Enhancing DF-GAN for Text-to-Image Synthesis: Improved Text-Encoding and Network Structure

Yixuan Wu[1], **Zhaonan Zhou**[2,3,*]

[1]School of Statistics and Mathematics, Shandong University of Finance and Economics, Shandong, China
[2]School of Electronic Engineering and Computer Science, Queen Marry University of London, London, United Kingdom

[3]zhaonan.zhou@hss23.qmul.ac.uk
*corresponding author

Abstract. Text-to-image synthesis is one of the most challenging and popular tasks in machine learning, with many models developed to improve performance in this area. Deep Fusion Generative Adversarial Networks (DF-GAN) is a straightforward but efficient model for image generation, but it has three key limitations. First, it only supports sentence-level textual descriptions, restricting its ability to extract fine-grained features from word-level inputs. Second, the structure of the residual layers and blocks, along with key parameters, could be optimized for better performance. Third, existing evaluation metrics, such as Fréchet Inception Distance (FID), tend to place undue emphasis on irrelevant features like background, which is problematic when the focus is on generating specific objects. To address these issues, we introduced a new text encoder that enhances the model having capacity to process word-level descriptions, leading to more precise and text-consistent image generation. Additionally, we optimized key parameters and redesigned the convolutional and residual network structures, resulting in higher-quality images and reduced running time. Lastly, we proposed a new evaluation theory tailored to assess the quality of specific objects within the generated images. These improvements make the enhanced DF-GAN more effective in generating high-quality, text-aligned images efficiently.

Keywords: Text-to-Image, DF-GAN, Generative Adversarial Networks, Fréchet Inception Distance.

1. Introduction

Compared to words, pictures can convey information more intuitively and vividly, making it clear at a glance. Because of its potential applications, text-to-image synthesis has attracted a lot of attention in recent years. [1]. One of the most well-liked and often utilized deep learning modules for this task is Generative Adversarial Networks (GANs) [2]. The module was first introduced by an American named Ian Goodfellow in 2014 and has since become a popular method of generative tasks. Specifically, one of the most common uses of GANs is text-to-image creation. It focuses on generating images based on textual descriptions which is very challenging since there is a significant gap between the visual image features and the corresponding textual descriptions.

Despite significant advancements in previous works, challenges remain in the areas of image quality, word-level descriptions, and overfitting. This study aims to address these issues by synthesizing bird images that outperform the original Deep Fusion Generative Adversarial Networks (DF-GAN) baseline and overcome its limitations. The first major improvement is the introduction of a more effective text encoder. Unlike the original DF-GAN, which struggled with complex descriptions, our enhanced encoder breaks down descriptions into sentence vectors, effectively handling a wider range of inputs. This allows the model to generate images from descriptions composed of short, simple words, which was a challenge for the original DF-GAN. Secondly, we redesigned the generator network, incorporating Deep text-image Fusion Block (DFBlock), residual blocks, and additional network structures to address issues related to overfitting and inadequate sampling. Even with more complicated inputs, the capacity to generate structurally sound images of this model is enhanced by this change. Thirdly, by fine-tuning key parameters, such as the truncation rate of the input white noise, the generated images achieve a more realistic appearance compared to those from the original model. These adjustments not only enhance image quality but also accelerate the image generation process. Experimental results demonstrate that these enhancements—specifically the improved text encoder, redesigned generator network, and optimized parameters—significantly boost the quality of the generated images while also increasing generation speed. These improvements are particularly beneficial for applications that require reliable image samples in fields with limited image resources, such as medicine. Additionally, the faster generation speed and simplified network structure reduce the overall cost of synthesizing images.

