Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Advances in Integrating GANs and NeRF for Image Generation and 3D Reconstruction

Shaoyang Qu *

Artificial Intelligence Institute, Tianjin Normal University, Tianjin, China

* Corresponding author: 2130330009@stu.tjnu.edu.cn

Abstract. This paper investigates recent advancements in integrating Generative Adversarial Networks (GAN) with Neural Radiance Fields (NeRF) for image generation and 3D reconstruction. This integration is pivotal for significantly enhancing the quality of image and 3D scene generation across diverse applications, including medical imaging, virtual reality, animation, and film production. The objective of this study is to summarize and compare existing methods, providing both theoretical and methodological insights. A detailed analysis of various methods is conducted, focusing on their strengths and weaknesses in terms of image quality, multi-view consistency, style migration, and texture transfer. Experimental results across different datasets reveal that while these methods perform effectively within their respective application contexts, which generally face challenges related to high computational resource consumption, training complexity, and the need for extensive multi-view data. Despite these issues, significant progress has been made in enhancing image quality and 3D consistency. This study contributes valuable theoretical insights and offers practical guidance for future research. Future efforts should focus on developing more efficient algorithms, refining training techniques, and exploring applications in medical imaging, virtual reality, animation, and film production to advance image processing and computer vision technologies.

Keywords: Generative Adversarial Networks (GAN); Neural Radiance Fields (NeRF); Image Generation; 3D Reconstruction.

1. Introduction

Nowadays, with the rapid development of computer vision, the Neural Radiation Field (NeRF) proposed by Ben Mildenhall et al. in 2020 has received more and more attention. NeRF is different from traditional 3D reconstruction [1], which utilizes the neural network's representation of a 3D scene to generate images of new perspectives through the body rendering. This is a relatively novel direction, but it lacks the detail and quality of the generated images. One method associated with this is the Generative Adversarial Network (GAN) [2], proposed by Ian Goodfellow et al. in 2014 as a way to train two neural networks through an adversarial process. It has good performance in image generation and image rubbing walls. The combination of GAN with NeRF is a novel approach to improving the quality of images generated by NeRF, and there have been many studies employing GAN networks and NeRF for other unique tasks. This review argues that this novel research needs a systematic induction and summary to compare the similarities and differences between them and effectively help future research in this direction.

As proposed by Erik Chan et al. in 2021, Representing Scenes as Compositional Generative Neural Feature Fields (GIRAFFE) achieves high-quality and controllable image synthesis by generating an adversarial network training method that utilizes generated neural feature fields to represent multiple objects [3]. In contrast, Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis (pi-GAN) has superior detail representation [4]. Whereas, A Style-based 3D-aware genera tor for high-resolution image synthesis (StyleNeRF) combines NeRF with a style generator using low-resolution feature maps for body rendering and then gradually sampling to high resolution over 2D space [5], the GAN network generates high-quality images. Geometry-enhanced Novel View Synthesis from Single-View Images (G-NeRF) proposed by Zixiong Huang in 2024 implements the generation of high-quality new-view images from single-view images [6]. In addition to this,

Boosting Novel View Synthesis for Monocular Images using Residual Encoders (ReE3D) has improved the quality and generation efficiency of new view synthesis of monocular images [7]. Next, there is Reference-Based Stylized Radiance Fields for Dynamic Scenes (S-DyRF) which generates high-quality dynamic scenes through the style transformation module of the GAN network [8]. In contrast, Achieve High 3D Consistency and Temporal Coherence for Face Video Editing on Dynamic NeRF (FED-NeRF) achieves high-quality facial video editing [9]. Elsewhere, Deformable Neural Radiance Fields for 3D Toonification (DEFORMTOON3D) injects geometric information into a pre-trained 3D GAN generator to achieve geometric stylization [10]. A Stylized Neural Field for 3D-Aware Cartoonized Face Synthesis (artNeRF) utilizes a contrast-learning style encoder to generate 3D-aware cartoonish faces with arbitrary styles [11]. There are also, for example, Generative Detail Compensation via GAN and Diffusion for One-shot Generalizable Neural Radiance Fields (GD2-NeRF) for high-quality new-view image generation through GAN and diffusion modeling for detail compensation [12], and Generative Radiance Fields Restoration (RaFE) [13], which excels in handling degraded images and detail recovery.

