Research on Text Based Generative Image Editing

Yiren Wang[a]
Jiangsu No.10 High School, Suzhou, China

Keywords: Image Editing, Deep Learning, Generative Adversarial Networks, Diffusion Model.

Abstract: In recent years, with the continuous development of artificial intelligence the quality of images generated by using deep learning ability has gradually improved. Image editing uses the deep learning model to use conditional information as a guide and modify some regions in the image, while other regions remain unchanged. Generated image is a kind of automatic image editing, which can automatically generate images that meet the requirements of the user. Editable content includes, but is not limited to, the color, shape, texture, overall style and other features of the image to be edited. Generative image editing is an important part in the field of computer vision, and generative image editing has important use and theoretical value. This paper is a review based on the generative image editing with text as the control condition. This paper classifies the implementation method of text-based generative image editing, which is divided into generative adversarial network (GAN), diffusion model, CLIP model, and autoregressive model. The advantages and disadvantages of various implementation methods are analyzed, and various evaluation indexes of image editing are introduced. Finally, the future of the field is discussed.

1 INTRODUCTION

Over the past few years, artificial intelligence has been continuously developing, the generative deep learning ability of artificial intelligence has continuously entered various fields. In the field of image editing, the quality effect of pictures generated by using generative deep learning is also gradually increasing, bringing progress and innovation to this field. Under this trend, generative image editing has gradually matured and developed new research directions in the field of image processing and computer vision. Image editing uses the deep learning model to use conditional information as guides, modifying some areas in the image, while other regions remain unchanged. Generated image is a kind of automatic image editing, which can automatically generate images that meet the requirements of users. Generative image editing can create more imaginative images, can generate images with new artistic style, make images more innovative, and text-based generative image editing refers to change some scenes in the form of text input according to user needs, while other scenes remain the same.

Nowadays, the application scenarios of generative image editing technology are very wide: (1) newly proposed texture expression in the field of clothing design: texture (Bi-Colored edge）（Li, 2020） can generate images with controllable texture and details and interact with images; (2) e-commerce platform through image editing technology to create Banner charts related to business, and create virtual AI characters for live broadcast. (3) In the medical field, special image editing models can automatically generate medical images that meet clinical needs; (4) In the field of public safety, some image editing models can improve the efficiency of the police and the identification of children by editing the age attributes of characters.

Text-based image editing follows text instructions for image editing, allowing iterative refinement of image editing capabilities with natural language. With the development of the generative model, the text conditional image model can synthesize the images according to the text prompts, and the irrelevant objects can be composed in a semantically reasonable way, making the model flexible. Given that text image editing has strict requirements for the authenticity of the generated images and the

[a] https://orcid.org/0009-0005-7423-6336

Wang, Y.
Research on Text Based Generative Image Editing.
DOI: 10.5220/0013241000004558
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 1st International Conference on Modern Logistics and Supply Chain Management (MLSCM 2024)*, pages 146-150
ISBN: 978-989-758-738-2

consistency with the text of the images, in addition to the powerful generation power, the editing methods must also have the ability to capture the high-dimensional semantics of the target.

However, the text based generative image editing literature review is relatively limited, in order to use more convenient method based on text based generative image editing, this paper based on text generative image editing implementation method is divided into through generative adversarial network (GAN), CLIP, self regression and diffusion model these four, and summarizes the four methods, analysis of the advantages and disadvantages of various methods. Finally, the shortcomings of the existing text-based generative image editing methods are summarized and analyzed, and the possible future research directions are discussed.