2. Literature Review

In recent years, significant progress in fields like Machine Learning and Computer Vision has opened up exciting opportunities for generating images from natural language descriptions. This area of research has garnered growing attention from scholars due to its potential applications across various domains, such as content creation, virtual reality, and human-computer interaction. However, generating high-quality images from text descriptions remains an exceptionally challenging task. The complexity arises from the need to translate abstract, often ambiguous language into detailed and coherent visual representations that capture not only the correct objects but also fine details such as textures, lighting, and context. One of the most groundbreaking advancements in this area came with the introduction of GANs in 2014. GANs revolutionized image generation by employing a system of two neural networks—the generator and the discriminator—that compete against each other to improve the quality of the generated images. This adversarial framework led to the creation of highly realistic images and made GANs the go-to model for many researchers working in text-to-image generation. Over the years, multiple variants of GANs, such as BigGAN and StyleGAN, further enhanced image quality and resolution, pushing the boundaries of what was achievable in this domain.

However, despite the success of GANs, they have limitations, particularly when it comes to the diversity of generated samples and the stability of the training process. GANs are known to suffer from mode collapse, where the model produces limited variations of images, neglecting other possible representations. Moreover, training GANs is notoriously difficult, requiring careful tuning of hyperparameters and often leading to unstable results. In 2021, a new model emerged that shifted the landscape of image generation: Denoising Diffusion Probabilistic Models (DDPM). DDPMs operate on an entirely different principle than GANs. Rather than relying on an adversarial framework, DDPMs generate images by gradually denoising a random noise vector through a series of probabilistic steps. This diffusion-based approach allows DDPMs to generate high-quality images by capturing more of the underlying sample distribution, effectively overcoming some of the limitations faced by GANs. One of the most notable achievements of DDPMs was their ability to outperform BigGAN on the ImageNet task, a widely used benchmark for image generation models. This success has positioned DDPM as a formidable competitor to GANs, proving its superior ability to produce diverse and high-quality images while offering a more stable training process. As DDPMs continue to evolve, they present a promising alternative to GANs, especially in tasks requiring complex, high-resolution image generation from natural language descriptions.

Aiming to generate realistic and semantic-consistent images, researchers have developed a lot of methods such as GANs, Variational Autoencoders (VAEs) [3], DDPM [4], and many variants of GANs: Conditional GANs (CGANs) [5] which add label class as input to both generator and discriminator, so that the trained generator can output an image of a given class. The cross-modal attention mechanism created by Attentional GAN (AttnGAN) [6] enhances the ability to synthesize more detailed images of the generator. MirrorGAN [7] uses generated images to reproduce text descriptions to improve text-image semantic consistency. In cases where the first images formed in stacked architecture are not well-generated, Dynamic Memory GAN (DM-GAN) [8] presents a novel structure called the Memory Network to refine fuzzy image contents. Although DDPM seemed to be a superstar in generation models recently and generative model Sora is also based on DDPM, GANs still have space to improve and many researchers are devoting themselves to it. DF-GAN [9] is a recent model that uses AttnGAN as a text-encoder and is unlike a large amount of GANs models that adopt the stack architecture as the backbone, it uses a one-stage backbone. A one-stage text-to-image backbone can directly synthesize high-resolution images without entanglements between multiple generators, in contrast to the previous module. Matching-Aware Gradient Penalty (MA-GP) and One-Way Output were used to create a Target-Aware discriminator, which the creators of DF-GANs hoped would increase both the authenticity of the output images and semantic consistency. Additionally, they created a DFBlock to combine the information from the images with the text descriptions. In a word, DF-GAN is a variant of GANs on text-to-image tasks, but it is a simpler but effective model.

3. Methodology

3.1. Dataset description and preprocessing

In this research, The Caltech-UCSD Birds-200-2011 (CUB-200-2011) is what we utilize [10]. The dataset is an upgraded version of CUB-200 and has 200 different species of birds. In addition, it has nearly 60 high-resolution images for each species in distinct environment situations and angles. We trained and assessed the generated images of birds using this commonly used benchmark dataset for bird identification studies. By evaluating the coherence between the produced image and text description as well as the caliber of the birds, we are able to determine the quality of the created photographs. The dataset has undergone training.