The main objective of this study is to summarize the general direction of combining GAN and NeRF, highlighting the current advanced methods and their effectiveness. By comparing results and data from existing literature on the combination of GAN networks and NeRF, this study aims to provide useful theoretical and methodological guidance for future research in this field. The unique 3D reconstruction method of NeRF and its operational principles are investigated, and the model results are thoroughly analyzed. Additionally, the study explores GAN networks, detailing their functionality and key features in image generation. Finally, a comparative analysis of the literature on combining GAN and NeRF is conducted to evaluate the methodologies and effectiveness of applying GAN networks in NeRF. The study concludes that there is significant potential for further research and improvement in combining GAN networks with NeRF. Future research can focus on optimizing these combination methods, enhancing training and inference efficiency, expanding application areas, and addressing practical challenges to advance image processing and computer vision technology.

This article carries out a summary, comparison, and analysis of research related to the combination of GAN and NeRF. Section 2 analyzes the basic working principles of NeRF and GAN and presents one of the more classical models from the extensive literature. Section 3 compares data, and effects and discusses the collected studies. Finally, Section 4 summarizes.

2. Methodology

2.1. Dataset Description and Preprocessing

The main datasets used now are the Blender Synthetic Dataset, Realistic Synthetic Indoor Dataset (RSID), Local Light Field Fusion (LLFF) Dataset, COCO Dataset, and Kaggle Datasets [14]. Blender Synthetic Dataset content contains 2D images and camera parameters from multiple viewpoints. It is mainly used for training and evaluating 3D scene reconstruction algorithms. RSID Content: contains high-quality synthetic indoor scene images. Its use is for training 3D reconstruction models of indoor scenes. LLFF Dataset contains multi-view photos of indoor and outdoor scenes. Its use is for training and evaluating view synthesis and 3D reconstruction algorithms. COCO Dataset contains multiple classes of 2D images and their annotation information. It is used for tasks such as image generation and style migration. Kaggle Datasets which has image datasets on various topics. It is mainly used for functions like image generation, style migration, 3D modeling, etc. These datasets usually require some preprocessing before use, such as image normalization, data enhancement, and camera parameter parsing, to ensure that the model can effectively learn and generate high-quality 3D images or scenes.

2.2. Proposed Approach

The study begins by examining NeRF's unique approach to 3D reconstruction, its operational principles, and model results. It identifies significant shortcomings and areas for improvement in

image generation. Simultaneously, the paper explores GANs, highlighting how they can effectively address NeRF's limitations and enhance its development in other aspects. The research includes a comparative analysis of the literature on the combination of GANs and NeRF, discussing the methodologies and effectiveness of applying GAN networks in NeRF. Through this analysis, the paper aims to draw meaningful conclusions and offer insights for future research, ultimately advancing the fields of image processing and computer vision. The pipeline is shown in the Fig. 1.

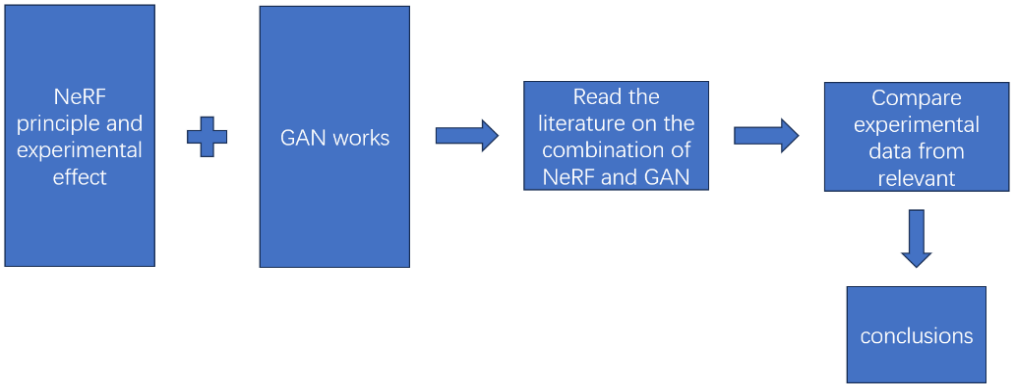

Figure 1. Research process

2.2.1. NeRF.

NeRF is a method of representing a 3D scene using neural networks to generate images of new perspectives through a body rendering technique. The NeRF model calculates the XYZ position of a sampling point by inputting the camera's position and combining it with the pixel points and the camera's internal reference matrix, together with the camera's position as an input into Multilayer perceptron (MLP). Finally, outputs are RGB value and opacity. The model is shown in Fig. 2. The core innovation of NeRF lies in its position encoding and hierarchical body rendering methods.