2 OVERVIEW OF THE TEXT-BASED GENERATIVE IMAGE-EDITING METHODS

2.1 GAN-Based Text-Image Editing

Text-driven image editing based on GAN uses the powerful generating ability of GAN to reverse the real image into a controllable potential space and obtain the target image through the forward transmission. Xu（Xu, Zhang, Huang, et al., 2018) proposed Attn-GAN, introduce the attention generation network to focus on the relevant text in the natural language description; Attn-GAN is capable of synthesizing fine-grained in different areas of the image by concentrating on the keywords in the user's natural language description, and realize fine-grained text-to-image generation through the attention generation network. And the author presents an idea for calculating the fine-grained image-text matching loss for generator training by utilizing a deep attention multimodal similarity model. The Attn-GAN model achieved an initial score of 4.36 on the CUB dataset, improving the initial score for the best report on the CUB dataset by 14.14%.25.89 on the more challenging COCO dataset, up 170.25% compared to the best initial score. It is evident that compared to the previous state-of-the-art methods , it has a novel attention mechanism and has better results in generating complex scenes.

Chen Ziyi (Cheng, 2023) uses the Transformer-based neural network instead of the convolutional neural network to introduce it into the GAN, and the discriminator integrates the word-level discriminator to give the generator more granular feedback. In the task of generating images for text, the diversity of training data is particularly important. In the qualitative experiment, bird images were generated on the two data sets of CUB and COCO. It can be seen that this model well realizes the multi-modal fusion of text and image. It can be seen that the introduction of Transformer makes the text model have more advantages in the two modal information extraction of text and image.

GigaGAN (Kang, Zhu, Zhang, et al., 2023) is the first GAN model that can train billions of parameters on large-scale datasets. It proposes a dynamic convolution kernel selection mechanism based on text conditions to increase generator capacity and expression capacity to increase inference time by several orders of magnitude. Self-attention and cross-attention are staggered between convolution layers to support various potential space editing applications. Multi-scale training is reintroduced to improve the generation quality of image text alignment and low-frequency details to generate high-resolution images.

The advantage of GAN-based text image editing is that only one forward transmission is required in the generation process, which takes less time than other methods. Disadvantages: it is difficult to control the output, the resulting image may not meet the user's expectations; GAN needs to judge whether the produced image belongs to the same category as other images, so the generated image is not innovative; the training is unstable and it is easy to crash.

2.2 Text-Image Editing Based on the Diffusion Model

The diffusion model gradually introduces noise to the image, thereby destroying the initial data distribution, and uses the prediction noise and denoising by learning to obtain the probability distribution of the generated image. Over time, the model learns to generate new images based on noisy input through multiple iterations.

Imagen (Saharia, Chan, Saxena, et al., 2022) A model for the diffusion of text into images is what it is. Imagen built on understanding the text of large transformer language model, and rely on the diffusion model of high fidelity image generation, using the frozen text encoder to get text features, text features into the image diffusion model get 64x64 images, after two super resolution diffusion model finally get 1024x1024 image. Imagen Usually operates directly in the pixel space, but generally requires a lot of time, and has a high reasoning cost.

Stable Diffusion (Rombach, Blattmann, Lorenz, et al., 2022) The diffusion model can be trained with limited resources, reducing both the complexity and retaining the details of the image when training the diffusion model, so that the generated image can achieve high fidelity.

ERNIE-ViLG 2.0 (Feng, Zhang, Yu, et al., 2023) is the first image text generation model in the Chinese field. It focuses on guiding the model to pay more attention to the alignment of image elements and text semantic in the learning process, so that the model can deal with the relationship between various objects and attributes. All of the above work is guided without the classifier, without training the classifier to guide the model learning.

2.3 Text Image Editing Based on the CLIP Model

The CLIP (Radford, Kim, Hallacy, et al., 2021) model uses contrast learning to collect 400 million image text pairs from the network for training, acquiring a multimodal embedding space that can be utilized to determine the semantic similarity between a text and image.

Patashnik (Patashnik, Wu, Shechtman, et al., 2021) combined the generating ability of StyleGAN generator with the visual language ability of CLIP, proposing three methods of combining CLIP with StyleGAN. Text-guided latent optimization is the first approach, and the CLIP model serves as the loss network. The second option is the latent residual mapper that has been trained for specific text prompts, based on the starting point of the latent space, the mapper produces local steps in the latent space. Finally, the method of mapping text prompts to input unknown direction in the StyleGAN style provides control over the strength of operation and the degree of entanglement.

