variable h with a Kullback-Leibler (KL) divergence penalty. The decoder $q(x|h)$ decodes h back into the original data space via minimizing the reconstruction error:

$$h\sim En(x) = p(h|x),\ x\sim De(h) = q(x|h) \tag{1}$$

Formally, the objective of VAE can be defined as follows:

$$\begin{aligned}\mathcal{L}_{VAE} &= -\mathbb{E}_{p(h|x)}\Big[\log \frac{q(x|h)q(h)}{p(h|x)}\Big] \\ &= -\mathbb{E}_{p(h|x)}\big[\log q(x|h)\big] + \mathcal{D}_{KL}\big(p(h|x)\,||\,p(h)\big)\end{aligned} \tag{2}$$

where $\mathcal{D}_{KL}$ denotes the KL divergence, and the log-likelihood $\log q(x|h)$ is converted into its lower bound by variational inference for efficient solution.

Similar to VAE, GAN also contains two sub-models, i.e., a generator G and a discriminator D. G is optimized for fitting the distribution of real data, which is realized via generating fake images to confuse D. Meanwhile, D aims to distinguish the fake data from the real one, which acts as a binary classifier for conveying informative signals to G. The objective of GAN can be defined by binary cross entropy as follows:

$$\mathcal{L}_{GAN} = log\big(D(x)\big) + \log\big(1 - D(G(z))\big) \tag{3}$$

where z is the prior noise fed into G.

B. Context-aware Conditional VAE

Conditional variational autoencoder is one of the most important variant of VAE, which is adopted in this paper as backbone model of our first component. In order to capture difference between background and foreground of images for more effective text-image alignment, we propose a context-aware conditional VAE (CVAE). On the one hand, it captures key layout and color of images for subsequent refinement. On

[1]It is noted that letter “C” in “CVAE” denotes “Context-aware” instead of “Conditional”.

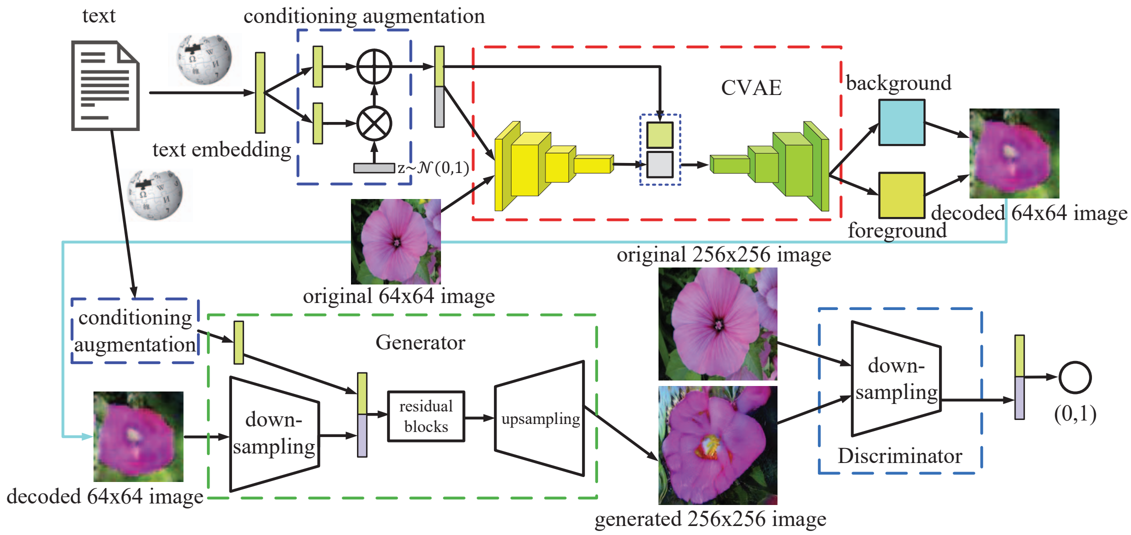


Fig. 1. An overview of our proposed framework, where the context-aware VAE (CVAE) captures key layout and color of images, and the conditional GAN aims to refine the output of CVAE for high-quality image generation.

the other hand, it counters the noisy nature in both image and text domain, for building the stable matching relations between these two heterogeneous domains.