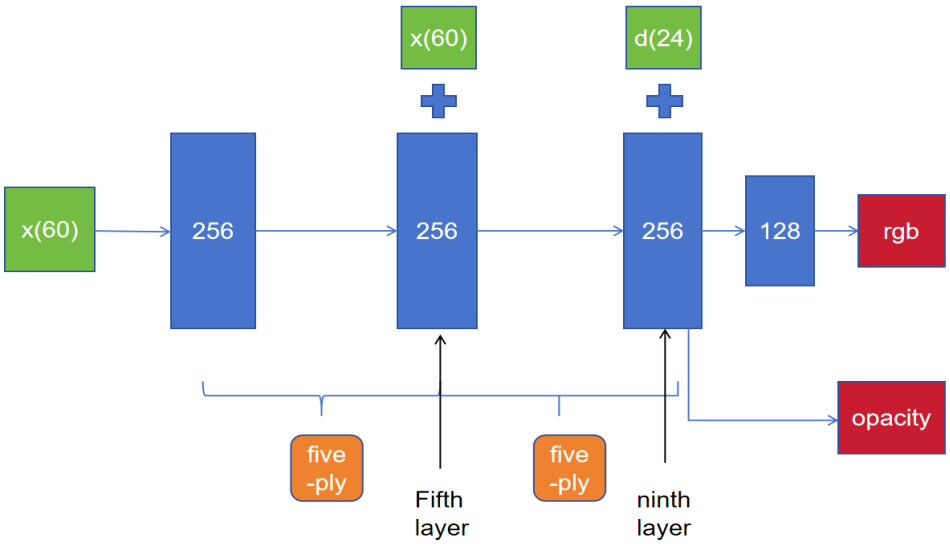

Figure 2. Structure of the NeRF model

First of all, NeRF introduces a position encoding technique. The main purpose of this technique is to prevent the generated image from being too smooth and losing detailed information. This process effectively transforms low-frequency spatial information into high-frequency features, enabling NeRF to generate more detailed and realistic images. Second, NeRF employs a hierarchical body rendering method, i.e., sampling through both coarse and fine networks, to improve the accuracy of image generation. Specifically, the coarse network first performs coarse sampling over the whole space to generate preliminary volume rendering results; then, the fine network performs denser sampling of the region of interest based on the coarse network. This dual-network architecture not only improves computational efficiency but also significantly enhances image detail and realism.

With this approach (See in Fig. 3), NeRF can generate high-quality 3D reconstructions and new-view images in a shorter period. In addition, NeRF is designed with computational resource constraints in mind. While positional coding and hierarchical body rendering methods increase computational complexity, the introduction of these methods allows NeRF to optimize the use of computational resources while maintaining high accuracy. Overall, by introducing positional coding and hierarchical body rendering methods, NeRF effectively solves the problems of overly smooth image generation and insufficient computational resources in traditional methods and significantly improves the accuracy and efficiency of 3D image generation. These innovations have led to the widespread attention and application of NeRF in both academia and industry, providing new ideas and methods for the future development of 3D image generation technology.

