Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/81/20241082

A Review of Facial Generation Technology Research

Haohua Wang

School of Electrical Engineering and Computer Science, Peking University, Beijing, China

whhpkuxky@stu.pku.edu.cn

Abstract. Face generation, as a cutting-edge generative technology, has made significant strides in various image synthesis tasks, including facial inpainting, text-to-image translation, and video-based facial animation. Generating realistic and diverse human faces is a critical task in computer vision, with a wide array of real-world applications. Given the remarkable success of face generation models, there has been a growing interest in leveraging advanced techniques to further improve the quality, diversity, and control of generated faces. This paper will provide a comprehensive overview of the state-of-the-art face generation techniques. Specifically, the paper reviews the key approaches in this field, including Generative Adversarial Networks (GANs), Vector Quantized methods, and the more recent Diffusion Models, each of which has contributed significantly to the advancement of face generation. The paper discusses how these models have evolved to handle the complexity of face generation, from capturing subtle facial details to enabling fine-grained control over facial attributes. The paper also explores the major applications of face generation technologies, with a particular emphasis on their use in entertainment, virtual reality, and security. Finally, the paper identifies promising directions for future research in face generation, such as improving the interpretability of models, addressing ethical concerns, and enhancing the ability to generate faces that are both highly realistic and diverse.

Keywords: Face generation, Diffusion model, Generative adversarial network, Variational autoencoder.

1. Introduction

Face generation, a significant generative task within the field of artificial intelligence, involves creating realistic and diverse human facial images from various input modalities. Among the predominant methodologies advancing the field of face generation are Vector Quantized methods [1], Generative Adversarial Networks (GANs) [2], and Diffusion Models [3-5]. Each of these approaches uniquely contributes to the development of this domain, demonstrating distinctive strengths in the synthesis of realistic and diverse facial images. Vector Quantized methods deploy a framework comprising an encoder and a decoder. The encoder compresses the input into a discrete latent space defined by statistical parameters. New data are then generated by sampling from this discrete latent space and reconstructing through the decoder, making it adept at synthesizing the inherent variability in facial features. GANs feature a dual-network architecture where the generator creates images intended to be indistinguishable from authentic data, and the discriminator evaluates their realism. This adversarial feedback loop enhances the generator's ability over time, enhancing its capacity to produce photorealistic facial images with complex textures and details. Diffusion Models simulate a process that gradually adds noise to data until only Gaussian noise remains. The model then learns to reverse this process, effectively denoising to regenerate the original data with added conditions. Therefore, the paper provides an overview of face generation methods respectively based on Vector Quantized methods, GANs and diffusion model, and then discuss the future research direction for face generation.

2. Face generation

In this section, the paper will classify the face generation methods into the three paradigms previously introduced.

2.1. Face generation based on Vector Quantized methods

In the realm of face generation, several advanced methodologies leveraging vector quantization have shown promising capabilities. Models based on vector quantization are initially designed to complete more comprehensive generation tasks, thus they can also be applied to the downstream tasks like face generation. (Vector-Quantised Variational AutoEncoder) VQ-VAE [1] is the very first vector quantized method, which integrates vector quantization with the traditional VAE [6] framework to encode input images into discrete latent spaces. This quantization simplifies the latent representation and enhances the robustness of the image reconstruction, aiding in producing diverse facial images with significant control over their features. Though VQ-VAE has strong generation capability, the objective only consists of one reconstruction loss and two regularization loss, which is still simple. To enhance the model's capability, VQ-GAN [7] improves the objective. VQ-GAN [7] combines the adversarial training mechanism of GANs with vector quantization to improve the output image quality. Due to the addition of the GAN loss, the discrete codebook becomes more interpretable. Lastly, to speed up the generation process, traditional autoregressive generation is improved. MaskGIT [8] utilizes a two-stage process that combines vector quantization with transformer-based generative modeling. Initially, the model employs a VQ-GAN [7] to encode images into a compressed, discrete format, significantly reducing the dimensionality and simplifying the image data into manageable sequences of discrete tokens. Subsequently, MaskGIT leverages a bi-transformer model in its second stage to generate or complete images by predicting masked segments based on the surrounding tokens. While MaskGIT is capable enough to complete simple face generation tasks, it could be further improved by fusing other modality information. This would involve modifying the transformer to handle multimodal inputs, or trains specific encoder to encode text, segmentation mask or sketch to the same domain of the image discrete tokens.