CLIP as a loss network: CLIPDraw (Frans, Soros, et al., 2022) The cosine distance between the text cue features and the generated image features obtained from CLIP is used to calculate loss. VQGAN-CLIP（Crowson, Biderman, Kornis, et al., 2022) modulates the intermediate potential vector by comparing the similarity of the generated image to the specified text in the CLIP model, thus generating images consistent with the text description.

CLIP acts as a joint embedding space: StyleGAN-NADA (Gal, Patashnik, Maron, et al., 2022) encodes the differences between domains as the text description direction of the CLIP embedded space, using two generators, one generator remains frozen to provide samples from the source domain, and the other generator is optimized to generate images different from the source domain in the CLIP space, but separate adhesion and distortion during face editing. HairCLIP (Wei, Chen, Zhou, et al., 2022) On the basis of StyleCLIP, image and text features are extracted by CLIP as conditions, and entanglement information injection makes the properties separate to different mappers.

No dot should be included after the section title number.

2.4 Text Image Editing Based on the Autoregressive Model

Early visual regression models executed visual synthesis through pixel-by-pixel execution., however, these methods are computationally costly on high-dimensional data. With the rise of VQ-VAE (Oora, Vinyals, Kavukcuoglu, et al., 2017), image visual synthesis tasks can benefit from using large-scale pre-training models, enabling autoregressive combination with VQ-VAE or other methods to achieve better image quality.

DALL-E (Ramesh, Pavlov, Gon et al., 2021) borrows from Transformer (Vaswani, Shazeer, Parmar, et al 2017) ideas to model image and text autoregressions as a single data stream. The images with the highest similarity were then screened by the CLIP (Cheng,2023) classifier. However, DALL-E (Ramesh, Pavlov, Gon et al., 2021) generates a relatively small image size, and a large amount of CPU memory is required to generate high-precision images.

Parti (Yu, Xu, Koh, et al., 2022) is a two-stage model composed of image marker and autoregressive model: the first stage designs the training marker; the second stage trains the autoregressive sequence of image marker from text tag to sequence model. Text-based generative image editing has become an important part of the image editing field.

The advantages of autoregressive-based text image editing are the advantages of clear probabilistic modeling and stable training. However, Converting the image into a token for autoregressive prediction is necessary for this method, which requires a slow inference speed and a large number of parameters. Text image editing based on diffusion model belongs to the likelihood model, so it is easier to train during training, and the image generation is innovative, which is more suitable for large-scale data set training compared with GAN. Although iterative methods can achieve stable training with simple goals, they can generate high computational costs during inference.

3 EVALUATING INDICATOR

At present, specific evaluation indicators are lacking in the field of image editing, so quantitative evaluation indicators are employed to evaluate image generation. Several quantitative evaluation indicators have been proposed in the field of image generation, which are used to measure the quality, authenticity, and diversity of the generated images.

Generated image diversity evaluation index: the diversity degree of the generated image can be measured by it, and evaluate it by calculating the degree of sample difference of the generated image.

Generated image authenticity evaluation index: It has the ability to determine how closely the generated image matches the actual image. Content consistency evaluation index of the generated image: Ensure that the content of the generated image is identical to that of the input image. To measure the degree of reconstruction error between the generated image and the input condition, this index can be utilized.

4 CHALLENGES AND PROSPECTS

At present, generative image editing based on text can achieve good results, but the relevant research is still in its initial stage, and there are still many valuable research questions that need to be solved urgently.

1. Diversified image editing methods: Existing image editing methods rely on text to generate images. DragGAN (Pan, Tewari, Leimkuhler, et al., 2023) proposes a new editing paradigm, based on point drag and drop interactive image editing, which can drag the content of the control points to the target points by defining multiple control points in the image and corresponding target points. This method uses neighbor search technology and GAN discriminative features to make precise handle point positioning using feature-based motion supervision loss to guide the handle points to the target point through the intermediate feature map of GAN, ensuring that the edited features remain stable around the control point. In the future, diverse editing approaches will attract wider academic attention in the research field.