Figure 3. Images after NeRF training

2.2.2. GAN.

GAN which proposed by Ian Goodfellow et al. in 2014, is a new framework for estimating generative models through an adversarial process. The framework trains two models simultaneously: a generative model G and a discriminative model D. The generative model G generates realistic data samples. In contrast, the discriminative model D tries to distinguish whether these samples come from real data or the generative model. The training goal is to make the samples generated by the generative model able to deceive the discriminative model, making it difficult for it to distinguish between real and fake samples. Mathematically, a GAN achieves this adversarial training through a tiny very large game, where the generative model G minimizes the probability that the discriminative model D correctly identifies the generated samples, while the discriminative model D maximizes the probability of distinguishing between real samples and generated samples. The main advantages of a GAN include the ability to generate samples with only forward propagation without the need for Markov chains or approximate inference networks. This simplifies the training process and allows the GAN to demonstrate its potential for generating high-quality samples on multiple datasets. Experimental results show that GAN-generated samples are visually competitive. Although GAN needs to synchronize discriminative and generative models during the training process and lacks diversity in generating samples, it provides an effective framework for generating high-quality samples. In addition, GAN can be extended to the field of conditional generative modeling and semi-supervised learning, e.g., by adding conditional variables to the generative model G and the discriminant model D to generate samples under specific conditions, or by utilizing the features of the discriminators to improve the performance of the classifiers. In conclusion, GAN is a powerful generative modeling framework with a wide range of applications.

2.2.3. G-NeRF.

G-NeRF is a novel view synthesis method designed to generate high-fidelity new view images from single-view images, combining the powerful 3D modeling and rendering capabilities of Nerf. Nerf provides an implicit representation of a 3D scene via a deep neural network, which efficiently models the geometry and appearance of the scene, enabling it to generate high-quality and realistic 3D rendered images observed from any viewpoint. However, traditional NeRF models have significant limitations when dealing with single-view data, requiring multi-view photos for training. In contrast, often only single-view images are available in real-world applications.

To overcome these limitations, G-NeRF employs two innovative approaches, Geometry-Guided Multi-View Synthesis (GMVS) and Depth-Aware Training (DAT). First, GMVS utilizes pre-trained 3D GAN models (e.g., EG3D) to generate multi-view data from single-view images to provide a rich geometric prior. To balance diversity and geometric quality, G-NeRF introduces a truncation

approach to create high-quality geometric data by sampling latent codes in the latent space. Secondly, G-NeRF devises a depth-aware training method (DaT) that introduces a depth-aware discriminator to enhance the geometric fidelity of the generated results by comparing the generated depth maps with the true depth maps to provide additional geometric supervision.

The G-NeRF model consists of three main components (See in Fig. 4): the scene encoder, the NeRF-based generator, and the depth-aware discriminator. The scene encoder adopts a ResNeXt structure for extracting features from the input image; the NeRF-based generator includes a pre-trained 3D GAN model and a super-resolution module for generating multi-view data and corresponding depth maps; and the depth-aware discriminator is used to differentiate between the generated depth maps and the real depth maps, providing geometric supervision. Combining these innovative approaches, G-NeRF performs well on multiple real-world datasets, successfully solves the challenge of generating new views from single-view images, and significantly improves the quality and geometric consistency of the generated images.