Specifically, we provide text condition not only for the encoder but also for the decoder of VAE. Given text condition $c \in \mathbb{R}^{N_c}$ and noise prior $z \in \mathbb{R}^{N_z}$, CVAE aims to maximize log-likelihood $\log(x|c,z)$ such that plausible image x can be generated. The variational lower bound of the conditional log-likelihood can be defined as follows:

$$\begin{aligned}\mathcal{L}_{CVAE} = &-\mathcal{D}_{KL}(q(x|c,z)||q(z)) \\ &+ \mathbb{E}_{q(z|x,c)}[\log(p(x|c,z))]\end{aligned} \tag{4}$$

The KL divergence $\mathcal{D}_{KL}$ in (4) is a regularization for drawing the distributions of z and $q(z|x,c)$, which forces the encoder to learn robust latent distribution during training. As for the log-likelihood of image (i.e. the second term in (4)), its variational lower bound is commonly converted into the reconstruction loss. However, we argue that ℓ_2-norm cannot really measure the similarity of natural images, and we replace the reconstruction loss via adversarial learning. Specifically, the image feature inferred by CVAE's decoder $De(x)$ is utilized to learn an auxiliary Gaussian model with identity covariance:

$$p(x|c,z) = \mathcal{N}\big(D(x)|D(De(x)), I\big) \tag{5}$$

where D denotes the discriminator for adversarial learning.

Object images are composed of background and foreground, and they describe unbalanced correspondence to text description. Inspired by this observation, we propose to capture text-to-image matching relations via decoupling the background and foreground for realistic image generation. Following [4], we decouple images with a matting equation:

$$x = x_f \odot (1-g) + x_b \odot g \tag{6}$$

where x denotes the image data, $\odot$ denotes the element-wise product, $g \in [0,1]^{N_x}$ is a gating function which performs as a mask for deciding the visibility of background pixels, and $(1-g)$ correspondingly decides the visibility of foreground pixels. In practice, we realize the gating function as a single layer in CVAE's decoder, aiming to learn different features w.r.t. background and foreground for finer text-image alignment.

Compared to existing methods that generate images in a holistic manner, our decoupling approach has two main advantages: (1) It forces the model to pay different attention on different parts of images, which captures more fine-grained information for high-quality image generation. (2) It stabilizes the training process for more robust text-image alignment, since the noise caused by background are reduced significantly.

C. Refinement Process via Conditional GAN

Images generated by our first component cannot meet the requirement of high-quality generation. First, they are commonly blurred, with only key layout and color that correspond roughly to text description. Second, as shown in Fig. 1, the resolutions of them are quite low, i.e., they have only 64×64 size. Third, various details are omitted in these images, since capturing fine-grained information from high-level text and fitting image data distribution simultaneously is intractable for single VAE model.

In order to address these problems, we adopt a conditional GAN topped uppon the CVAE, which refines the output of CVAE for both high-resolution (i.e. 256×256) and realistic image generation. Specifically, the GAN is provided with both the low-resolution images in the CVAE stage and the corresponding text descriptions as conditions, while the noise prior z is omitted as we believe the randomness has been stored in the outputs of CVAE.

Moreover, we adopt conditioning augmentation techanique method proposed by [6] for text embedding smoothing. As stated in [6], the dimension of text embeddings is commonly high, while the text data is quite rare, leading to the manifold of text embedding is discontinuous. The conditioning augmentation method aims to produce additional conditioning variables, which is sampled from an independent Gaussian distribution $\mathcal{N}\big(\mu(\phi(c)), \sum\phi(c)\big)$, where $\phi(c)$ is the embedding of text c, μ and $\sum$ denote the mean and the diagonal covariance matrix of the Gaussian distribution respectively, and both of them are functions of c. By sampling from such Gaussian distribution randomly, we can obtain more text conditions for supplementing text embeddings' manifold, so that the text-image alignment can be more smooth.