3.2. Proposed approach

The research identified several limitations with DF-GAN, prompting key improvements to enhance its performance in generating high-quality images. The first issue was related to the text-encoder, which originally relied on AttnGAN. While AttnGAN performs adequately for sentence-level descriptions, it struggles with word-level details, limiting its ability to generate fine-grained image features. To address this, the text-encoder was replaced with a Transformer-based architecture, which is more adept at processing word-level descriptions, improving the model's ability to capture intricate details from text inputs. In addition to changing the text-encoder, several key parameters within the model were optimized to improve the quality of the generated images. By fine-tuning hyperparameters such as the learning rate, batch size, and the number of training iterations, the model was made more efficient at generating realistic images with improved resolution and visual fidelity. These adjustments were aimed at ensuring that the model could generate clearer and more accurate images from the given text descriptions.

Further, architectural modifications were made to the convolutional layers and residual blocks of the model. The convolutional layers were adjusted to enhance the feature extraction process, while the structure of the residual layers was redesigned to streamline the flow of information through the model. These changes increased the efficiency of the model, allowing it to generate higher-quality images in less time while maintaining or even enhancing the richness of details. Another critical improvement was related to the evaluation of the generated images. The original model used the Fréchet Inception Distance (FID) as the primary metric to assess image quality. However, FID tends to overemphasize irrelevant features, such as background details, leading to inaccurate assessments, particularly when the primary object in the image is well-generated but the background is less detailed. This resulted in higher FID scores for images where the background was blurred, even if the object itself was generated well. To overcome this limitation, the Fréchet Inception Distance for Objects (O-FID) was introduced. O-FID focuses more on the target object in the image and reduces the impact of irrelevant features like background, providing a more accurate evaluation of the quality of the generated object itself [11,12].

In summary, by addressing these issues—upgrading the text-encoder, fine-tuning key parameters, redesigning the model architecture, and improving the evaluation metric—the modified DF-GAN offers more precise and higher-quality image generation capabilities, particularly when dealing with detailed, word-level descriptions and object-focused image generation tasks. The pipeline of these improvements is outlined in Figure 1.

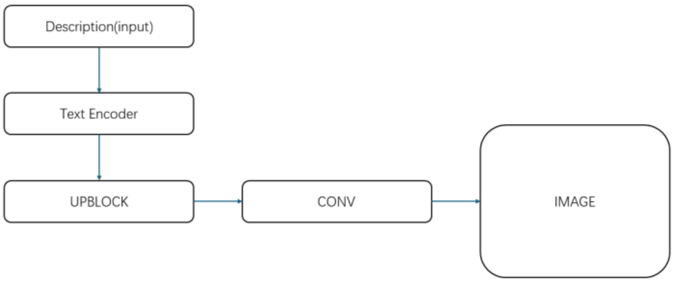
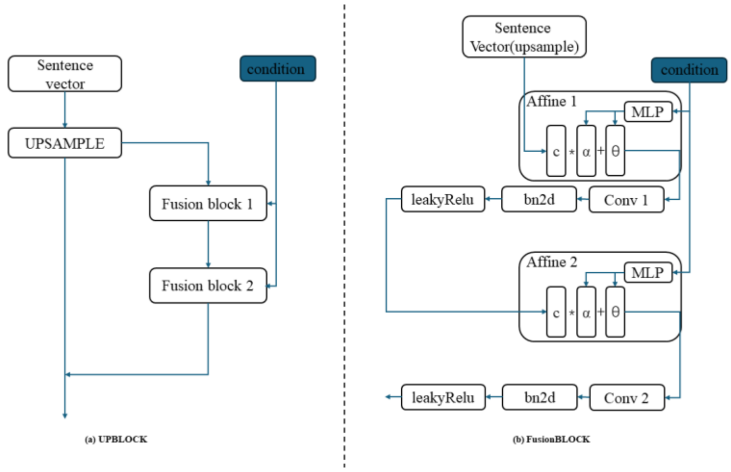


Figure 1. Pipeline of the model.