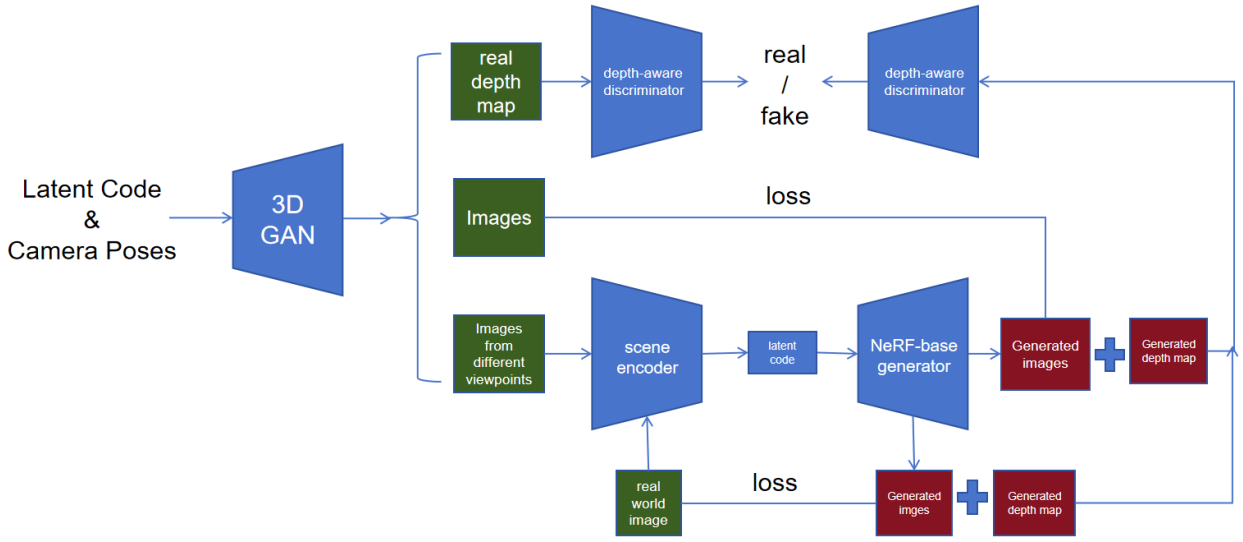

Figure 4. G-NeRF Modeling Architecture

G-NeRF overcomes the limitations of the traditional NeRF model in processing single-view data and realizes the goal of high-quality new view synthesis on single-view images. Experimental results show that G-NeRF performs well on multiple real-world datasets, successfully solves the problem of generating new views on single-view images, and significantly improves the quality and geometrical consistency of the generated images. The innovative approach of G-NeRF provides new ideas and technical support for 3D modeling and rendering of single-view images and demonstrates a wide range of application prospects.

3. Result and Discussion

This chapter categorizes the main current research areas and research methods into three areas by analyzing the literature on the combination of GAN and NeRF. Based on the well-sorted model types, the corresponding data are compared to analyze the superiority of each and summarize the shortcomings and future research directions.

3.1. Image Generation Quality and Multi-view Consistency

As show in Table 1. A literature search reveals that GIRAFFE achieves high-quality and controllable image synthesis through a generative adversarial network training approach that utilizes generated neural feature fields to represent multiple objects. Experimental results show that it has excellent Fréchet Inception Distance (FID), Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM) metrics in various datasets with fast generation speed and high scene complexity. In contrast, pi-GAN improves detail representation and multi-view consistency by introducing periodic activation

functions and body rendering techniques. The results show that pi-GAN outperforms existing methods regarding image quality and multi-view consistency metrics on multiple datasets. Then there is StyleNeRF which combines NeRF with a style generator to use low-resolution feature maps for body rendering, and then gradually samples to high resolution over 2D space. styleNeRF significantly outperforms the traditional NeRF model in terms of style consistency and detail reproduction over multiple datasets. Compared to GIRAFFE and pi-GAN, StyleNeRF has advantages in stylistic consistency and detail reduction of the generated images but is slightly inferior in scene complexity and controllability. Finally, G-NeRF generates multi-view images through pre-trained 3D GAN models to achieve high-quality new-view images from single-view images. G-NeRF excels in 3D consistency and image quality for single-view image generation tasks. However, G-NeRF is not as flexible as GIRAFFE in dealing with complex scenes. ReE3D improves the quality of new view synthesis and the generation efficiency of monocular images by the residual encoder and 3D GAN. Experimental results show that ReE3D performs well in image reconstruction quality and new perspective synthesis tasks while processing time is significantly reduced. Compared with other literature, ReE3D has an advantage in processing time but is slightly inferior to pi-GAN and StyleNeRF in multi-view consistency and detail representation.

Table 1. Image quality comparison

literature	Data set	FID	PSNR	SSIM
GIRAFFE	CelebA-HQ	21	-	-
	FFHQ	32	-	-
	Cars	26	-	-
	Churches	30	-	-
	Clevr-2	31	-	-
pi-GAN	CelebA	14.7	-	-
	Cats	16.8	-	-
	CARLA	29.2	-	-
StyleNeRF	FFHQ	8	-	-
	AFHQ	14	-	-
	CompCars	8	-	-
G-NeRF	ShapeNet Cars	-	22.31	0.88
	ShapeNet Chairs	-	21.03	0.86
ReE3D	CelebA	-	19.9102	0.5305

3.2. Dynamic Scene Processing

S-DyRF applies the styles of reference images to the radial field generation of dynamic scenes and generates high-quality dynamic scenes through the style transformation module of a GAN network. In contrast, FED-NeRF utilizes the dynamic NeRF structure to achieve high-quality facial video editing that maintains multi-view consistency and temporal coherence. Experimental results show that FED-NeRF excels in 3D view consistency and temporal consistency, while S-DyRF has an advantage in generating stylized dynamic scenes.

