Text-To-Image and Image-To-Image Synthesis With Generative Adversarial Networks

Arda Tümay
Bilkent University
arda.tumay@bilkent.edu.tr

Naci Dalkıran
Bilkent University
naci.dalkiran@ug.bilkent.edu.tr

Tolga Çatalpınar
Bilkent University
tolga.catalpinar@ug.bilkent.edu.tr

Abstract

Text-to-image synthesis can be easily described as generating an image based on a given text description. The main difficulty involved in this process, however, is making the generated image to convey the meaning of the text description both semantically and visually. There are extensive works in the literature about the field of text-to-image synthesis that successfully produce an image with the same meaning of text description. The major domain of generated images are generally flowers and birds. Generating images from different domains such as faces based on text description, however, is rarely studied. In this paper, we concentrate on generating various images from different domains such as faces or classrooms.

1. Introduction

Text-to-image generation is a technique to produce a non-existing image from a given text that describes the attributes of the image. It may also be seen as the reverse of the image captioning where a sentence that describes the image content is produced. Text-to-image synthesis can be adapted to different domains ranging from creating real-looking bird images to flower images. It is kind of a multimodal learning where the data distribution is learnt from both text descriptions and images. In the literature, there are various types of neural networks to accomplish this task. As we did literature search, the most suitable technique for text-to-image-synthesis is using Generative Adversarial Networks that is introduced by GoodFellow et. al. in 2014 [1]. GANs are also very suitable for multimodal learning tasks because of its nature that is going to be discussed in the next sections. Besides GANs, other network types such as Recurrent Neural Networks(RNN) and Convolutional Neural Networks(CNN) might also be used, however, these networks are not suitable for such tasks due to their nature. As we did literature search, it is revealed that RNNs are most suitable for sequential data where we can talk about timestamps within the data. Possible application areas of RNNs may be speech recognition, machine translation or image captioning. Furthermore, CNNs are coming forward for learning tasks applied on images. However, CNNs are only able to create a single image with a given input according to learnt data distribution which performs badly when it comes to creating non-existing images from learnt distribution. Because there is not any criteria for generated images to measure whether images are real looking or not. Since Generative Adversarial Networks comprise two different networks that are discriminator and generator, there is a chance for the generator to improve itself for creating better looking images in accordance with the discriminator.

In our work, we aimed to produce non-existing classroom, bird and face images by using two different networks. First network is a state-of-the-art generative network which is named as StackGAN-v2 [2]. We have used StackGan to produce classroom and bird images and by implementing changes in the network's architecture. Another network is a convolutional network that is created for image creation which is named as MobileNet[3]. Different from GANs and text-to-image synthesis, we wanted to observe how convolutional neural networks perform at creating images. That's why we choose MobileNet network to train and observe its effectiveness for image creation.

We have used MobileNet for creating face images and StackGan for creating bird and classroom images. We have used 3 different datasets. For face creation, we have used the CelebA [4] dataset. It consists of more than 200k celebrity images with 5 landmark locations and 40 binary attributes. There are more than 10k identities within images. For classroom creation, we have used the LSUN dataset. It includes millions of images for 10 scenes and 20 objects. Lastly, we have used Caltech-UCSD Birds 2000(CUB) [5] dataset for bird creations. This dataset includes more than 6000 images of 200 bird species with annotations and bounding boxes.

This report will contain several sections. In the related work section, we will be presenting literature about text-to-image synthesis. Then in the implementation section, we will mention implementation details. Then we will mention generative adversarial networks. After that, we will finish with results and conclusion.

2. Related Work

Generating real looking fake images that humans cannot differentiate from real images is a demanding research area of computer vision. Text-to-image synthesis is one of the subcategories that relies on generating real looking fake images based on natural text descriptions. As nature of the problem brings, it is a multimodal learning challenge which means there may be numerous different images synthesized from the same text description and real looking images.. With the advances in the field of deep learning and generative adversarial network, it becomes more convenient to produce such images. GANs as proposed by Goodfellow et. al. [1] are very suitable for such multimodal learning problems.

