Portrait3D: Text-Guided High-Quality 3D Portrait Generation Using Pyramid Representation and GANs Prior

SIGGRAPH 2024

YIQIAN WU, State Key Lab of CAD&CG, Zhejiang University, China
HAO XU, State Key Lab of CAD&CG, Zhejiang University, China
XIANGJUN TANG, State Key Lab of CAD&CG, Zhejiang University, China
XIEN CHEN, Yale University, United States of America
SIYU TANG, ETH Zürich, Switzerland
ZHEBIN ZHANG, OPPO US Research Center, United States of America
CHEN LI, OPPO US Research Center, United States of America
XIAOGANG JIN∗, State Key Lab of CAD&CG, Zhejiang University, China

arXiv:2401.15687v2 [cs.CV] 30 Jan 2024

Contents

- Background and Motivation
- Research Problem
- Main Contribution
- The Results of Portrait3D
- Comparative Results
- Reference

Background and Motivation

- **Background**:
  - **Challenges**: Traditional 3D portrait modeling is time-consuming and limited in quality, with rising demand for better 3D generation in VR and related fields.
  - **Technological Advances**: Neural rendering and diffusion models have advanced, but existing methods still struggle with texture consistency and details, often producing artifacts.

Background and Motivation

- **Motivation**:
  - **Need**: There is a demand for generating high-quality, realistic 3D portraits from text.
  - **Challenges**: Overcoming limitations in geometric detail and texture handling, while reducing artifacts.
  - **Solution**: Introduce pyramid tri-grid representation, combining GANs and diffusion models to generate high-quality 3D portraits from text.

Research Problem

- **Limitations of Existing Methods**:
  - Current text-to-3D generation methods rely heavily on geometric priors, leading to issues with inconsistent and unrealistic textures in the generated portraits.
  - These methods often suffer from oversmoothing, oversaturation, and grid-like artifacts, especially when handling high-frequency information.

Research Problem

- **Research Problem**:
  - How to design an effective joint prior that incorporates both geometric and appearance information to generate high-quality 3D portraits.
  - The method should avoid texture inconsistencies and artifacts seen in current approaches, while ensuring realism and consistency across different viewing angles.

Main Contribution - 1

- **Development of 3DPortraitGAN Generator**:
  - The authors introduced a **3DPortraitGAN generator**, which uses the pyramid tri-grid to produce high-quality, 360-degree full-head 3D portraits, serving as a robust joint geometry-appearance prior .

Main Contribution 1

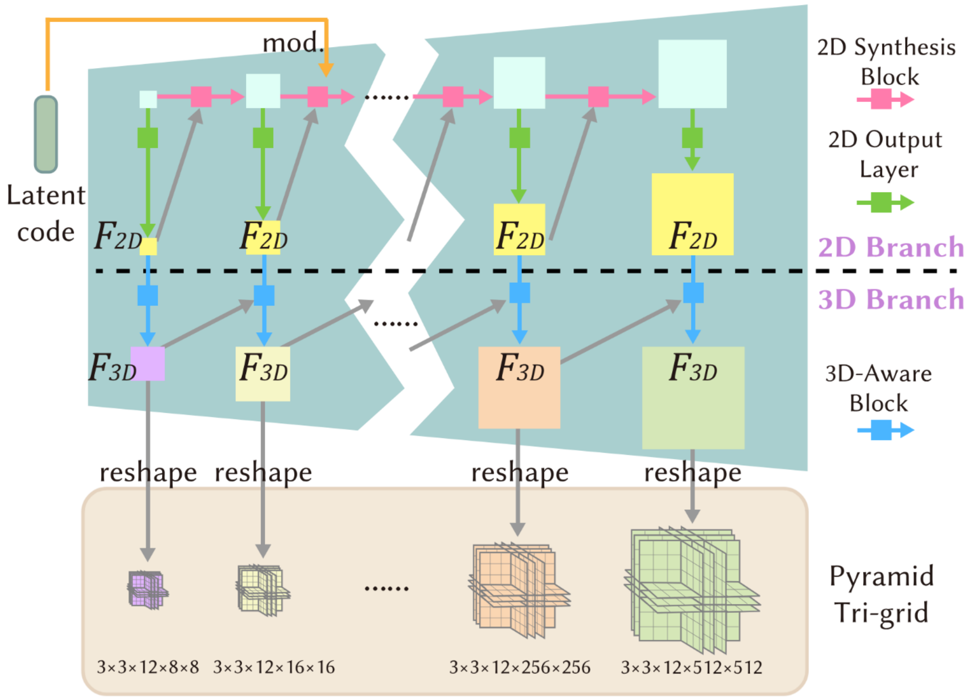


Fig. 3. The architecture of the 3D-aware *pyramid tri-grid* generator in 3DPortraitGAN. The *pyramid tri-grid* is composed of *tri-grids* generated at different layers. For the sake of simplicity and clarity, we omit the latent code modulation applied to each block.