DyRF and FED-NeRF have their advantages: S-DyRF generates stylized dynamic scenes through the style conversion module, which is suitable for application scenarios that require stylistic consistency, while FED-NeRF performs well in facial video editing with high 3D view consistency and temporal consistency, which is suitable for tasks that require high-quality dynamic scene reconstruction.

3.3. Style Migration and Texture Migration

StyleNeRF combines NeRF and a style generator to improve the fidelity of the generated images through an adversarial network, showing excellent style consistency and detail recovery. In the experiments, StyleNeRF achieves a FID score of 8 on the FFHQ and CompCars datasets, and a FID

score of 14 on the AFHQ dataset. In contrast, DEFORMTOON3D achieves geometric stylization by introducing the StyleField module, which injects geometric information into the pre-trained 3D GAN generator and performs well in identity preservation and image quality. Its FID score of 27.6 and identity preservation (IP) score of 0.781 significantly outperform CIPS-3D (FID: 50.6, IP: 0.681), E3DGE (FID: 34.0, IP: 0.707), and StyleGAN-NADA (FID: 59.3, IP: 0.535). All data is displayed in Table 2. ArtNeRF utilizes the contrasting learning style encoder to generate 3D-aware cartoonized faces with arbitrary styles, ensuring style consistency and identity preservation, achieving an FID score of 14.42. These models made significant progress in both style migration and texture migration, excelling in fidelity, identity retention, and image quality, respectively.

Table 2. Comparison of data related to style migration and texture migration

literature	Data set	FID	IP (Identity Preservation)
StyleNeRF	FFHQ	8	-
	AFHQ	14	-
	CompCars	8	-
DEFORMTOON3D	-	27.6	0.781
	CIPS-3D	50.6	0.681
	E3DGE	34.0	0.707
	StyleGAN-NADA	59.3	0.535
ArtNeRF	-	14.42	-

In terms of style migration, StyleNeRF achieves high-quality stylized image generation through adversarial networks, DEFORMTOON3D excels in geometric stylization and identity preservation, while ArtNeRF has superior style consistency and identity preservation in cartoonized face generation. In terms of texture migration, both StyleNeRF and DEFORMTOON3D perform well, with the former focusing on image fidelity and detail restoration, and the latter having an advantage in identity preservation and geometric stylization. ArtNeRF ensures that the generated cartoonized faces are highly consistent in terms of texture and style by using the contrast learning method.

3.4. Image Quality and 3D Consistency

GD2-NeRF achieves high-quality new view image generation through detail compensation combined with GAN and diffusion modeling and exhibits excellent 3D consistency and image quality on ShapeNet and Database Transaction Unit (DTU) datasets. For example, GD2-NeRF has high PSNR and SSIM scores on the ShapeNet dataset, showing the clarity and consistency of its generated images. On the DTU data set, GD2-NeRF also shows good recovery in detail. In contrast, RaFE combines GAN and diffusion modeling to deal with different types of degraded scenes to generate high-quality NeRF, significantly outperforming the existing methods. RaFE has a PSNR of 24.99 and an SSIM of 0.901 on the Blender dataset, demonstrating its superiority in high-quality image generation. In addition, RaFE also performs well on the LLFF data set with a PSNR of 23.85 and an SSIM of 0.752, demonstrating its consistency and adaptability across different datasets. Both are shown in Table 3. Overall, these models have made significant progress in both image quality and 3D consistency, demonstrating great potential for detail restoration and image quality enhancement. GD2-NeRF and RaFE both achieve excellent results in their respective datasets and evaluation metrics, respectively, proving their superior performance in new view image generation.