Reed et. al. proposed one of the initial architectures for generating images based on text description by using CNN and GAN. Unlike traditional encoders, their architectures condition network on the text descriptions rather than class labels. As a result, they achieved to produce 64x64 real looking images. They utilize CUB bird and Oxford-102 flower dataset for training. Their architecture is a single end-to-end trainable network with GAN and CNN.

Zhang et. al. has come with the idea of generating high resolution, 256x256 images based on text descriptions and created a new model that consists of two stacked GANs. StackGAN-v1 has offered a model that can achieve producing high resolution images by dividing the problem into two sub-problems. The first stage generates an image with the traditional way, by the conditional tasks from the given texts. Then the second stage focuses on correcting the defects and increasing the resolution of the image that is generated in the first stage. In that novelty, StackGAN-v1 is the first GAN that can generate high resolution images, compared to the other GANs. The COCO and CUB datasets are used in training StackGAN.

Then, AttnGAN model was proposed that outperforms the state-of-art GAN models at the time. The model executes with fine-grained text-to-image generation. Its difference from the other GANs is that it synthesizes fine-grained details on the parts of the images. This attentional generative network promotes the best reported inception scores by 14.14% on CUB dataset and by 170.25% on the more challenging COCO dataset. AttnGAN is the first instance of layered attentional GAN that can automatically select the condition at the word level for synthesizing the desired attributes in the parts of the generated images.

After coming up with StackGAN-v1[6], researchers had produced a more successful GAN than StackGAN-v1, which is, StackGAN-v2. Whereas StackGAN-v1 works on the conditional parameter only, StackGAN-v2 operates on both conditional and unconditional parameters. It achieves generating images with higher quality and in a more stable way. It uses a GAN with multiple generators and discriminators that are placed in a tree-like structure. In this way, it "jointly approximates multiple related distributions, including multi-scale image distributions and jointly conditional and unconditional image distributions." [6] StackGAN-v2 uses COCO, CUB and Oxford-102 datasets for utilizing its conditional tasks, and, ImageNET and LSUN for utilizing its unconditional tasks. Similar to the StackGAN-v1, this improved version of it outperforms the similar GANs in the field.

Similar to AttnGan[**7**], MirrorGan is another architecture created for text-to-image synthesis that utilizes the attention module [8] . They approach the problem as an inverse process of text-to-image synthesis by stating an image that is generated based on the text description should also generate the same text redescription that has the same semantic meaning with the initial description. Based on this idea, they develop a novel solution with three different modules which are a semantic text embedding module, a global-local collaborative attentive module and semantic text regeneration and alignment module. They also used sentence level text embeddings as well as word level embeddings to reinforce the semantic meaning of the generated image. They utilize RNN to create word embeddings, generative adversarial networks to generate images and CNN and LSTM to create redescription. They experiment on CUB bird and MS COCO dataset. They utilize inception score, R-precision as quantitative metrics and human perceptual test as qualitative metrics.

There are successful solutions for generating high-resolution images that convey the semantic meaning of text description. Whereas, none of them provides control over the location of objects in the generated image. Motivated by this fact, Reed et. al. [9] comes up with a new architecture named Generative Adversarial What-Where

Network (GAWWN) that draws an image based on a text description that defines the location of objects as well as objects to be drawn. Their architecture paves the way for text- and location-controllable text to image synthesis with the help of spatial masking and bounding boxes for the coarse location of the object and set normalized coordinates of keypoints for the object parts. They benefit from CNN and RNN to encode text and learn a function that maps between images and text features. Similar to AttnGan, GAWWN also utilizes sentence-level word embeddings. For experiments, they use the CUB bird dataset and MPII Human Pose dataset.

Image-Text-Image (I2T2I) was proposed as a model to integrate image-to-text and text-to-image synthesis by Zhang *et al.* [10]. I2T2I provides opportunities to learn more textual descriptions of each image to enhance text-to-image synthesis by implementing textual data augmentation [10]. Its image captioning model, one of its three models, enables description to be augmented [10]. MSCOCO and MPII Human Pose (MHP) datasets are used in the experiment.