As shown in Fig. 1, the generator G is fed into the low-supervision image $\hat{x} = CVAE(x, \phi(c), z)$ and corresponding text embedding $\phi(c)$, whose down-sampling modules extract the feature of $\hat{x}$ with convolutional architectures. Then the image feature is concatenated with $\phi(c)$, which feeds forward to residual blocks and upsampling modules. The adversarial objective of the conditional GAN can be defined as follows:

$$\begin{aligned}\mathcal{L}_D = &\mathbb{E}_{(x,c)\sim p_{real}} \log\big[D\big(x, \phi(c)\big)\big] \\ &+ \mathbb{E}_{\hat{x}\sim p_{CVAE}, c\sim p_{real}}\big[\log\big(1 - D(G(\hat{x}, \phi(c))), \phi(c)\big)\big]\end{aligned} \tag{7}$$

$$\mathcal{L}_G = \mathbb{E}_{\hat{x}\sim p_{CVAE}, c\sim p_{real}}\big[\log\big(1 - D(G(\hat{x}, \phi(c))), \phi(c)\big)\big] \tag{8}$$

It is noted that although we use c to denote text condition in both the CVAE stage and the GAN stage, the instances in two stages are different, since they are randomly sampled from Gaussian distribution stated above. Such setting encourages that sub-models in two stages can capture complementary text-image relations, ensuring a robust text-to-image generation.

IV. Experiments

A. Datasets

To verify the effectiveness of our proposed approach, we conduct experiments on 2 widely-used datasets, namely CUB [17] of bird images and Oxford-102 [18] of flower images. CUB contains 11788 bird images of 200 categories, and Oxford-102 has 8189 flower images of 102 categories. 10 captions for each image in both datasets provided by Reed et al. [2] are used in this paper. We follow the zero-settings in [3] and [6] for fair comparison. Namely, we split CUB with 150 training categories and 50 testing categories, and split Oxford-102 with 82 training categories and 20 testing categories.

B. Mode Architecture

Our CVAE component contains an encoder En, a decoder De and an auxiliary discriminator D_a. En adopts 5×5 convolution with stride 2 to extract image features, followed with three fully-connected (FC) layers whose output dimension is 2048. In De, we use 5×5 fractional striding to upscale the images, where striding operation realizes upsampling via transposing the convolution direction. It is noted that all the convolutional layers in En and De are followed by batch normalization [19] and ReLU activation, expect the last layer of De is `tanh` activation without batch normalization. As for D_a, we use four 5×5 convolution layers followed by batch normalization and ReLU activation, and two FC layers topped with sigmoid function for probability estimation.

In our GAN component, the generator G mainly contains three modules, namely down-sampling module, residual module and upsampling module. The down-sampling module uses 4×4 stride 2 convolutions, and all of them are followed by batch normalization and LeakyReLU activation. The residual module consists of two residual blocks, where 3×3 stride 1 convolutions are used. The upsampling module adopts bilinear upsampling followed by 3×3 stride 1 convolution. Similarly, the convolution layers in residual and upsampling module are followed by batch normalization and ReLU activation.

C. Implementation details

We directly use the text embedding that is released by [2]. We sample noise from Gaussian distribution and the dimension of noise vector is 100. For training, we firstly train the VAE for 400 epochs. Once the training procedure is done, the encoder is discarded and only the decoder is involved in the GAN component, and they are trained end-to-end for text-to-image generation. Specifically, we train GAN component for 500 epochs. All the training processes are powered by Adam optimizer [20] with an initial learning rate 0.0002, and we decay it every 50 epochs by a factor of 0.2.

D. Experimental Results

In order to evaluate our approach comprehensively, we conduct both quantitative and qualitative experiments for comparing our approach with two state-of-the-art text-to-image generation methods, namely StackGAN [6] and GAN-INT-CLS [3], and the details are elaborated as below.