3.2.1. Text encoder. As shown in Figure 1. The input description will be divided into sentence vectors by the text encoder. It turns the text into the word embedding vectors relating to the context, so the model will receive a series of numeric vector which could represent semantemes and grammatical features. In order to extract the semantic vector from the provided text, the text encoder in the previous model is based on a bi-directional Long Short-Term Memory model. It is an attnGAN pre-trained model. The later text encoder is based on the transformer framework, self-attention mechanism. These structures allow the encoder to capture the long-distance dependence relationship. also lets the model focus more on the bird itself rather than its surroundings. The text encoder outputs the sentence vector to the UPBlock for upsampling operation to synthesize the image.



(a) UPBLOCK

(b) FusionBLOCK

Figure 2. UPBLOCK and FusionBLOCK.

3.2.2. UPBLOCK. The structure of UPBLOCK is shown in Figure 2, which consists of an upsampling operation and two DFBLOCK. DFBLOCK is used for deeper text and image fusion, which contains two affine layers, two leakyRelu activation layers, two batch normalization layers, and two convolutional layers. These affine layers. Sentence vectors will be upsampled in the upsample operation and fuse with feature of the image in two fusion blocks. The DFBLOCK compared to the former one, could handle the fusion of text and image more effectively by adding the batch normalization layer and activation layers. Its conditional input to the MLP (Multilayer Perceptron) in the affine layers to calculate the scaling parameters α and shifting parameters θ using sentence vector c [1].

$$\alpha = MLP_1(c) \tag{1}$$

$$\theta = MLP_2(c) \tag{2}$$

Then output varied vector:

$$affine(x_i \mid c) = \alpha \cdot x_i + \theta \tag{3}$$

The i-th channel of the image map channel is represented by x_i, while affine stands for the affine transformation. So, the block could shift the upsampled sentence vectors based on the predicted scaling parameters and shifting parameters. Meanwhile, we also improve the Generalization ability of the model by adding the batch normalization layers. It may solve the overfitting problems.

3.2.3. Convolutional layer. The Figure 3 shows the basic structure of the DF-GAN. The convolutional layer transforms the image feature into the image with these preprocessed sentence vectors.

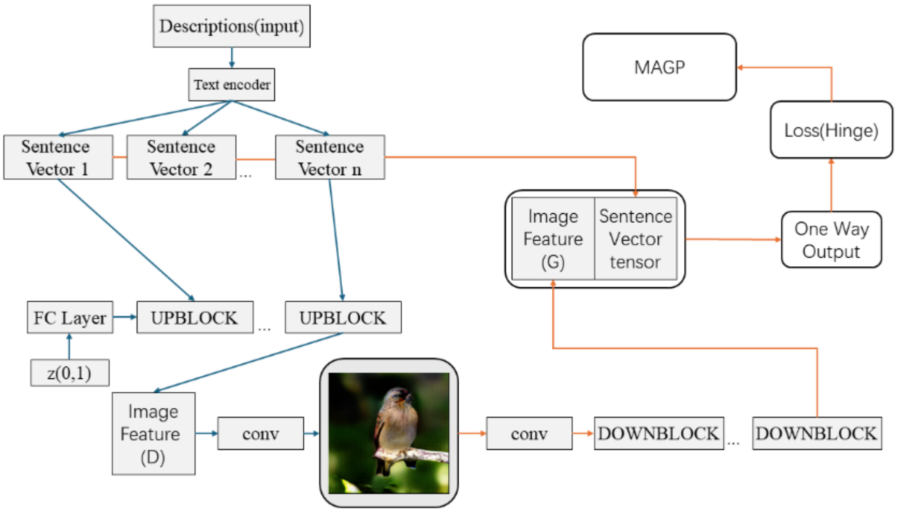


Figure 3. DF-GAN structure.