High-Definition and Hierarchically-nested Discriminators Generative Adversarial Network (HDGAN) was proposed to cope with the obstacles of generating photographic images by Dong *et al.* [11]. Their method regularizes mid-level representations and helps the generator training to capture the complex image statistics inside the network hierarchies [10]. Also, to enhance semantic consistency and image fidelity at the same time, a multi-purpose adversarial loss is adopted [11]. Moreover, to interpret the semantic consistency of the generated image, a new visual -semantic similarity measure is introduced [11]. They experiment on CUB birds, Oxford-102 flowers, and MS COCO dataset. To evaluate the method, three quantitative metrics are used which are Inception score to measure objectiveness and diversity of generated image, Multi-scale structural similarity metric for further validation, and visual semantic similarity introduced by them. Vanilia GAN, without additional class label supervision and requiring multi-stage training, is used as a generator. Also, to aid accompanying discriminators more effectively, new CNN is proposed for the generator. Consecutive stride-2 convolutional layer with LeaklyReLU and BN is utilized in the discriminator.
Dong *et al.* present a new method to manipulate a person image's visual appearance according to natural language descriptions. The method performs two-stage. In the first stage, the description is guided pose generation. In the second stage, a person's visual appearance is transferred to image synthesis. In the experiment, CUHK-PEDES dataset is used. To evaluate attribute transfer correctness, VQR perceptual score is proposed as a novel metric by them. Also, In the experiment, GAN-based pose inference network proposed for the first time to generate human pose with a given description is used as model architecture.

3. Implementation

In this section, implementation details will be discussed in detail. However, before moving to the implementation process, we would like to mention an important point about our project. First, as we explained in the introduction section, our aim was doing a text-to-image synthesis with different datasets and different architectures. As we complete our literature search, we learnt that generative adversarial networks are best suited to create non-existing real-looking images from real images that are conditioned on corresponding text descriptions. As we dive deep into the process of conditioning the training process to a text description, we see that it is not the text description itself that the network condition on but the text embedding that is created from text description. We then started to implement a network that created text embeddings, however, we failed at creating such a network since we have insufficient knowledge about it. Then, we decided to find a network to create text embeddings, we found a network that is supposed to create embeddings from sentences but this time we were never able to make the network run and give text embeddings. This point is where we spent much of the time on. From this point, we have decided to do image-to-image synthesis. For this aim, we again used state-of-the-art model StackGan and MobileNet networks as we mentioned in the introduction section.

4. Generative Adversarial Networks

GAN's consists of two networks: generator and discriminator. Both networks are simply different kind neural networks.

Generator is supplied with random noise input and expected to generate an image from noise input. With this way, the generator basically tries to synthesize non-existing real-looking images to "fool" the discriminator.

Then the discriminator is fed with two images which are real image sampled from the training set and fake image that is produced by the generator network. Discriminator is trained to make correct differentiation between real image and fake image. In each iteration, the generator is trained so it produces more accurate results to fool the discriminator. At the end, this process becomes a two player minmax game with two different loss functions for each of the network, generator and discriminator. The rough architecture of a gan and its loss functions are given below.

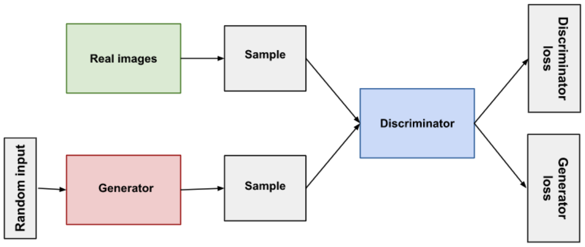

Figure 1: Architecture of a simple generative adversarial network.

$$\max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

Figure 2: Loss function for discriminator.

As it is seen from figure 2, discriminator tries to maximize probability for real images and minimize probability for fake images from generator.

$$\min_{\theta_g} \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z)))$$

Figure 3: Loss function for generator.