1) Quantitative experiment: Similar to [6], [7], we adopt inception score [7] as quantitative evaluation metric. Inception score is a recently proposed numerical assessment metric for evaluating generative models [6], [7], [21]–[23], which resorts to recognition model for scoring the quality of the generated images. It can be defined as follows:

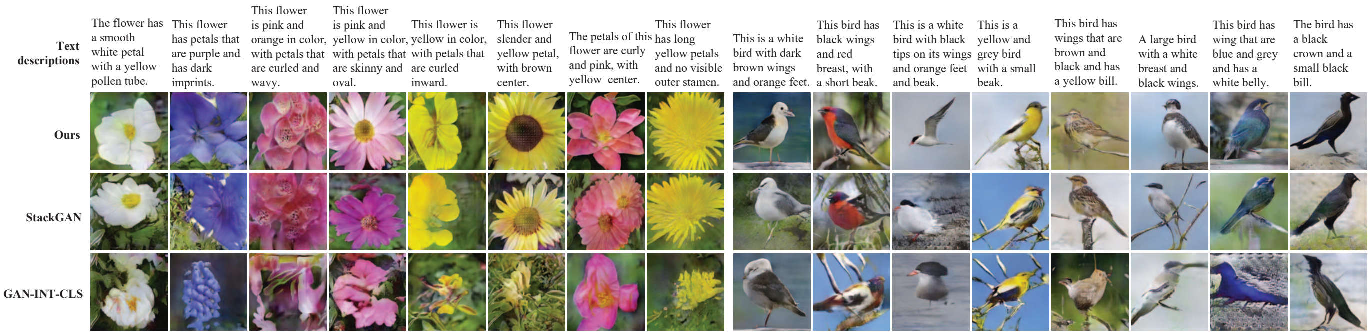
$$I = exp\Big(\mathbb{E}_x\mathcal{D}_{KL}\big(p(y|x)||p(y)\big)\Big) \tag{9}$$

where x denotes the generated images and y is the category label predicted by Inception model [24]. The key idea behind this metric is that good generative model can generate diverse but meaningful images, and thus the KL divergence between the marginal distribution $p(y)$ and the conditional distribution $p(y|x)$ should be large. Therefore, a higher inception score denotes a better generation performance.

As shown in Table I, our approach outperforms the two compared methods on Oxford-102, while achieves comparable results to them on the CUB. Furthermore, the variance of our approach is smaller than two compared methods, demonstrating that our approach is more stable than them.

TABLE I
Inception scores of our approach and compared methods.

Dataset	Method	Inception score
CUB	Ours	4.97 ± 0.03
	StackGAN	4.95 ± 0.04
	GAN-INT-CLS	**5.08 ± 0.08**
Oxford-102	Ours	**4.21 ± 0.06**
	StackGAN	4.17 ± 0.07
	GAN-INT-CLS	3.54 ± 0.07



(a) Oxford-102 (b) CUB

Fig. 2. Some image examples generated by our proposed approach and compared methods conditioned on text descriptions from testing set of two datasets.

2) Qualitative experiment: Fig. 2 is qualitative results generated by our approach and two compared methods. We randomly select text descriptions from test categories and use them as conditions for all models. Our observations are as follows. Although GAN-INT-CLS use correlation constraints for reducing the heterogeneity between images and text, images generated by this method are unrealistic, which lost some fine-grained visual details, and some examples do not even match the text descriptions (e.g., the second and sixth column in Fig. 2 (a)). StackGAN can generate more realistic images with better text-image alignment. However, when the background is complex, the objects are easily disturbed by background (e.g., the second column of Fig. 2 (b)). Our proposed approach can generate more vivid images with fine-grained visual parts, as well as is more robust for complex background, due to the advantages of our context-aware mechanism.