The generator begins with a latent code to produce 2D feature maps in the 2D synthesis block, which are transformed into tri-grid representations by the 3D branch. The pyramid tri-grid is built across layers with different resolutions and aggregated to form the full 3D portrait for rendering.

2D Branch:

--2D Synthesis Block: This block is responsible for generating 2D feature maps, which are then passed to the 2D output layer. This block operates similarly to the synthesis blocks in traditional 2D generation networks, such as those in StyleGAN.

--2D Output Layer: This layer generates 2D feature maps , which are passed into the 3D branch. The 2D feature maps contain important information in the generation of the 3D representation in the next step.

3D Branch:

--3D-Aware Block: The 3D branch processes the 2D feature maps using a 3D-aware block. This block performs an upsampling operation and reshapes the processed feature maps into a 3D tri-grid representation.

--Generated Tri-grid: After processing by the 3D-aware block, the feature maps are converted into a tri-grid, which stores the color and density information of the 3D portrait. This tri-grid is essential for creating a high-quality 3D representation.

Pyramid Tri-grid Generation:The pyramid tri-grid generator generates features at different resolutions across multiple layers, progressively building up the final 3D representation. By generating tri-grids at multiple resolutions and aggregating them, the system can better capture multi-scale details and reduce high-frequency noise and artifacts, resulting in a high-quality 3D portrait.

Main Contribution - 2

- **Text-Driven 3D Generation with Diffusion Models**:
  - The paper presents a framework that combines diffusion models for **text-guided 3D portrait generation**, refining the results to create high-quality, view-consistent 3D portraits from text inputs
- **Efficient Optimization Method**:
  - The authors propose an optimization method using 21 different view renders to further improve the quality and consistency of 3D portraits