3.2.4. Loss function. This model uses Hinge loss as the loss function. The hinge loss is defined as:

$$max\left(0,1 - y \cdot \langle w, x \rangle\right) \tag{4}$$

where y is the true label which depends on the discriminator. w is white noise that follows the Gaussian distribution, and x is the sentence vector.

3.3. Implementation details

The study used Python 3.10 and the Colab environment, and the whole research was based on Pytorch. We first changed some baselines to launch DF-GAN on the Colab environment. Then, we changed the text-encoder from AttnGAN to Transformer module to give DF-GAN a stronger ability to capture text features from sentence level to word level. Thirdly, by optimizing the hyperparameters of adaptive moment estimation (Adam) optimizer, the quality of the generated image has an improvement. We also adjusted many key parameters including Wordtoix, Red, Green, Blue (RGB) module activation function which declined the figure of FID from 35.89 to 35.62, and Noise truncation rate to further improve the FID to 35.07. In addition, we redesigned the residual network and residual blocks, which declined the running time of generating images from 4.24 to 3.81 and got a lower score of 33.87 by optimizing the structure with FID. Finally, we failed to change FID to O-FID.

4. Result and Discussion

This chapter shows the details of improvements to the generated image and how these variations make sense. It includes improvements to the quality of the image, better performance on word-level descriptions, the stable shape of the birds and a more realistic image.

4.1. Improvements in the images

In Figure 4, the shape of the bird and quality are better than the former ones with the same description after changing the text encoder. As formerly mentioned, the text encoder uses transformer structures and attention mechanisms to allow the encoder to focus more on the bird itself. Hence, after replace encoder the generator will generate an image which will have a better performance on birds rather than the surroundings. It also helpful for generating speed.

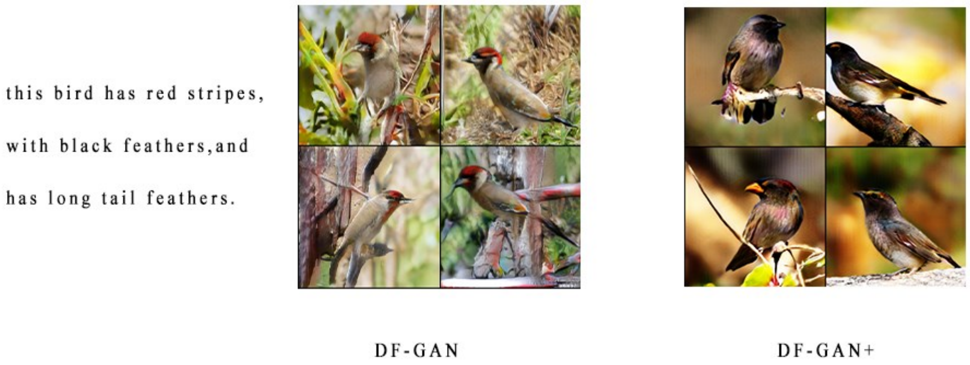


Figure 4. Better quality.

4.2. Better performance on word-level text

From Figure 5, it can be seen that original DF-GAN model has a poor performance on word-level text. Except for the second, all the generated image has an unrealistic shape that is caused by the original text encoder not comprehending the word-level text well. It has a significant improvement after replacing the text encoder, which has a stronger capability of word embedding and tokenization. Hence, the model could generate images through easier descriptions, which is also helpful to the cost of generating.