2.2. Face generation based on GANs

The evolution of face generation has been prominently marked by the development of GANs, particularly with the inception of StyleGAN [9-11] and its subsequent iterations. Each generation of StyleGAN has introduced significant enhancements that have continuously pushed the boundaries of realism, control, and diversity in generated facial images. StyleGAN [9] was the first to introduce a sophisticated system of layered control through a disentangled latent space, represented as z , which is mapped to an intermediate latent space w. This mapping, $f: z \rightarrow w$, allows for more stable and controllable style changes at different levels of the synthesis process. Each level of detail is controlled by a different segment of the w vector, influencing the generation through adaptive instance normalization (AdaIN) at each layer of the generator: $AdaIN(x_i, y) = y_{s,i} \frac{x_i - \mu(x_i)}{\sigma(x_i)} + y_{b,i}$, where x_i is the feature map of the style i, and y is the style vector derived from w, split into scale $y_{s,i}$ and bias $y_{b,i}$ components. This approach enabled precise manipulation of facial features across different levels of detail. StyleGAN2[10], a direct successor, addressed several of the shortcomings of the original model, including removing certain artifacts like water droplet-like effects and improving the consistency of the features across different resolutions. It improved by revising the normalization process, introducing weight demodulation to replace AdaIN: $w' = \frac{w}{\sqrt{\sum w^2 + \epsilon}}$. This version also refined the style mixing abilities, allowing even more nuanced control over the generated outputs, enhancing the model's capacity to produce highly realistic and varied facial expressions and identities. StyleGAN3 [11] was developed to tackle the issue of translational and rotational equivariance. The transformational equivariance is achieved by ensuring that the learned features F in the network transform predictably under rotations and translations, which can be mathematically represented as: $F\big(G(x)\big) = G(F(x))$, where g represents a geometric transformation applied to the input x. This improvement was crucial for applications requiring consistent facial orientation, such as animated characters.

Building upon these foundational advancements, several notable variants have emerged: StyleGAN-XL [12] further pushes the capabilities in handling extremely large datasets and enhancing the diversity of the generated images. This variant is particularly adept at scaling up the training process without compromising the detail and quality of the generated faces. HyperStyle [13] uses a hypernetwork approach where an external network generates parameters for the primary StyleGAN model, dynamically adjusting styles based on additional input conditions, modeled as: $H(x;\theta) \rightarrow \{w^{'} \mid \theta \in \Theta\}$. Here, H represents the hypernetwork function parameterized by θ, producing style parameters $w^{'}$ that directly influence the primary StyleGAN generator, allowing dynamic style manipulation without retraining the network. This method makes it possible to adapt the synthesis process according to specific requirements without retraining the network. StyleRig [14] combines StyleGAN with a 3D morphable model, adding rig-like control over the generated faces. This variant is particularly useful for animation and VR applications. StyleFlow [15] uses conditional continuous normalizing flows in the StyleGAN latent space to allow attribute-conditioned traversal, making changes such as altering age or adding accessories without affecting other intrinsic attributes like identity.

2.3. Face generation based on diffusion model

Diffusion models have rapidly ascended to prominence in the field of generative models, proving highly effective for tasks such as face generation. Compared to traditional GANs, diffusion models introduce a continuous noise process that allows for fine control over the quality of generated images, reducing instability and mode collapse during training. Diffusion models based on single-modal information [16, 17], such as textual descriptions, have already developed powerful capabilities and they are commercially viable. Recently, models that can exploit complex information have emerged. Composable Diffusion [18] is an innovative approach that enables the combination of multiple instances of diffusion models to enhance the capability of generating complex images conditioned on various inputs. It typically employs the same text-to-image diffusion model in a layered manner, allowing for the composition of visual elements in a stepwise and controlled manner. The model operates by integrating intermediate results from different stages or conditions, which makes it particularly useful for scenarios where a clear layering or staging of visual elements is beneficial, such as in detailed scene constructions or complex character designs. However, Composable Diffusion relies on the staging of multiple diffusion processes, which adds significant complexity to the model's architecture and training procedures. In contrast, The Collaborative Diffusion [19] framework represents a significant evolution in multi-modal image synthesis and editing. It integrates pretrained uni-modal diffusion models to work in parallel, leveraging their individual strengths without the need for re-training. This is achieved through a novel component called the "dynamic diffuser," which adaptively modulates the influence of each uni-modal model based on the specific requirements of the multimodal input conditions. Though dynamic diffuser could predict an appropriate weight to each uni-modal, the information from different modal could not be more fully integrated.

