Research Progress of Image Generation Methods Based on Deep Learning

Senwei Zhang *

Department of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

* Corresponding Author Email: 21722076@bjtu.edu.cn

Abstract. This paper provides an overview of deep learning-based image generation methods and image style migration techniques. The focus is on text-to-image generation, unconditional and conditional image generation methods, and various approaches for image style migration. The review covers the underlying models, including Generative Adversarial Networks (GAN), Variational Autoencoders (VAE), and diffusion models, while also discussing conditional image generation methods such as Conditional Generative Adversarial Networks (CGAN) and Stacked Generative Adversarial Network (StackGAN). Furthermore, it explores deep learning-based image style migration methods, including both image iteration-based and model iteration-based approaches. The possibilities and opportunities for the advancement and use of these techniques in the field of picture production are highlighted in the review's conclusion.

Keywords: Deep Learning, Image Generation, GAN, VAE, CGAN, StackGAN.

1. Introduction

Deep learning has made it possible for machines to create graphical content for sharing and understanding based on textual descriptions. Image generation techniques have a wide range of applications in the field of artificial intelligence [1], such as image generation, image transformation, and style transformation data.

The process of creating a text-generated image can be thought of as the opposite of creating an image caption. It involves three primary tasks: 1) extracting the information embedded in the text; 2) generating an image based on the extracted information; and 3) testing the authenticity of the image and the text [1]. In early research, text-generated images relied heavily on a combination of search methods and supervised learning [2]. This method generates images from text by first identifying text units with information and then searching for the most likely image regions, thus optimizing the image layout conditional on text and image regions. However, the image features can only be retrieved again by altering the provided text because this traditional image production approach lacks diversity, and it is challenging to ensure the quality of the generated images.

Fortunately, deep learning-based approaches have made huge breakthroughs. Image generation through text is mainly divided into two parts: the natural language processing stage and the natural language processing stage [3]. During the initial phase of natural language processing, the model will interpret and process the written description. In order for the computer to comprehend the details found in the text description, such as the subject, noun, adjective, and other details, it is necessary to first extract the text features from the text description. In the second stage of the image generation process, the model is based on the text description to generate the corresponding image. The text information is converted by the computer into specific image information, which should reach the degree of fakeness. From the two parts of the text to generate image technology, it is clear that this process contains two parts of technology are natural language processing and image synthesis technology. The accuracy and diversity of the generated images are enhanced by this two-stage approach [3].

1.1. Unconditional Image Generation Method

Three primary approaches are used in conditional picture generation: autoregressive models, variational autoencoders (VAE) [4], and generative adversarial networks (GAN). A discriminator

network and a generator network collaborate to create realistic samples through ongoing competition in generative adversarial networks, or GANs. On the other hand, Variational Autoencoders (VAE) operate as generative models with the goal of obtaining latent variable representations in order to learn the underlying distribution of the input data and produce new samples that are similar to the training set. Autoregressive Models generate outputs based on previous outputs during the generation process. Examples include Recurrent Neural Networks (RNN) and Transformers, which are useful for generating sequence data like text and images.

1.2. Conditional Image Generation

Conditional image generation introduces additional conditional information, such as category labels, text descriptions, etc., to guide the generation process when generating images. The conditional image generation model enhances the controllability of image generation and has higher practical application value. In the application of unconditional and conditional image generation methods, it is necessary to comprehensively consider the relevant conditions and make appropriate choices according to the actual situation. The majority of generated material in practical applications has specific criteria, and the condition-guided picture generation approach can improve process control to ensure that the generated content meets target requirements.

2. Unconditional Image Generation Methods

Unconditional image generation refers to the direct generation of images from random noise through techniques such as Generative Adversarial Networks (GAN), Diffusion Modeling, etc. The goal of unconditional image generation methods is to fit the data distribution of the training set and generate images directly at the time of inference. The structure and design of the model are relatively simple, and the training difficulty is relatively small. However, its controllability is poor, and it can only generate images randomly according to the learned data distribution. On the plus side, it can concentrate on improving the quality and diversity of image generation and other features. Conditional image generation methods fully consider the given conditions in both training and inference to generate images with specific requirements. Compared with unconditional image generation methods, the structure of the network has one or more conditional branches that introduce conditions into the model, and also has to consider how to make the conditions effectively affect the generation results, which leads to certain limitations in the design of such models and higher training difficulty.