Main Contribution – 2

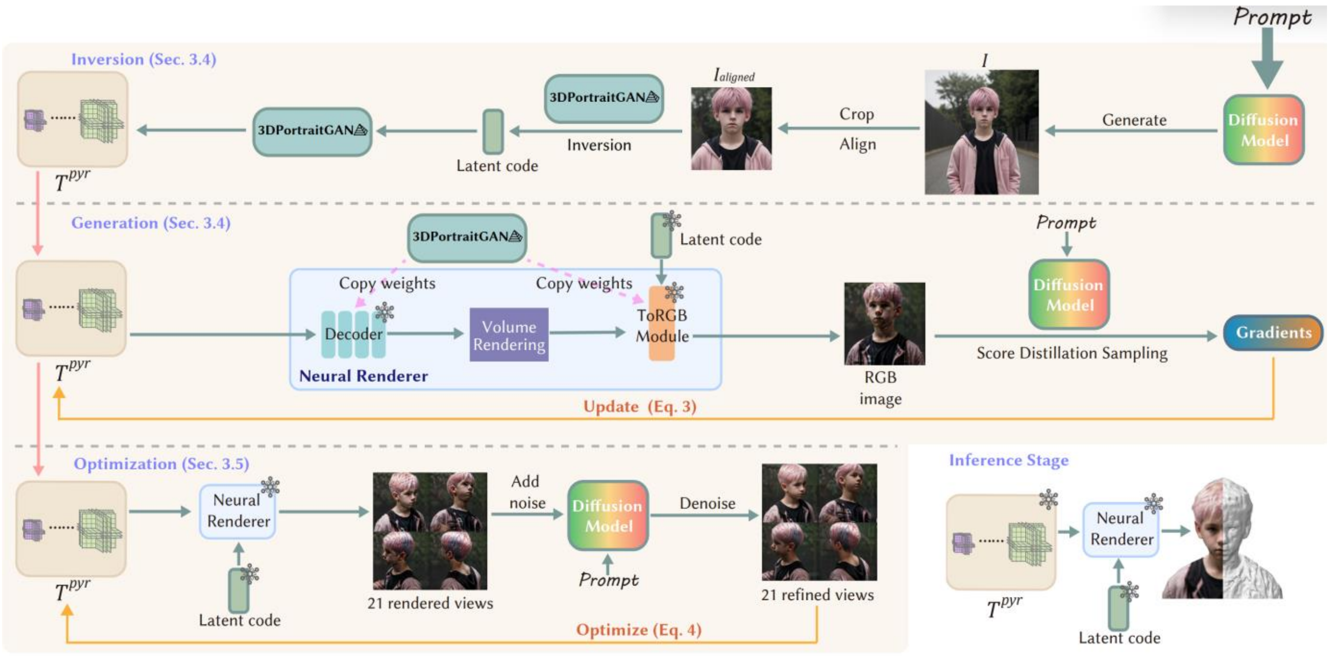


Fig. 4. The 3D portrait generation pipeline of Portrait3D. The “❄” denotes that the submodule or representation is frozen.

The figure illustrates the 3D portrait generation pipeline in Portrait3D, combining 3DPortraitGAN and a diffusion model.

And it clearly depicts each step through arrows and labeled modules, showing the entire process from the text input to the generation of a high-quality 3D portrait.

Here's a detailed explanation of each part:

1-Generate:This step begins with a text prompt, where a random 2D portrait image is generated using a diffusion model. This serves as the initial input for the 3D generation process.

2-Align:The generated random image undergoes an alignment process to ensure it fits the standards of the 3DPortraitGAN model. The resulting image is referred to as the "Aligned Image."

3-Inversion:Through latent code optimization, the aligned image is projected into the latent space of 3DPortraitGAN, resulting in a latent code. This latent code forms the basis for the subsequent 3D generation.

4-Pyramid Tri-grid Generation:The latent code is used to generate the corresponding Pyramid Tri-grid, which serves as the initial 3D representation of the portrait. The pyramid tri-grid contains multi-level features that are aggregated into a full 3D representation at different resolutions.

5-Score Distillation Sampling (SDS):On top of the pyramid tri-grid, SDS is applied. It uses the knowledge from the diffusion model to optimize the tri-grid. By freezing the neural renderer's parameters, the gradients produced by SDS are backpropagated into the pyramid tri-grid to enhance the 3D representation.

6-Render 21 Views:The pyramid tri-grid is rendered from 21 different viewpoints, chosen based on different azimuth and elevation angles. These views provide a comprehensive visualization of the 3D portrait's shape and details.

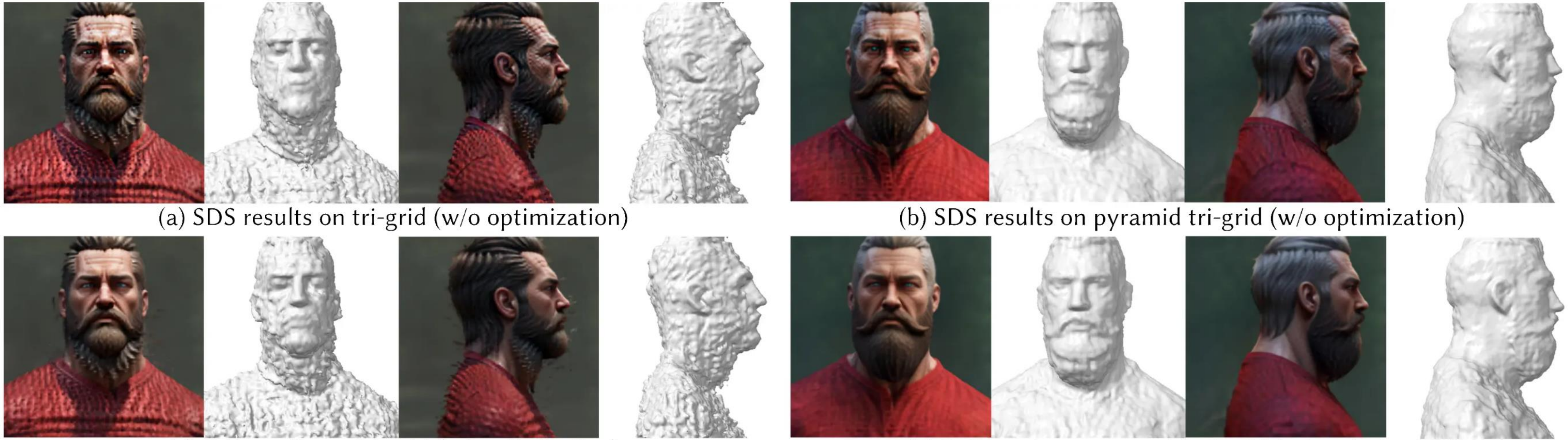
7-Add Noise and Denoise:Random noise is added to the rendered views, and the diffusion model is used to denoise the images. This process helps eliminate artifacts and further improves the visual quality of the 3D portrait.