Table 3. Comparison of image quality and 3D consistency. FFHQ, AFHQ represents Flickr-Faces-High-Quality, Animal Faces-HQ.

literature	Dataset	FID	IP (Identity Preservation)
StyleNeRF	FFHQ	8	-
	AFHQ	14	-
	CompCars	8	-
DEFORMTOON3D	-	27.6	0.781
	CIPS-3D	50.6	0.681
	E3DGE	34.0	0.707
	StyleGAN-NADA	59.3	0.535
ArtNeRF	-	14.42	-

3.5. Discussion

Research methods combining GAN and NeRF each have their strengths and weaknesses. NeRF is renowned for its high-quality 3D reconstruction, positional coding, and hierarchical body rendering but demands multi-view data and significant computational resources. Conversely, GAN excels in high-quality image generation but is complex to train and often lacks sample diversity. Specific models have demonstrated various advantages: GIRAFFE, pi-GAN, StyleNeRF, and G-NeRF are noted for their detail performance and multi-view consistency, while ReE3D stands out in generation efficiency. S-DyRF and FED-NeRF are exceptional in dynamic scene processing and 3D view consistency, and GD2-NeRF and RaFE excel in image quality and detail recovery. Future research should prioritize optimizing training and inference efficiency, expanding application areas, addressing the single-view data problem, and enhancing image details and realism. Current challenges include high computational resource consumption, complex training processes, and the need for multiple views. Solutions lie in optimizing algorithm structures, improving training methods, and developing single-view data generation techniques. These discussions and research directions offer robust theoretical and methodological guidance for advancing image processing and computer vision technology.

4. Conclusion

The purpose of this study is to summarize and compare the state-of-the-art methods for integrating GAN and NeRF technologies, offering both theoretical and methodological insights. This paper examines methods such as GIRAFFE, pi-GAN, StyleNeRF, G-NeRF, ReE3D, DyRF, FED-NeRF, GD2-NeRF, and RaFE, assessing their strengths and weaknesses in terms of detail performance, multiview consistency, generation efficiency, dynamic scene processing, and detail recovery. The results indicate that while these methods excel in their respective application scenarios, they are also hindered by high computational resource demands, training complexity, and a substantial need for multi-view data. Despite significant advances in improving image quality and 3D consistency through the synergy of GAN and NeRF, there remains considerable room for optimization in training and inference efficiency, as well as in expanding application areas and addressing single-view data challenges. Future research should focus on developing more efficient algorithms, enhancing training methodologies, and exploring applications in fields such as medical imaging, virtual reality, animation, and film production, to further advance image processing and computer vision technologies.

References

[1] Mildenhall B. et al. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 2021, 65 (1): 99 - 106.

[2] Creswell A. White T. Dumoulin V. et al. Generative adversarial networks: An overview. IEEE signal processing magazine, 2018, 35 (1): 53 - 65.

[3] Niemeyer M. et al. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields, 2020, arXiv preprint: 2011. 12100.

[4] Chan E.R. Monteiro M. Kellnhofer P. et al. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021: 5799 - 5809.

[5] Gu J.T, et al. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. 2021, arxiv preprint: 2110. 08985.

[6] Huang Z. Chen Q. Sun L. et al. G-NeRF: Geometry-enhanced Novel View Synthesis from Single-View Images. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 10117 - 10126.

[7] Guo K.H, et al. ReE3D: Boosting Novel View Synthesis for Monocular Images using Residual Encoders. IEEE Transactions on Multimedia, 2023.

[8] Li X. Cao Z. Wu Y. et al. S-DyRF: Reference-Based Stylized Radiance Fields for Dynamic Scenes. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 20102 - 20112.

[9] Zhang H. Yu W. and Chi K. FED-NeRF: Achieve High 3D Consistency and Temporal Coherence for Face Video Editing on Dynamic NeRF. 2024, arxiv preprint: 2401. 02616.

[10] Zhang J. Lan Y. Yang S. et al. Deformtoon3d: Deformable Neural Radiance Fields for 3d Toonification. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 9144 - 9154.

[11] Tang Z. and Hong Y. ArtNeRF: A Stylized Neural Field for 3D-Aware Cartoonized Face Synthesis. 2024, arxiv preprint: 2404.13711.

[12] Pan X. Yang Z. Bai S. et al. GD^ 2-NeRF: Generative Detail Compensation via GAN and Diffusion for One-shot Generalizable Neural Radiance Fields. 2024, arXiv preprint: 2401. 00616.

[13] Wu Z. Wan Z. Zhang J. et al. RaFE: Generative Radiance Fields Restoration. 2024, arXiv preprint: 2404. 03654.

[14] Li X. Zhang Q. Kang D. et al. Advances in 3d generation: A survey. 2024, arXiv preprint: 2401. 17807.