2.1. Generative Adversarial Network (GAN)

The generative adversarial network (GAN) was first introduced by Goodfellow et al. [5] in 2014 and is one of the primary network models for deep learning. Consequently, GAN and its derivative networks have emerged as one of the most popular paths in CV research in recent years, and GAN has drawn increased interest and study in the picture synthesis domain [6]. In contrast to conventional generative networks, which are just composed of generators, GANs are unable to produce meaningful data from a small training set. A GAN, on the other hand, consists of a discriminator and a generator. While the discriminator ascertains whether the data is genuine, the generator produces bogus information. It is possible to establish whether the final data generated is real or false by adjusting the two.

The generator and discriminator deep neural networks, which are trained separately, make up the unsupervised architecture known as a GAN. In order to fool the discriminator, the generator is fed random noise vectors to create images, and the loss function is optimized to make the distribution of the created images resemble the distribution of the genuine images. To enhance its discriminative power, the discriminator makes distinctions between the created image and the real image. The entire training process is an ongoing mutual game in which both parties optimize each other.

People see the world through text and image, and how these two modalities connect is a key area of study in these two disciplines. Images can yield rich and intricate textual semantic characteristics, and the development of GANs offers an unsupervised model [7] for picture generation. It is possible to embed the features into the image by first extracting important qualities (such space, relationships, and the state of objects) from the text and then using the discriminator and generator's game results in the GAN.

After the generative adversarial networks have demonstrated their excellent generative performance, people are not satisfied with their random generative properties, and there are still problems such as unstable network training process, insufficient feature extraction, and poor quality of the generated images for the GANs based text generating image methods [8]. StyleGAN [9] is inspired by style migration, and it redesigns the structure of generative networks to control the image generation process. Inspired by StyleGAN [9], it redesigns the structure of the generative network to control the image generation process, and directly controls the features of the generated image by adjusting the image style of the convolutional layer in the generator and discriminator, thus realizing the migration of the image style. Generative Adversarial Networks have been able to generate very high pixel counts and even fake images through learning, but the process of generating textual images, which is a cross-modal data mapping, is much more difficult than just generating images.

2.2. Diffusion Model

For generative modeling challenges, the diffusion model is a deep generative model. There are two stages to it: the forward phase and the retrograde phase of diffusion. Adding Gaussian noise step-by-step perturbs the input data in the forward diffusion phase. The generative model gradually reverses the diffusion process to retrieve the original input data during the reverse diffusion phase. Many tasks in computer vision, including image synthesis, super-resolution, restoration, editing, and translation, can be accomplished via diffusion modeling. Furthermore, the diffusion model's latent representations can be helpful in discriminative tasks including anomaly detection, image segmentation, and classification. The widespread use of diffusion modeling demonstrates its broad applicability [10] and that there are many undiscovered application areas.

The primary limitation of the diffusion models in use today is the requirement to carry out several steps in the inference procedure in order to produce a sample. Even with a lot of research in this area, GANs are still able to generate images more quickly than diffusion models. Furthermore, it is troublesome for diffusion models to use CLIP embeddings as a common technique for text-to-image generation. For instance, some research has shown that the model might not be able to produce legible text when utilizing CLIP embeddings to condition the denoising process since CLIP embeddings lack spelling information [11]. Diffusion models might so inherit this issue.

The quality and usefulness of the learned representation space for discriminative tasks can be assessed, applications to other computer vision tasks can be investigated, and multipurpose models can be learned for a variety of tasks. These are some ways to enhance the diffusion model. To make the diffusion model generation (sampling) process more efficient, the update rules of an efficient optimizer are borrowed. As a result, there is less uncertainty and fewer significant sampling steps, guaranteeing that the learnt Gaussian distribution can explain the data samples produced at each stage. Subsequent studies may investigate the diffusion model's application to additional computer vision applications as visual quizzing, video anomaly detection, and image defogging.

