An Investigation of Face Generation Methods Based on Encoders, GAN and Diffusion models

Timothy Tianlin Zhang*

Computer Science, University of British Columbia, V6T1Z4 Vancouver, Canada

Abstract. Face generation from natural language input has rapidly emerged as a pivotal research area in computer vision, bridging the gap between creative design and practical commercial applications. This review paper synthesizes findings from ten seminal works to map the methodological evolution of text-to-face generation, tracing the progression from encoder-decoder architectures and Generative Adversarial Networks (GANs) to the current state-of-the-art dominated by diffusion models. This paper examines how these paradigms address core challenges such as generation fidelity, controllability, and cross-modal alignment. The analysis reveals that while early GAN-based approaches pioneered conditional control, modern diffusion models, often integrated with autoencoders, offer superior stability and detail. Furthermore, this paper explores extensions into multi-modal conditioning using audio and pose, as well as applications in synthetic data generation for privacy-preserving facial recognition. Despite significant advances, critical challenges persist, including ensuring precise semantic alignment, maintaining identity across edits, achieving temporal consistency, and mitigating ethical risks associated with deepfakes. This review concludes by identifying key trends and outlining promising future directions for developing more robust, efficient, and ethically sound face generation systems.

1 Introduction

Face generation based on natural language input has rapidly emerged as an important research direction within computer vision. Advances in computational hardware and generative modeling now make it possible to perform graphic generation tasks that were previously unattainable. From producing anime- or comic-style illustrations for narrative storytelling, to adjusting model faces in online retail platforms such as clothing marketplaces, text-to-face generation illustrates how artificial intelligence can bridge creative design, personalization, and commercial applications.

The current progress in generative modeling is driven primarily by diffusion models, autoencoders, and related approaches [1, 2]. In parallel, research on improved autoencoder architectures provides robust latent spaces where faces can be represented compactly without sacrificing reconstruction quality [3]. Diffusion models have become the leading framework for high-fidelity image synthesis, offering stable training, wide mode coverage, and strong

* timmyz@student.ubc.ca

controllability compared with earlier paradigms such as Generative Adversarial Networks (GANs) [4]. Their combination with latent autoencoders has significantly reduced the computational burden of training and sampling, enabling practical generation at high resolutions while preserving facial identity and fine-grained details, as seen in models like Lynx for personalized video generation [1]. Complementary directions, including GAN-based hybrids, 3D-aware models, and consistency or flow-based formulations, further enrich the design space. For instance, frameworks like AnimPortrait3D enable the generation of high-quality, animatable 3D avatars from text, ensuring multi-view consistency . Other innovations include the use of landmark assistance and heatmap-guided denoising losses to achieve high-fidelity controllable facial generation .

Despite these advances, challenges remain. Ensuring semantic alignment between prompts and outputs, achieving reliable identity preservation during editing, maintaining multi-view and temporal consistency, and delivering results at interactive speeds are active research problems. At the same time, the widespread availability of face generation technologies raises questions of security, authenticity, and ethics, motivating the development of watermarking, provenance tracking, and deepfake detection systems.

The remainder of this paper is organized as follows. Section 2 (Methods) reviews the methodological approaches adopted in the ten selected papers, with emphasis on architectural design choices, conditioning strategies, and evaluation protocols. Section 3 (Discussion) highlights the different methodologies' strengths and weaknesses. Finally, Section 4 (Conclusion) summarizes the key insights from this review and outlines promising directions for future research.

2 Method

The field of face generation has evolved through several key architectural paradigms, each building upon the last. This section reviews the methodological approaches in the selected papers, beginning with the foundational encoder-decoder architectures, moving to Generative Adversarial Networks (GANs) and their conditional variants, and concluding with the current state-of-the-art represented by diffusion models.

2.1 Encoder-Decoder Architectures

O. Wiles et al. presents a unified framework for controlling face generation through multiple modalities including images, audio, and pose codes [1]. Based on an encoder-decoder architecture, the system can generate facial animations driven by different input types. The encoder processes reference images to create embedded representations, while the decoder generates output frames conditioned on control signals. This flexible approach allows for cross-modal face animation, where for example audio input can drive facial movements. The architecture demonstrates how shared latent spaces can enable unified control of facial generation across different input modalities, paving the way for more adaptable and controllable face animation systems.