8-Final Optimization:The final step computes the loss between the denoised images and the original rendered views, further optimizing the pyramid tri-grid's parameters. This ensures that the final 3D portrait maintains high quality and consistency across all views.

Main Contribution - 3

- **Introduction of Pyramid Tri-Grid 3D Representation**:
  - The paper proposes a novel **pyramid tri-grid 3D representation** to alleviate the "grid-like" artifacts caused by high-frequency information, producing more realistic 3D portraits

Main Contribution – 3



(a) SDS results on tri-grid (w/o optimization)

(b) SDS results on pyramid tri-grid (w/o optimization)

(c) Full results on tri-grid (w/ optimization)

(d) Full results on pyramid tri-grid (w/ optimization)

This Figure illustrates the effectiveness of the pyramid tri-grid representation in reducing "grid-like" artifacts, divided into four sections:

(a) and (b): These show the results of Score Distillation Sampling (SDS) using the traditional tri-grid and pyramid tri-grid representations without further optimization.

In (a), the tri-grid results show pronounced grid-like artifacts, while (b) shows that the pyramid tri-grid significantly reduces these artifacts.

(c) and (d): These depict the final results after optimization based on the tri-grid and pyramid tri-grid representations.

The optimized pyramid tri-grid (d) produces smoother and more realistic images, while the traditional tri-grid (c) still exhibits visible artifacts, although improved.

This Figure highlights the advantages of the pyramid tri-grid in mitigating high-frequency artifacts and demonstrates its capability to produce higher quality, more realistic 3D portraits