3. Conditional Image Generation Methods

Current text-to-image methods mainly include variational autoencoder (VAE), align-DRAW and generative adversarial networks. Since generative adversarial networks are excellent in generating images, current text-to-image models are improved and researched based on them. Currently, the main objects of text generated images are faces, flowers and birds. Text generated image usually consists of two phases, i.e., natural language processing and image generation phase. In the natural

language processing stage, the accuracy of the text vectors first determines the validity of the constraint information, and whether the computer can understand the feature information expressed in the text is the key to generating high-quality images. A good natural language processing model can convert text vectors more efficiently and help computers capture features and learn for understanding. Two primary research questions at the image production stage are how to enhance the generated image's resolution and the degree of textual and image matching.

After much research, the problem of creating text-generated graphics has progressively advanced in recent years, and several research models, like CGAN, have shown some degree of success. but the generated images still have some obvious problems. The images are more like a simple splicing of text attributes and lack consideration of the overall layout. Sometimes the blurring and weirdness of the background of the described objects make the images look less realistic. A text generation image model based on multi-stage generative adversarial network combined with spatial attention mechanism is developed on this.

3.1. Conditional Generative Adversarial Network (CGAN)

Based on the Generative Adversarial Network (GAN), the Conditional Generative Adversarial Network (CGAN) is an enhanced generative model that adds more conditionally restricted variables to precisely regulate the generated images. The network structure of CGAN is similar to that of the original GAN, but its key innovation is the addition of conditional constraints to the generator and discriminator by adding a conditional variable c as input. This condition variable can contain category information, auxiliary information, etc., which enables the generator to generate images with desired categories or attributes according to specific conditions. When the condition variable is a category attribute, CGAN can be regarded as a neural network model with supervised learning.

The emergence of CGAN has greatly expanded the application area of generative adversarial networks, making it a generalized network modeling method. By introducing conditional variables, CGAN is not only able to generate images of specific categories, but also images with specific attributes, which satisfies the precise demand for image generation [12].

Nevertheless, CGAN is not without its challenges and shortcomings. Since the quality and diversity of the training data determine how well CGAN performs, first a sizable amount of labeled data is required to ensure the variety and quality of the generated images. Furthermore, CGAN could have pattern collapse, when the generator is unable to generate a variety of images. Furthermore, the generated images' realism still needs work, and both their realism and clarity need to be further enhanced.

3.2. Stacked Generative Adversarial Network (StackGAN)

Zhang et al. presented StackGAN, a Stacked Generative Adversarial Network (SGAN). The model uses a step-by-step method to produce high-quality photos and has two pairs of Conditional Generative Adversarial Networks (CGAN). In the first stage, a text encoder converts text descriptions into text vectors corresponding to the features and passes them as conditional constraints along with random noise into the generator to produce a 64×64 low-resolution image. In the second stage, the text vectors and low-pixel images are passed to the generator network in the second stage, and the generator is trained to produce a high-resolution image with 256×256 pixels. The discriminative model of StackGAN uses multiple convolutional and pooling operations for down sampling to form a tensor of 4×4×512, and the discriminator determines that only the samples of the real image and the corresponding textual descriptions are positive samples during the training. pairs as positive samples.

Despite its achievements in the field of text-generated images, StackGAN still has some drawbacks, such as the fact that the generated images sometimes lack consideration of the overall layout and the blurriness and weirdness of the background of the described objects make the images look less real [13].

StackGAN++ is an improved version based on StackGAN, which is a tree-structured network model consisting of a combination of multiple pairs of generators and discriminators that can generate

progressively finer images, including 64 × 64, 128 × 128, and 256 × 256-pixel images. Each pixel generator corresponds to a discriminator trained alternately, thus realizing an end-to-end generative adversarial network. This improvement allows StackGAN++ to generate better quality images [14].

3.3. ControlNet

A neural network architecture called ControlNet is used to give massive, previously trained text-to-image diffusion models spatial conditioning control. Large, production-ready diffusion models are the focus of the architecture, which repurposes their robust, deep coding layers—pre-trained on billions of images—as a potent foundation for learning a wide range of conditional controls. To ensure that no detrimental noise affects the fine-tuning, the parameters are gradually increased through "zero-convolution" connections. ControlNet has been demonstrated to train robustly on both small and large datasets by testing a variety of conditioning controls, including edges, depth, segmentation, and human pose. The findings of this study could contribute to improving image diffusion models' control performance across a larger range of applications.