V. Conclusion

In this paper, we have proposed a context-aware and stacked framework for high-quality text-to-image generation, which contains a context-aware conditional VAE and a conditional GAN. The context-aware VAE decouples background and foreground of images, forcing the VAE to pay different attention on different parts, and conditional GAN is adopted for refining the generation of VAE, which improves the quality of the generated images significantly. Experimental results on 2 datasets verify the effectiveness of our proposed approach.

As for the future work, we will focus on exploring more efficient training algorithms of VAE and GAN, to improve the computational efficiency of our model.

Acknowledgment

This work was supported by National Natural Science Foundation of China under Grant 61771025 and Grant 61532005.

References

[1] Karpathy, A., & Fei-Fei, L. "Deep visual-semantic alignments for generating image descriptions". In CVPR, 2015.

[2] Reed, S., Akata, Z., Lee, H., & Schiele, B. "Learning deep representations of fine-grained visual descriptions". In CVPR, 2016.

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

[4] Porter, T., & Duff, T. "Compositing digital images". In ACM Siggraph Computer Graphics (Vol. 18, No. 3, pp. 253-259).

[5] Radford, A., Metz, L., & Chintala, S. (2015). "Unsupervised representation learning with deep convolutional generative adversarial networks". arXiv preprint arXiv:1511.06434.

[6] Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., & Metaxas, D. "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks". In ICCV, 2017.

[7] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. "Improved techniques for training gans". In NIPS, 2016.

[8] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. "Generative adversarial nets". In NIPS, 2014.

[9] Mirza, M., & Osindero, S. (2014). "Conditional generative adversarial nets". arXiv preprint arXiv:1411.1784.

[10] Arjovsky, M., Chintala, S., & Bottou, L. (2017). "Wasserstein gan". arXiv preprint arXiv:1701.07875.

[11] H. Dong, J. Zhang, D. McIlwraith, and Y. Guo. "I2t2i: Learning text to image synthesis with textual data augmentation". In ICIP, 2017.

[12] E. Mansimov, E. Parisotto, L. J. Ba, and R. Salakhutdinov. "Generating images from captions with attention". In ICLR, 2016.

[13] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Z. Afzal, and M. Liwicki. "TAC-GAN-Text Conditioned Auxiliary Classifier Generative Adversarial Network". arXiv preprint arXiv:1703.06412, 2017.

[14] X. Wang and A. Gupta. "Generative image modeling using style and structure adversarial networks". In ECCV, 2016.

[15] D. P. Kingma and M. Welling. "Auto-encoding variational bayes". In ICLR, 2014.

[16] D. J. Rezende, S. Mohamed, and D. Wierstra. "Stochastic backpropagation and approximate inference in deep generative models". In ICML, 2014.

[17] 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.

[18] M.-E. Nilsback and A. Zisserman. "Automated flower classification over a large number of classes". In ICCVGIP, 2008.

[19] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

[20] D. Kinga and J. B. Adam, "A method for stochastic optimization", in ICLR, 2015.

[21] David Berthelot, Thomas Schumm, and Luke Metz. 2017. "BEGAN: Boundary equilibrium generative adversarial networks". arXiv preprint arXiv:1703.10717 (2017).

[22] Xun Huang, Yixuan Li, Omid Poursaeed, John E. Hopcroft, and Serge J. Belongie. "Stacked generative adversarial networks". In CVPR, 2017.

[23] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. "Pose guided person image generation". In NIPS, 2017.

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. "Rethinking the inception architecture for computer vision". In CVPR, 2016.

[25] M. Arjovsky and L. Bottou. "Towards principled methods for training generative adversarial networks". In ICLR, 2017.

[26] C. K. Snderby, J. Caballero, L. Theis, W. Shi, and F. Huszar. "Amortised map inference for image super-resolution". In ICLR, 2017.

[27] Bao, Jianmin and Chen, Dong and Wen, Fang and Li, Houqiang and Hua, Gang. "CVAE-GAN: Fine-Grained Image Generation through Asymmetric Training". In ICCV, 2017.