2.2 Generative Adversarial Networks (GANs)

GANs revolutionized image generation by framing it as an adversarial game between a generator and a discriminator, leading to unprecedented levels of realism.

This pioneering work by M. Mirza. et al. explores conditional GANs for face generation, where the generative process is guided by specific conditions or attributes [2]. Unlike unconditional GANs that generate random faces, conditional GANs can produce faces with

predefined characteristics such as age, gender, or expression. The convolutional architecture enables the generation of high-resolution faces through hierarchical feature learning. Despite being surpassed by more recent diffusion approaches in terms of image quality and training stability, this work laid important groundwork for controlled face generation.

K. K. Singh et al. addresses the challenging task of generating faces from detailed textual descriptions [3]. Using a GAN framework conditioned on textual embeddings, the model can generate faces that match fine-grained descriptions such as "a young Asian male with curly black hair, narrow eyes, and a wide nose." This enables semantic face manipulation through textual input, allowing precise control over generated facial attributes. Despite the challenges of capturing the complex relationship between language and visual features, Text2FaceGAN demonstrates the potential for cross-modal face generation from textual descriptions.

L. Li et al. presents an approach that combines attribute guidance with CycleGAN architecture for face generation and enhancement [4]. The method takes a low-resolution image and an attribute vector as input to generate a high-resolution face with desired characteristics. The conditional CycleGAN framework employs cycle-consistency loss to preserve identity and other unrelated attributes during the transformation process. This enables applications such as face super-resolution with controlled attribute manipulation, such as enhancing image quality while simultaneously modifying age or expression.

2.3 Diffusion Models

Diffusion models have emerged as the leading framework for high-fidelity image synthesis, offering stable training and strong controllability. They work by iteratively denoising a random noise vector to generate a coherent image.

W. Lv et al. proposes a framework that combines uni-modal models into a multi-modal diffusion model called "Collaborative Diffusion" [5]. The purpose of this framework is to leverage multiple input modalities (e.g., text and an image) to produce more accurate face generations without the cost of retraining a single, combined model. The key component is a "dynamic diffuser" that determines how much weight to give to each pre-trained uni-modal model based on spatial and temporal data during the denoising process.

M. Kim et al. attempts to eliminate the use of real face photos for training Face Recognition models due to privacy concerns [6]. The proposed Dual Condition Face Dataset Generator includes two stages. In the first stage, a diffusion-based generator creates a high-fidelity identity image. In stage two, this image is blended with the stylistic information (lighting, shadows) from a "style photo" using another diffusion model. A key innovation is the use of a patchwise style extractor to ensure no identity information is leaked from the style image.

M. S. Sheikh et al. proposes "Diffused Heads" to generate audio-driven facial animation from a single image, outperforming traditional GAN-based approaches [7]. To resolve temporal jitter (unnatural sudden movements), the team introduced "Motion Frames," which use immediately preceding frames as conditioning to maintain temporal momentum. To prevent the accumulation of color errors in an autoregressive frame-by-frame generation setting, the model generates only grayscale frames, which are later fused with a high-fidelity identity frame to produce the final color video.

R. Wang et al. designs a diffusion-based framework for face animation that differs from autoregressive models [8]. To maintain motion consistency and avoid error accumulation, the team first uses an autoencoder to produce a low-level latent vector that serves as a stable, blurry blueprint for all diffusion steps. Another key innovation is the Attribute Guided Conditioning Network (AGCN), which processes the inputs (e.g., identity image and driving signals) to intelligently determine how much conditioning weight to assign to each input for the diffusion model.

Z. Zhang et al. represents a sophisticated approach by combining diffusion models with autoencoders [9]. The framework employs a Denoising Diffusion Implicit Model (DDIM) autoencoder to first encode faces into latent factors. A conformer architecture—which combines convolutional neural networks with transformers—then maps audio features to these latent factors. Finally, the DDIM decoder reconstructs high-fidelity facial images from the predicted latents. This approach effectively bridges the audio-visual correlation gap, generating realistic lip synchronization and expressive facial animations.