3. Discussion

The exploration of face generation through various generative models, including vector quantized methods, GANs, and diffusion models, reveals a rapidly advancing field. Each class of models brings unique strengths to the challenges of creating realistic, diverse, and controllable facial images. Several key directions emerge for further research and development in this area.

3.1. Enhanced Multimodal Capabilities

As seen in models like Composable and Collaborative Diffusion, the integration of multimodal inputs (e.g., text, sketches, existing images) holds significant promise. While current models are exploring methods to exploit multimodal information, there still exists techniques like contrastive learning to integrate information better. Future work could focus on enhancing the synergy between different modalities to create more coherent and contextually accurate facial images. Research could also explore deeper into unsupervised or self-supervised methods that learn to leverage multimodal data without extensive labeled datasets.

3.2. Improving Model Accessibility and Efficiency

While models like StyleGAN-XL and Imagen show exceptional capabilities, they often require substantial computational resources. Future developments could focus on making these models more accessible by optimizing their architectures or utilizing pre-trained models for efficiency without sacrificing output quality. Reducing computational demands of training and deploying these sophisticated models is a priority research direction.

3.3. Addressing Bias and Fairness

Bias in AI-generated faces remains a concern, as models often perpetuate biases present in their training data. Future research should prioritize the development of algorithms that can identify and reduce biases, ensuring that face generation models produce fair representations of all human features. This involves not only technical improvements, but also careful consideration of the data used for training these models.

3.4. Regulatory and Ethical Considerations

As face generation technology becomes more widespread, it also raises ethical concerns, particularly in the contexts of deepfakes. Future research should include the development of regulatory frameworks and ethical guidelines to manage the use of face generation technologies. Additionally, technology that can detect synthetic images could be an important area of development.

4. Conclusion

This paper has provided a comprehensive review of the advancements in face generation technologies, focusing on three primary classes of models: vector quantized methods, GAN-based methods, and diffusion models. The technical challenges discussed include training stability, scalability, control over generated attributes, and efficient integration of multimodal inputs. Addressing these challenges is essential for improving the robustness and applicability of face generation models. Future research should focus on refining model architectures to enhance efficiency, optimizing training processes to ensure stability, and developing techniques to better handle diverse and complex input conditions. Besides, the growing impact of face generation technologies in practical applications should be emphasized. The ethical considerations, including privacy concerns and potential misuse of generated faces, must be addressed. Ensuring that these technologies are used responsibly is crucial as they become more integrated into everyday applications.

References

[1] Van Den Oord, A., & Vinyals, O. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, *30*.

[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144.

[3] Jascha, S. D., Eric, W., Niru, M., and Surya, G. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pages 2256–2265, Lille, France, PMLR.

[4] Jonathan, H., Ajay, J., and Pieter, A. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33: 6840–6851.
[5] Yang, S., and Stefano, E. (2019). Generative modeling by estimating gradients of the data distribution. Advances in Neural Information Processing Systems, 32.
[6] Kingma, D. P. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
[7] Huang, Z., Chan, K. C., Jiang, Y., & Liu, Z. (2023). Collaborative diffusion for multi-modal face generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6080-6090).
[8] Esser, P., Rombach, R., & Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12873-12883).
[9] Chang, H., Zhang, H., Jiang, L., Liu, C., & Freeman, W. T. (2022). Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11315-11325).
[10] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401-4410).
[11] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8110-8119).
[12] Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-free generative adversarial networks. *Advances in neural information processing systems*, *34*, 852-863.
[13] Sauer, A., Schwarz, K., & Geiger, A. (2022). Stylegan-xl: Scaling stylegan to large diverse datasets. In ACM SIGGRAPH 2022 conference proceedings (pp. 1-10).
[14] Alaluf, Y., Tov, O., Mokady, R., Gal, R., & Bermano, A. (2022). Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In Proceedings of the IEEE/CVF conference on computer Vision and pattern recognition (pp. 18511-18521).
[15] Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H. P., Pérez, P., ... & Theobalt, C. (2020). Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6142-6151).
[16] Abdal, R., Zhu, P., Mitra, N. J., & Wonka, P. (2021). Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, *40*(3), 1-21.
[17] Rombach, R., et al. (2022). High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.
[18] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., ... & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, *35*, 36479-36494.
[19] Liu, N., Li, S., Du, Y., Torralba, A., & Tenenbaum, J. B. (2022). Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision* (pp. 423-439). Cham: Springer Nature Switzerland.
[20] Huang, Z., Chan, K. C., Jiang, Y., & Liu, Z. (2023). Collaborative diffusion for multi-modal face generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6080-6090).