An end-to-end neural network architecture called ControlNet is used to teach massive, previously trained text-to-image diffusion model conditional control. ControlNet locks the parameters of the huge model and simultaneously generates a trainable replica of its coding layer, maintaining the quality and capacity of the model. Large pre-trained models are seen by the architecture as a potent tool for learning a variety of conditional controls. Zero convolutional layers connected the trainable replicas and the original locked model. The weights of these layers are initialized to zero and grow progressively during training. This architecture prevents damaging noise from being introduced to the big diffusion model's deep features at the start of training and shields the trainable replica's large-scale pretrained support from being tainted by it. Tests demonstrate that ControlNet is capable of managing Stable Diffusion with a wide range of conditioning inputs, such as segmentation maps, depth, Hough lines, user graffiti, Canny edges, and human key points. Experiments demonstrate how the technique enables combinations of many conditions and are conducted using individual conditioned images, either with or without textual cues [15].

4. Deep Learning Based Image Style Migration Method

Image style migration techniques are being used more and more in daily life. Typically, they are used to transfer an artist's style to a regular photo, giving the migrated image a sense of art; or they are used to transfer the style of two real photos, changing the daytime tonight and the cloudy day to sunny day, and so forth. Numerous academics have put out various techniques for visual style migration in support of these concepts. Nevertheless, content distortion and edge-to-edge detail loss plague the photos produced using so many different techniques.

The primary foundations of conventional non-parametric image style migration techniques are texture synthesis and physical model drawing. By splicing and rearranging sample textures, simple texture algorithms can create new textures. Conversely, Hertzman et al. proposed an analogy-based technique to generate novel textures in images by the use of correlations between image feature mapping [16]. Gatys et al. developed the first method of visual style migration based on convolutional neural networks with the advent of deep learning [17]. By processing high-level abstract feature representations individually, this approach generates artistic effects. Starting with a random noise image, the method creates a synthetic image with the original content and the new style using a pre-trained VGG model to extract the high-level abstract feature representations of the content and style images. Industrial applications successfully employ these techniques.

Currently the main deep learning-based image style migration methods include two categories: image iteration based, and model iteration based.

4.1. Image Iteration Based Style Migration Methods

Three basic types of image iteration-based style migration algorithms are based on deep image analogies, Markov random fields, and maximum mean difference. The most widely used approaches are those that rely on depth image analogy, Markov random field, and maximum mean difference. By extracting the stylistic feature representations of an image using the VGG19 network and reducing the maximum mean difference to match the stylistic features, Gatys et al. suggested a maximum mean difference-based style migration approach. Also, color histogram matching and luminance-only migration strategies were introduced in order to preserve the color information of the original image [18].

In addition, Li et al. proposed a style migration method based on Markov Random Fields, which achieves the capture of spatial distributional links of images by combining convolutional neural networks and Markov regularization models, thus improving the visual reasonableness of the synthesized images. However, this method is more demanding on the input image, requiring the content image to have a similar shape to the style image.

Meanwhile, Liao et al. explored a style migration method based on deep image analogies, using a pre-trained VGG19 network combined with the PatchMatch algorithm to achieve semantic-level visual attribute migration [19].

Although the method performs well on semantically similar content, it has some limitations in terms of generalization ability. It is characterized by high quality of synthesized images, good controllability, easy parameter tuning, no need to train the network, long computation time, and high dependence on pre-trained models [20].

4.2. Model Iteration Based Style Migration Algorithm

Although the method based on image iteration can produce style synthesized images with excellent results, it suffers from the problem of inefficiency. While model iteration-based image style migration methods train generative models through a large amount of data, model iteration-based style migration algorithms are mainly generative model-based style migration and image reconstruction decoder-based style migration. Generative model-based style migration optimizes the high-level abstract feature representations extracted from pre-trained VGG models by using residual networks as the basic components of the generative model and combining them with a perceptual loss function. This method has pioneered a new idea of style migration and has been widely used in the field of image processing [21]. And the style migration based on the generative model proved experimentally that using instance normalization instead of batch normalization can significantly improve the quality of the generated images. Furthermore, Zhang et al. developed a generative model that enables quick style migration for many styles and can train numerous styles [22].