As it is seen from figure 2, the generator tries to minimize the likelihood of the discriminator being correct. However, this loss function does not work well in reality since gradient signal starts too low that causes vanishing gradient problem. Instead, generator tries to maximize likelihood of discriminator being wrong that corresponds to a loss function in figure 4 which gives a high gradient signal at the beginning of training.

$$\max_{\theta_g} \mathbb{E}_{z \sim p(z)} \log(D_{\theta_d}(G_{\theta_g}(z)))$$

Figure 4: Actual loss function of generator.

5. Methodology

As we mentioned before, since we cannot accomplish text-to-image synthesis, we perform image-to-image synthesis with StackGan-v2[2] and MobileNet[3]. We have used these two networks to train with our new datasets and to do modification in the network. In this section, we will be explaining implementation details of these architectures in detail.

StackGAN-v2 is a state-of-the-art model [2] that performs both conditional and unconditional image synthesis. Unlike similar GANs that consist of one layer. StackGan-v2 includes 2 different stacked GANS. By this way, it decomposes a difficult problem of image synthesis into two more manageable subproblems.

Its layers' names are stage-1 GAN and stage-2 GAN. It conveys very similar properties from StackGan-v1 [6]. Conditioning augmentation layer at the beginning of the stage-1 and stage-2 networks are presented in StackGan-v1 and StackGan-v2. Aim of the conditioning augmentation layer is yielding more training pairs given a small number image text pairs which results in robustness and eliminates discontinuity in the latent conditioning manifold. With conditioning augmentation technique, the network produces additional conditioning variables in contrast to fixed conditioning text variable. Additional conditioning variables are randomly sampled from a gaussian distribution where the mean and covariance matrix are the functions of the text embedding. Moreover both versions of the network introduces a new regularization term which is Kullback-Liebler divergence between a standard gaussian distribution and conditioning gaussian distribution. By this way, network enforce smoothness over conditioning manifold and prevent overfitting.

Both StackGan-v1 and StackGan-v2 consist of two layers that are mentioned above. Instead of directly generating high-resolution images, stage-1 GAN creates low-resolutions images with rough shapes, background layer and correct colors for object. In the stage-1 GAN, generator is fed with two inputs. First one is random noise input and second one is text conditioning variable that is obtained from text embedding of text description. Then noise input image and conditioning variable are concatenated along network.

By getting low-resolution image from generator and conditioning variables from conditioning augmentation layers, stage-2 GAN produces high-resolution image with vivid object parts. Both stage-1 and stage-2 are conditioned on text variables. While in stage-1 GAN convolutional and upsampling layers are used, in stage-2 GAN residual layers are also utilized to learn multimodal representations across image and text features as well as convolutional and upsampling layers. Architecture of StackGan-v1 is showed in figure 5.

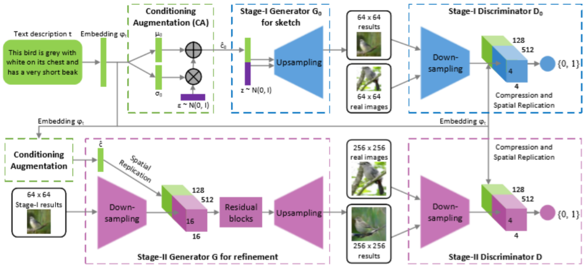

Figure 5: Architecture of StackGan-v1

Then on top of the StackGan-v1, StackGan-v2 introduces unconditional image synthesis using stacked generative adversarial networks. Moreover, in contrast to StackGan-v1, in StackGan-v2 multiple generators and discriminators are introduced in a tree-like structure to model a series of multi-scale image distributions.

Images from low-resolution to high-resolution are generated from different branches of the tree. At each branch, the generator captures the image distribution at that scale and the discriminator estimates the probability that a sample came from training images of that scale rather than the generator.

The motivation of StackGan-v2 is that by modeling data distributions at multiple scales, if any one of those model distributions shares support with the real data distribution at that scale, the overlap could provide good gradient signal to expedite or stabilize training of the whole network at multiple scales. Architecture of StackGan-v2 is showed in figure 6.