2. Further fusion between modes: Although the current generative image editing technology can generate high-quality images, there are some challenges in the control ability, especially when using text to guide image generation. The difficulty mainly stems from the divergent nature of the generative model, leading to uncertainty between the generative results and the user's expectations. In the future, the closer integration of modes can be further implemented in the following directions: text combined with line draft, text combined with line draft and physical example images, and text combined with other conditions, including but not limited to semantic map, depth map, normal vector map, and other semantic correlation geometric features.

3. Ensure that the image editing area and the non-editing area transition is natural. In the physical paradigm guiding image editing, the model is easy to learn a simple mapping function, The network is unable to comprehend the relationship between the content of the reference image and the source image due to this.. The transition area has obvious artifacts, texture blur, structure distortion and other problems. In the future, attempts can be made to smooth image editing boundaries, including image priors, data enhancement and classifier guidance. Therefore, ensuring a smooth transition between boundaries should be a key study of image editing.

5 CONCLUSION

This paper classifies the method of generative image editing based on text and introduces the method, sorts out the advantages and disadvantages of image editing through different methods, and introduces the evaluation indexes that can evaluate the quality of image editing. In the future, in the field of image editing, it is hoped to develop more image editing methods, and also need to strengthen the control ability of generative image editing. And develop specific evaluation indicators in the field of image editing.

REFERENCES

Li, Y., 2020. Clothing image generation and interactive editing based on deep learning.

Xu, T., Zhang, P., Huang, Q., et al. 2018. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Network Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1316-1324.

Chen, Z., et al. 2023. Research on text-oriented image editing algorithm. Jiangxi University of Science and Technology.

Kang, M., Zhu, J., Zhang, R., et al. 2023. Scaling up gans for text-to-image synthesis Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 10124-10134.

Saharia, C., Chan, W., Saxena, S., et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language

Understanding. Advances in Neural Information Processing Systems, 35: 36479-36494.

Rombach, R., Blattmann, A., Lorenz, D., et al. 2022. High-Resolution Image Synthesis with Latent Diffusion Models Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 10684-10695.

Feng, Z., Zhang, Z., Yu, X., et al. 2023. ERNIE-ViLG 2.0: Improving Text-to-Image Diffusion Model with Knowledge-Enhanced Mixture-of-Denoising-Experts Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 10135-10145.

Radford, A., Kim, J., Hallacy, C., et al. 2021. Learning transferable visual models from natural language supervision international conference on machine learning. PMLR, 8748-8763.

Patashnik, O., Wu, Z., Shechtman, E., et al. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery Proceedings of the IEEE International Conference on Computer Vision, 2085-2094.

Frrans, K., Soros, L., 2022. WITKOWSKI O. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. Advances in Neural Information Processing Systems, 35: 5207-5218.

Crowson, K., Biderman, S., Kornis, D., et al. 2022. VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance European Conference on Computer Vision. Cham: Springer Nature Switzerland, 88-105.

Gal, R., Patashnik, O., Maron, H., et al. 2022. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. ACM Transactions on Graphics, 41(4): 1-13.

Wei, T., Chen, D., Zhou, W., et al. 2022. Hairclip: Design your hair by text and reference image Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,18072-18081.

Oord, A., Vinyals, O., Kavukcuoglu, K., 2017. Neural discrete representation learning. Advances in neural information processing systems.

Ramesh, A., Pavlov, M., Goh, G., et al. 2021. Zero-Shot Text-to-Image Generation International Conference on Machine Learning. PMLR, 8821-8831.

Vaswani, A., Shazeer, N., Parmar, N., et al. 2017 Attention is All you Need. Neural Information Processing Systems.

Yu, J., Xu, Y., Koh, J., et al. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. A Computational Study. arXiv:2206.10789[EB/OL].

Pan, X., Tewari, A., Leimkühler, T., et al. 2023. Drag your gan: Interactive point-based manipulation on the generative image manifold ACM SIGGRAPH 2023 Conference Proceedings.