5. Conclusion

In the subject of picture generation, deep learning-based text-based image generation techniques, unconditional image generation techniques, conditional image generation techniques, and image style migration techniques are all crucial. These techniques use a variety of cutting-edge models and algorithms, including CGAN and GAN, to cover the entire process from natural language processing to image production. Through continuous technological innovation and development, researchers have gradually solved the challenges in text-to-image generation and image style migration, providing important ideas and methods for image generation and artistic effects. The continuous improvement and application of these methods will further promote the development of image generation technology and expand its application prospects in various fields.

There is a lot of room for advancement in deep learning-based text creation image methods, unconditional image generation methods, conditional image generation methods, and image style migration methods. In the future, more efficient and accurate models and algorithms may appear, which can realize more realistic and diverse image generation effects. At the same time, with the

deepening understanding of the generation model, it will also help to improve the quality and efficiency of image generation. Deep learning-based image generation technology will continue to innovate, injecting more surprises and possibilities into the future digital world.

References

[1] LI Yueyang, TONG Guoxiang, ZHAO Yingzhi, LUO Qi. A Survey of Text-to-Image Synthesis Based on Generative Adversarial Network. Electronic Science and Technology, 2023, 36(10): 39-55.

[2] Zhu X, Goldberg A B, Eldawy M, et al. A text-to-pictur. Van-synthesis system for augmenting communication cover: Proceedings of the Sixth International Conference on Learning Representations, 2018: 1710-1735.

[3] YI Yaokuan. Research on Generative Adversarial NetworkResearch on Algorithm of Text to Image Based. Changchun University of Technology, 2021.

[4] Kingma DP, Welling M. Auto-encoding variational Bayes. IEEE Computer Society, 2021, 43(12): 4217-4228.

[5] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Advances in neural information processing systems. Cambridge MIT Press, 2014: 2672-2680.

[6] AGNESE J, HERRERA J, TAO H, et al. A survey and taxonomy of adversarial neural networks for text-to-image synthesis. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2020, 10(4): 1345.

[7] PAN Z, YU W, YI X, et al. Recent progress on generative adversarial networks (GANs): A survey. IEEE Access, 2019, 7: 36322-36333.

[8] WANG Zhendong, ZHENG Huangjie, HE Pengcheng, et al. Diffusion-GAN: Training GANs with diffusion. arXiv: 2206.02262, 2022.

[9] Karras T, Laine S, Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 4401-4410.

[10] F.-A. Croitoru, V. Hondru, R. T. Ionescu and M. Shah, Diffusion Models in Vision: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 10850-10869.

[11] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents, 2022.

[12] H. Zhang, V. Sindagi and V. M. Patel, Image De-Raining Using a Conditional Generative Adversarial Network, IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(11) : 3943-3956.

[13] Zhang H, Xu T, Li H, et al. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017: 5907-5915.

[14] H. Zhang, StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8) :1947-1962.

[15] Zhang, L., Rao, A., Agrawala, M. Adding Conditional Control to Text-to-Image Diffusion Models. IEEE/CVF International Conference on Computer Vision, 2023, 3836-3847.

[16] Hertzmann A, Jacobs C E, Oliver N. Image analogies, Proc of the 28th Annual Conference on Computer Graphics and Interactive Techniques. New York: ACM Press, 2001: 327-340.

[17] Gatys L A, Ecker A S, Bethge M, et al. Controlling perceptual factors in neural style transfer, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3985-3993.

[18] Gatys L A, Ecker A S, Bethge M. A neural algorithm of artistic style. 2015.

[19] Li Chuan, Wand M. Combining Markov random fields and convolutional neural networks for image synthesis, IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2479-2486.

[20] Liao Jing, Yao Yuan, Yuan Lu, et al. Visual attribute transfer through deep image analogy. ACM Trans on Graphics, 2017, 36(4) : 120.

[21] Justin J, Alexandre A, Li Feifei. Perceptual losses for real-time style transfer and super-resolution, European Conference on Computer Vision, 2016: 694-711.

[22] Li Yijun, Fang Chen, Yang Jimei, et al. Universal style transfer via feature transforms. 2017.