H. Park et al. addresses synthetic face recognition using innovative diffusion-based approaches [10]. The key innovation involves "fuzzy" identity conditioning, which generates variations of faces while preserving essential identity characteristics. This creates diverse training data that improves recognition model robustness. The method achieves a remarkable 98% accuracy on the challenging Labeled Faces in the Wild (LFW) benchmark, significantly outperforming other synthetic-based solutions and representing substantial progress toward viable, privacy-conscious synthetic data alternatives.

3 Comparative Discussion

The evolution of face generation technologies has progressed through several distinct phases, each with characteristic strengths and limitations. Generative Adversarial Networks (GANs) represented a breakthrough in synthetic image quality, enabling unprecedented realism in generated faces. Their adversarial training process allows for high-fidelity synthesis and sharp image quality. However, GANs suffer from training instability, mode collapse issues, and limited diversity in outputs. Additionally, GAN-based approaches often struggle with precise attribute control and can produce artifacts when manipulating specific facial characteristics .

Autoencoders and particularly variational autoencoders (VAEs) offer advantages in learning structured latent spaces and enabling more controlled generation through manipulation of latent codes. Their encoder-decoder architecture provides a framework for compressed representation learning and attribute disentanglement. However, autoencoder-based approaches typically generate blurrier imagescompared to GANs and often lack the fine-grained texture details necessary for photorealistic face generation .

Diffusion models represent the current state-of-the-art, combining the best attributes of previous approaches while addressing many of their limitations. Diffusion models excel at generating high-fidelity details and diverse outputs while maintaining stable training characteristics. Their iterative refinement process allows for precise control and high-quality synthesis, albeit at increased computational cost during inference. Recent advances in distillation techniques and accelerated sampling algorithms are addressing these efficiency concerns .

The future of face generation technology points toward increasingly unified architectures capable of handling multiple modalities and tasks within single frameworks. Approaches like Collaborative Diffusion and UniDiffuser demonstrate how different modalities can be combined in flexible ways without retraining. The integration of reinforcement learning with diffusion models opens new possibilities for optimizing generated faces toward specific objectives or rewards. As these technologies continue to advance, it can be expected that further improvements in efficiency, controllability, and accessibilityof face generation systems.

4 Conclusion

This review has charted the significant evolution of text-to-face generation, highlighting a clear trajectory from foundational encoder-decoder models and GANs to the current dominance of diffusion-based approaches. The analysis demonstrates that diffusion models, particularly when hybridized with autoencoders, provide a powerful framework for achieving high-fidelity, controllable, and diverse facial synthesis. They have effectively addressed many limitations of earlier models, such as training instability and limited mode coverage. Looking forward, the field must tackle persistent challenges in semantic precision, identity preservation, and computational efficiency. The integration of 3D awareness and the development of robust ethical safeguards, including provenance tracking and deepfake detection, will be critical. Future progress hinges on creating unified, multi-modal systems that are not only more powerful and efficient but also transparent and responsible, ensuring the technology's positive impact across creative and commercial domains.

References

1. O. Wiles, A. S. Koepke, A. Zisserman, X2Face: A network for controlling face generation using images, audio, and pose codes. In Proceedings of the European Conference on Computer Vision (ECCV) (2018)
2. M. Mirza, S. Osindero, Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
3. K. K. Singh, A. S. Yamaha, S. K. P., Text2FaceGAN: Face Generation from Fine Grained Textual Descriptions. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (2019)
4. L. Li, H. Lin, Attribute-Guided Face Generation Using Conditional CycleGAN. In Proceedings of the Asian Conference on Computer Vision (ACCV) (2018)
5. W. Lv, et al., Collaborative Diffusion for Multi-Modal Face Generation and Editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
6. M. Kim, et al., DCFace: Synthetic Face Generation with Dual Condition Diffusion Model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
7. M. S. Sheikh, et al., Diffused Heads: Diffusion Models Beat GANs on Talking-Face Generation. In Proceedings of the International Conference on Learning Representations (ICLR) (2024)
8. R. Wang, et al., Face Animation with an Attribute-Guided Diffusion Model. In Proceedings of the AAAI Conference on Artificial Intelligence (2023)
9. Z. Zhang, et al., DAE-Talker: High Fidelity Speech-Driven Talking Face Generation with Diffusion Autoencoder. arXiv preprint arXiv:2303.17550 (2023)
10. H. Park, et al., iDiff-Face: Synthetic-based Face Recognition through Fuzzy Identity-Conditioned Diffusion Models. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2024)