The Results of Portrait3D

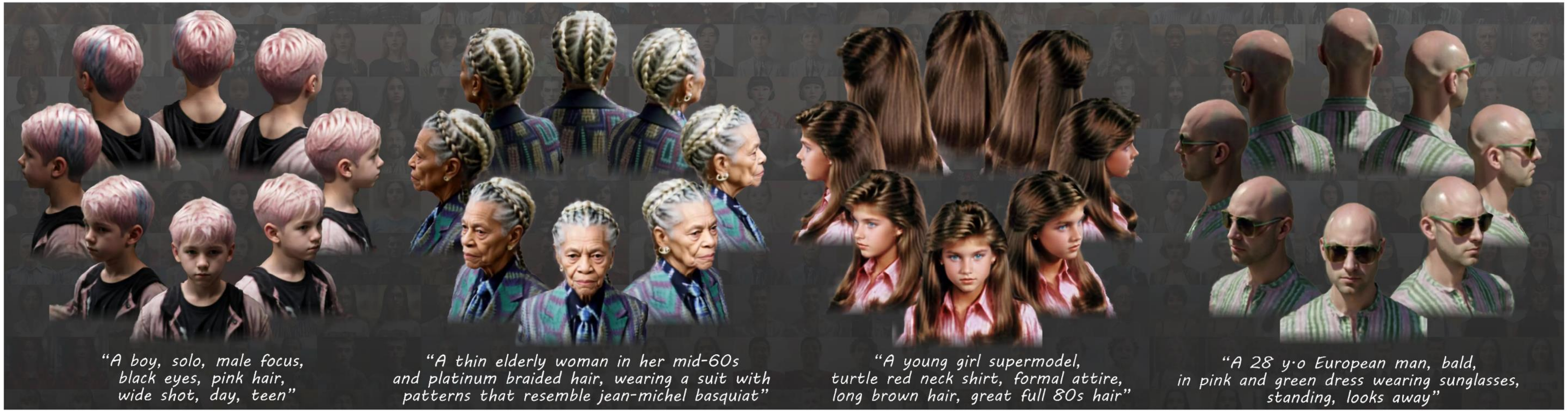
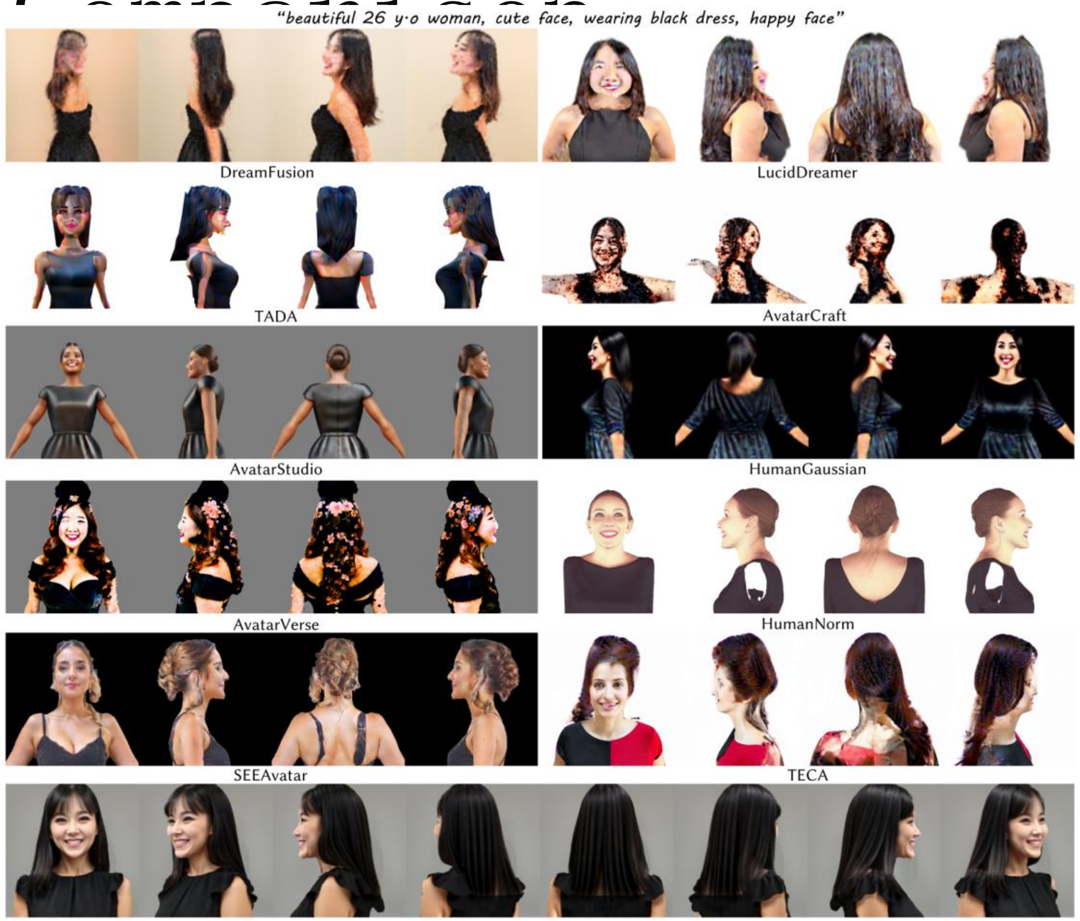

"A boy, solo, male focus, black eyes, pink hair, wide shot, day, teen"
"A thin elderly woman in her mid-60s and platinum braided hair, wearing a suit with patterns that resemble jean-michel basquiat"
"A young girl supermodel, turtle red neck shirt, formal attire, long brown hair, great full 80s hair"
"A 28 y·o European man, bald, in pink and green dress wearing sunglasses, standing, looks away"

Fig. 1. Using text as input, our text-to-3D-portrait method, Portrait3D, can automatically generate a variety of realistic textured 3D portraits. Portrait3D consistently produces high-quality 3D portraits that are aligned with the provided text prompts. Each portrait is rendered from eight different views using volume rendering.

Comparative Results – Qualitative Comparison



This figure provides a visual comparison between Portrait3D and other state-of-the-art 3D portrait generation methods, by showcasing examples of generated 3D portraits.

DreamFusion and **AvatarCraft** both suffer from the **Janus problem**, where the generated 3D portraits exhibit inconsistent facial features from different angles. This issue significantly impacts the realism and consistency of the generated portraits.