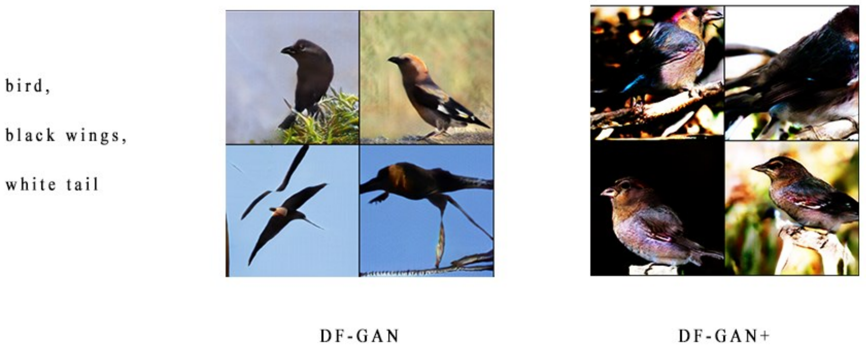


Figure 5. Word level text.

4.3. More realistic and reasonable shape

After serval generations there still exists some problems that the shape of the bird is still not that realistic, it may be caused by overfitting or the lack of samples from real images. So redesigned generator network is applied to the former model. Meanwhile, increasing the sample times is also helpful to this question. It can be seen that in Figure 6 some abnormal and unrealistic structures on the bird have been fixed, like in the second image, the former bird has two beaks, after applying the changes, it is more realistic. That will benefit reliable image generation.

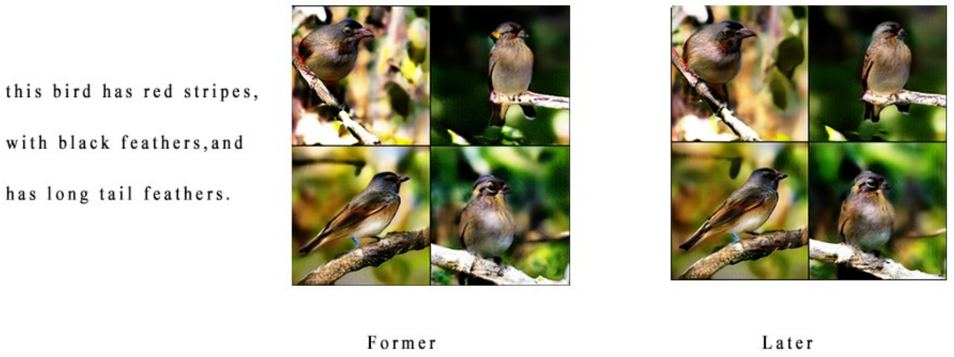


Figure 6. More realistic and reasonable shape.

4.4. Quantitative evaluation of FID Score and generating speed with existing method

From table 1 and table 2 it can be seen that after fine-tuning and redesigning the network, the FID score has a significant improvement. And the generating speed also improves. That means less consumption on graphics processing unit (GPU), which will lower the cost of generating images. The enhanced DF-GAN will be called DF-GAN+ for simplicity.

Table 1. FID Score result.

FID Score (↓)	initial	later
DF-GAN+ (encoder of AttnGAN)	14.81	12.00
DF-GAN+ (encoder based on transformer)	38.00	33.87

Table 2. Generating times consumes.

Time (↓)	1	2	3	4	5	average
DF-GAN	4.25	4.16	4.3	4.27	4.23	4.24
DF-GAN+	3.87	3.8	3.9	3.86	3.62	3.81

The enhanced DF-GAN will be called DF-GAN+ for simplicity. We use FID Score to make a quantitative evaluation. From table 3, it can be seen that after fine-tuning and redesigning the Generator network, the FID Score metric significantly decreased from 23.98 to 12.00 compared to existing methods, including DF-GAN. Also, as table 1 shows, compared with the baseline, DF-GAN+ has better performance with both encoders. That means DF-GAN+ produced images that were more similar to real images. However, the FID Score is focused on the similarity between the distribution of fake and real images. Due to the images produced by transformer-based encoders lacking information about their backgrounds, the similarity between these real and fake images may decrease, as well as their distribution. This could explain the strange increase in the FID Score between the model using the transformer-based encoder and AttnGAN. Therefore, an attempt was made to introduce O-FID into the evaluation system, which is described as a metric that focuses more on the main body of the image rather than the whole distribution. However, it did not work. Therefore, it may be addressed in future research.