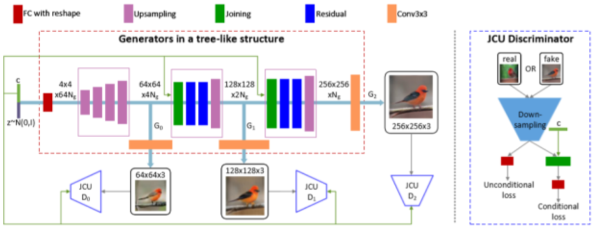

Figure 6: Architecture of StackGan-v2

MobileNet [3] is an architecture that is useful for mobile and embedded based vision implementations. MobileNet-v2 [3] is the augmented version of it. MobileNet simply operates depthwise convolution followed by pointwise convolution. That means, the convolutional kernels are applied to every input channel, which then reduces the number of parameters. Then, pointwise convolution applies a 1x1 convolution over the outputs of the depthwise convolution to combine them.

6. Experiments

We conducted various experiments with StachGan-v2. First, we used a CUB bird dataset to train StackGan-v2. Then we used the LSUN classroom dataset to train StackGan-v2. We also made changes in the architecture of the StackGan-v2 by adding more convolutional layers to upsampling and residual blocks, however, results are not so much satisfactory as it will be seen in the results section. Lastly, we have used MobileNet to produce faces based on CelebA face dataset.

7. Results

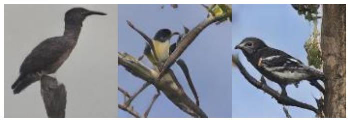

Figure 7: StackGan-v2 PPPCaltech-UCSD Birds 200 dataset results from left: Black Footed Albatross, Groove Billed Ani, Least Auklet.

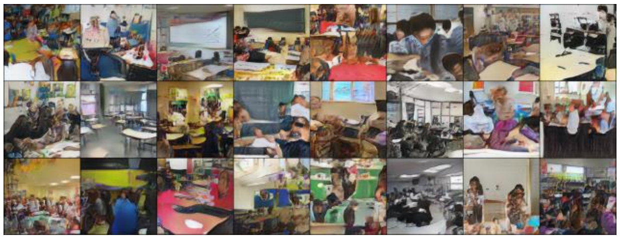

Figure 8: StackGan-v2 Lsun-classroom dataset results in 25th epoch.

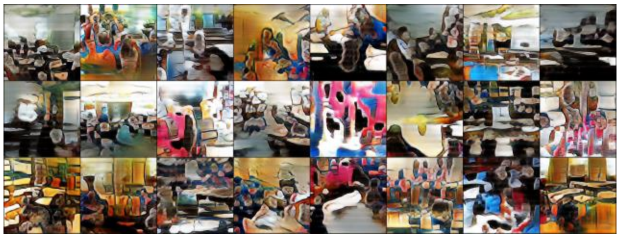

Figure 9: UretGan (first trial) Lsun-classroom dataset results in 9th epoch. 3 more convolutional layers in upsampling blocks.

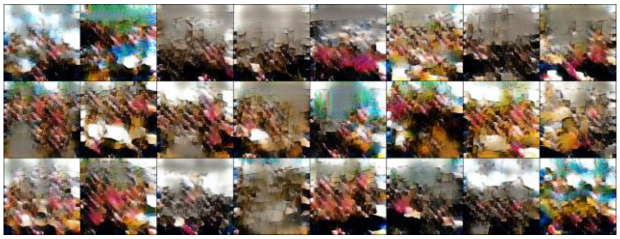

Figure 10: UretGan (second trial) Lsun-classroom dataset results in second epoch. 3 more convolutional layers in upsampling blocks and residual blocks.

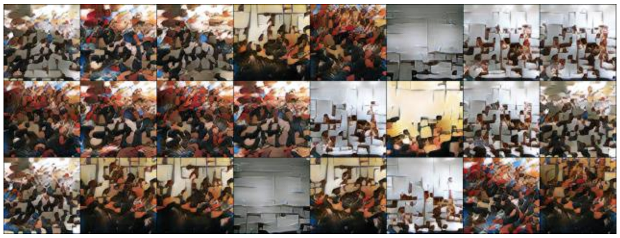

Figure 11: UretGan (third trial) Lsun-classroom dataset results in second epoch. 3 more convolutional layers in upsampling blocks and residual blocks.