LucidDreamer shows geometric distortions, resulting in unnatural shapes in the generated figures.

TADA produces over-saturated results with distortions, lacking realism.

HumanGaussian and **AvatarStudio** fail to capture enough detail, resulting in unrealistic geometry and appearance.

AvatarVerse also generates portraits with unnatural geometry and appearance.

Advantages of Portrait3D:Compared to these methods, **Portrait3D** consistently generates high-quality, realistic 3D portraits that avoid issues such as the Janus problem, geometric distortion, and over-saturation. The portraits generated by Portrait3D maintain consistency across different views and exhibit high levels of detail and realism. Additionally, Portrait3D demonstrates its ability to generate a diverse range of 3D portraits, including different races, ages, genders, hairstyles, and more, showcasing its strength in diversity and detail.

This Figure presents several examples of 3D portraits generated from text prompts (e.g., "a beautiful 26-year-old woman wearing a black dress, smiling"). The comparison highlights how Portrait3D outperforms other methods in terms of appearance, detail, and alignment with the input text prompts.

Comparative Results - Quantitative Comparison

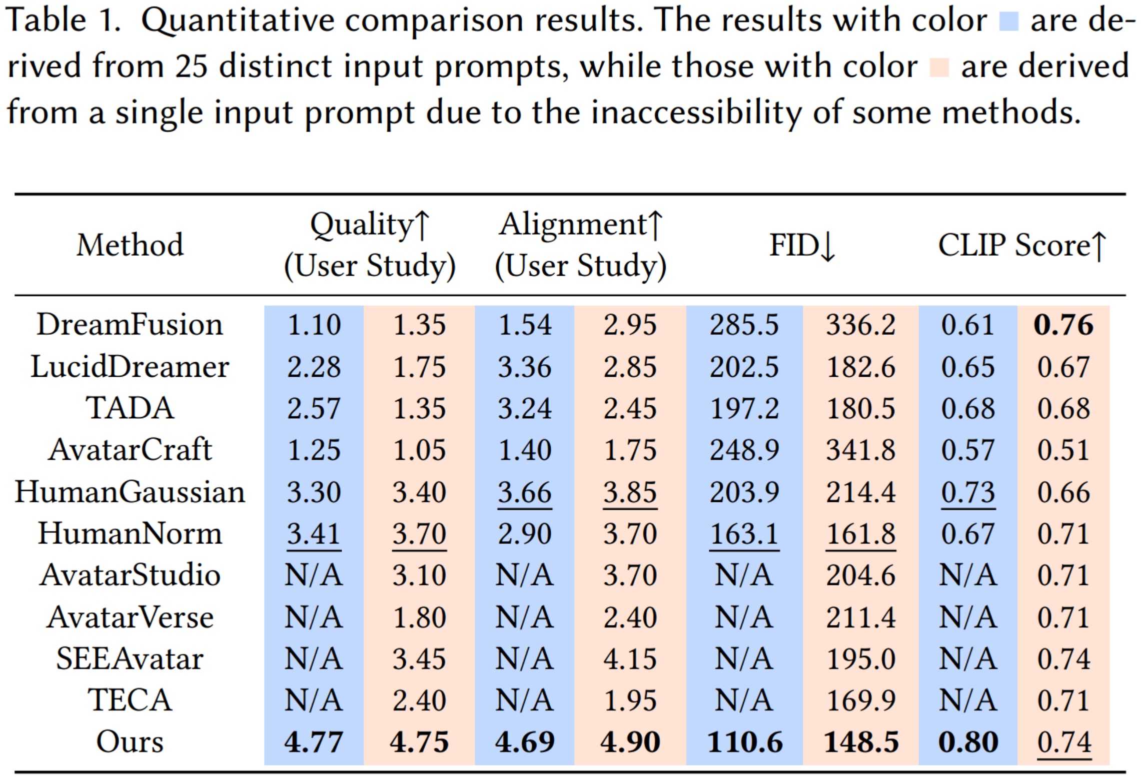


Table 1. Quantitative comparison results. The results with color 🟦 are derived from 25 distinct input prompts, while those with color 🟧 are derived from a single input prompt due to the inaccessibility of some methods.