From table 2, it shows an increase in generating speed, from 4.24 seconds to 3.81 seconds. This means less consumption of the GPU, which will reduce the cost of generating images. In summary, these experiments improve the quality of the generating and lower the generating speed by changing the text encoder, redesigning the netG, and optimizing some key parameters in the models. Finally, the performance of the model has a significant improvement.

Table 3. FID Score compared with exisiting method

MODEL	FID Score(↓)
AttnGAN	23.98
MirrorGAN	18.34
DM-GAN	16.09
DAE-GAN	15.19
DF-GAN	14.81
TIME	14.30
DF-GAN+(Ours)	12.00

5. Conclusion

In this paper, we introduced several key improvements to the DF-GAN model for text-to-image synthesis tasks. First, we upgraded the text encoder, addressing the limitation that DF-GAN could only process sentence-level textual descriptions. With the new text encoder, the model can now handle word-level descriptions, significantly enhancing its ability to synthesize fine-grained visual features. This improvement allows for more precise control over the generated images, leading to more detailed and accurate results. Second, by carefully adjusting a range of key parameters, we achieved a noticeable increase in the quality of the generated images. These changes improved realism to the images while also strengthening the generating process and guaranteeing consistency from different sources. Third, we optimized the architecture by redesigning the networks of the residual and structural layouts of residual blocks. This redesign resulted in a reduction in running time without compromising the quality of the generated images. The improved efficiency makes the model more practical for real-world applications, where both speed and quality are crucial. Additionally, our study highlighted the need for a new evaluation metric focused on the quality of specific objects within generated images. While we explored a similar concept known as O-FID, our implementation did not yield the expected results, suggesting that further research is needed in this area. Future work will aim to refine this evaluation method to better assess and improve the performance of text-to-image models.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Ramesh A Pavlov M Goh G Gray S 2021 Zero-shot text-to-image generation. In International conference on machine learning pp 8821-8831

[2] Goodfellow I Pouget-Abadie J Mehdi M Xu B Warde-Farley D Sherjil O Aaron C and Yoshua B 2014 Generative adversarial nets In Advances in Neural Information Processing Systems pp 2672–2680.

[3] Chen Y Liu J Peng L Wu Y 2024 Auto-encoding variational bayes Cambridge Explorations in Arts and Sciences 2(1)

[4] Ho J Jain A Abbeel P 2020 Denoising diffusion probabilistic models Advances in neural information processing systems 33 pp 6840-6851

[5] Takeru M Masanori K 2018 cGANs with projection discriminator arXiv Preprint:1802.05637

[6] Tao X Pengchuan Z Qiuyuan H Han Z Zhe G Huang X L and He X D 2018 Attngan: Fine-grained text to image generation with attentional generative adversarial networks In Proceedings of the IEEE conference on computer vision and pattern recognition pp 1316-1324

[7] Qiao T T Zhang J Xu D Q and Tao D C 2019 Mirrorgan: Learning text-to-image generation by redescription. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 1505–1514

[8] Zhu M F Pan P B Chen W and Yang Y 2019 Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 5802-5810

[9] Tao M Tang H Wu F Jing X Y 2022 Df-gan: A simple and effective baseline for text-to-image synthesis In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp 16515-16525

[10] Catherine W Branson S Welinder P Perona P Belongie S 2011 The Caltech-UCSD Birds-200-2011 Dataset California Institute of Technology

[11] Heusel M Ramsauer H Unterthiner T 2017 Gans trained by a two time-scale update rule converge to a local nash equilibrium Advances in neural information processing systems 30

[12] Dinh T M Nguyen R Hua B S 2022 Tise: Bag of metrics for text-to-image synthesis evaluation European Conference on Computer Vision. Cham: Springer Nature Switzerland pp 594-609