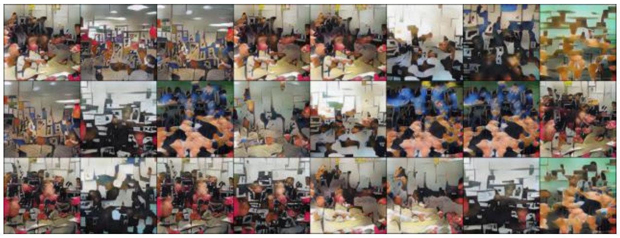

Figure 12: UretGan (fourth trial) Lsun-classroom dataset results in 11th epoch. 3 more convolutional layers in upsampling blocks and residual blocks.

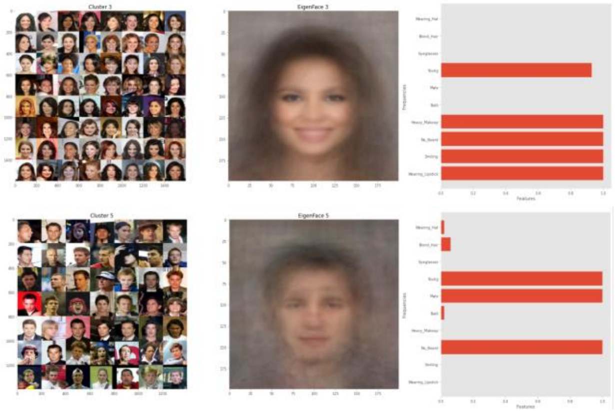

Figure 13: MobileNet CelebA dataset results

8. Conclusion

We developed a model based on StackGAN-v2 architecture. We trained this model with the LSUN Classroom Dataset and compared it with StackGAN-v2's original architecture. We expected to get more accurate results, however, StackGAN-v2's architecture seemed to give better results. We have restricted our results with small numbers of epochs, since training requires so many times. We also trained MobileNet-v2 with CelebA dataset to test our model's performance. From the conducted experiments, it can easily be seen that both our model and StackGAN-v2 generates better images with high resolutions. Throughout the project we have learnt so much knowledge about generative adversarial networks, text-to-image and image-to-image synthesis. Project was very beneficial for us to enlarge our vision in the field of deep learning and generative adversarial networks.

Our github link:
https://github.com/tolgacatalpinar/UretGAN

References

[1] S. A. Israel et al., "Generative Adversarial Networks for Classification," 2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington, DC, 2017, pp. 1-4, doi: 10.1109/AIPR.2017.8457952.

[2] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1947–1962, 2019.

[3] Howard, Andrew & Zhu, Menglong & Chen, Bo & Kalenichenko, Dmitry & Wang, Weijun & Weyand, Tobias & Andreetto, Marco & Adam, Hartwig. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.

[4] *Large-scale CelebFaces Attributes (CelebA) Dataset*. [Online]. Available: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html. [Accessed: 10-Jun-2020].

[5] “UCSD Birds 200,” *Caltech*. [Online]. Available: http://www.vision.caltech.edu/visipedia/CUB-200.html. [Accessed: 10-Jun-2020].

[6] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, “StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks,” *2017 IEEE International Conference on Computer Vision (ICCV),* 2017.

[7] 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,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[8] T. Qiao, J. Zhang, D. Xu, and D. Tao, “MirrorGAN: Learning Text-To-Image Generation by Redescription,” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, H. Lee, “Learning What and Where to Draw” *2016 IEEE International Conference on Computer Vision (ICCV),* 8 Oct 2016.

[10] H. Dong, J. Zhang, D. Mcilwraith, and Y. Guo, “I2T2I: Learning text to image synthesis with textual data augmentation,” *2017 IEEE International Conference on Image Processing (ICIP),* 2017.

[11] Z. Zhang, Y. Xie, and L. Yang, “Photographic Text-to-Image Synthesis with a Hierarchically-Nested Adversarial Network,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.