Method	Quality↑ (User Study)		Alignment↑ (User Study)		FID↓		CLIP Score↑	
DreamFusion	1.10	1.35	1.54	2.95	285.5	336.2	0.61	**0.76**
LucidDreamer	2.28	1.75	3.36	2.85	202.5	182.6	0.65	0.67
TADA	2.57	1.35	3.24	2.45	197.2	180.5	0.68	0.68
AvatarCraft	1.25	1.05	1.40	1.75	248.9	341.8	0.57	0.51
HumanGaussian	3.30	3.40	3.66	3.85	203.9	214.4	0.73	0.66
HumanNorm	3.41	3.70	2.90	3.70	163.1	161.8	0.67	0.71
AvatarStudio	N/A	3.10	N/A	3.70	N/A	204.6	N/A	0.71
AvatarVerse	N/A	1.80	N/A	2.40	N/A	211.4	N/A	0.71
SEEAvatar	N/A	3.45	N/A	4.15	N/A	195.0	N/A	0.74
TECA	N/A	2.40	N/A	1.95	N/A	169.9	N/A	0.71
Ours	**4.77**	**4.75**	**4.69**	**4.90**	**110.6**	**148.5**	**0.80**	0.74

The paper compares Portrait3D with other state-of-the-art (SOTA) 3D portrait generation methods using different quantitative metrics, including user studies, Fréchet Inception Distance (FID), and CLIP Score.

User Study:

The paper conducted a user study involving participants who evaluated videos rendered from 3D portraits generated by different methods. Participants scored the portraits based on two criteria:

--Quality: The overall visual quality of the generated 3D portraits.

--Alignment: How well the generated portrait aligns with the input text prompt.

The scores range from 1 to 5, with higher scores indicating better performance.

Fréchet Inception Distance (FID):FID is used to measure the visual similarity between the generated images and real images. A lower FID score indicates better image quality, meaning the generated images are closer to the real image distribution.

CLIP Score:The CLIP Score measures the semantic alignment between the generated images and the input text prompts. A higher CLIP score indicates that the generated portraits better match the semantic meaning of the input prompts.

The results show that Portrait3D outperforms other methods in terms of overall quality and semantic alignment with the prompts. Across diverse input prompts, Portrait3D achieves higher scores in the user study, lower FID, and higher CLIP Scores, demonstrating its superiority in generating high-quality and text-aligned 3D portraits.

Reference

Badour Albahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, andJia-Bin Huang. 2023. Single-Image 3D Human Digitization with Shape-GuidedDiffusion. In SIGGRAPH Asia 2023 Conference Papers. Article 62, 11 pages.

Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. 2021. imGHUM: ImplicitGenerative Models of 3D Human Shape and Articulated Pose. In Proceedings of theIEEE/CVF International Conference on Computer Vision (ICCV). 5461–5470.

Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. 2023.PanoHead: Geometry-Aware 3D Full-Head Synthesis in 360deg. In IEEE/CVF Con-ference on Computer Vision and Pattern Recognition, CVPR. 20950–20959.

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini DeMello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, TeroKarras, and Gordon Wetzstein. 2022. Efficient Geometry-aware 3D GenerativeAdversarial Networks. In IEEE/CVF Conference on Computer Vision and PatternRecognition, CVPR. 16102–16112.

Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021.Pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware ImageSynthesis. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR.5799–5809.

Xingyu Chen, Yu Deng, and Baoyuan Wang. 2023a. Mimic3D: Thriving 3D-Aware GANsvia 3D-to-2D Imitation. In Proceedings of the IEEE/CVF International Conference onComputer Vision (ICCV). 2338–2348.

Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao,and Yebin Liu. 2023b. MonoGaussianAvatar: Monocular Gaussian Point-based HeadAvatar. CoRR abs/2312.04558 (2023).

Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, OrLitany, Zan Gojcic, and Sanja Fidler. 2022. GET3D: A Generative Model of HighQuality 3D Textured Shapes Learned from Images. In Advances in Neural InformationProcessing Systems, Vol. 35. 31841–31854.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley,Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative AdversarialNets. In Advances in Neural Information Processing Systems, Vol. 27. 2672–2680.

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2022. StyleNeRF: A Style-based 3D Aware Generator for High-resolution Image Synthesis. In The 10th Inter-national Conference on Learning Representations, ICLR.

Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. DensePose: DenseHuman Pose Estimation in the Wild. In IEEE/CVF Conference on Computer Visionand Pattern Recognition,