3DFaceController: Region-Controllable Face Synthesis via Decomposed and Recomposed Neural Radiance Fields

Kangneng Zhou
Nankai University
zkn@mail.nankai.edu.cn

Yaxing Wang
Nankai University
yaxing@nankai.edu.cn

Shuang Song
University of Science and Technology Beijing
ssong@ustb.edu.cn

Jie Zhang*
Macao Polytechnic University
jpeter.zhang@mpu.edu.mo

Ping Li
The Hong Kong Polytechnic University
p.li@polyu.edu.hk

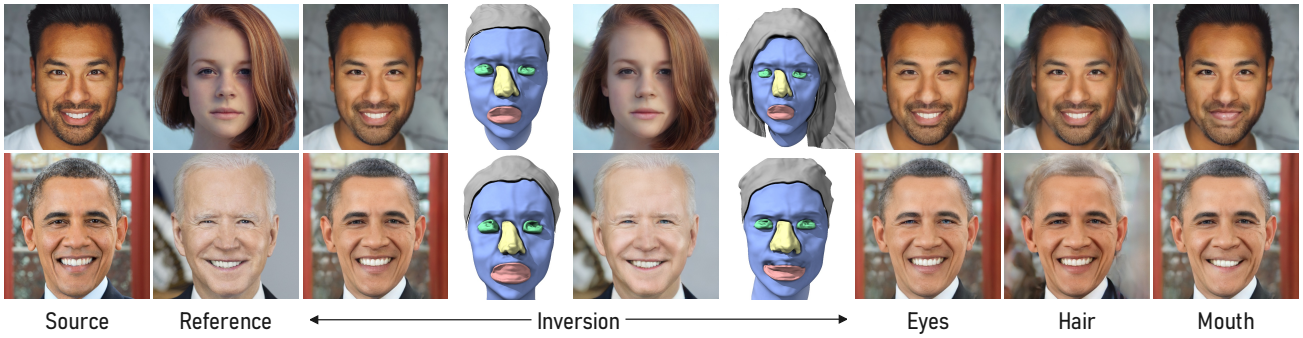

Figure 1: Our 3DFaceController results on two distinct examples. We use frequently-used GAN inversion to invert the given images and recover global and local 3D shapes. 3DFaceController can achieve face generation under region-controllable ability.

Abstract

Advancements in neural radiance fields (NeRFs) have enhanced 3D face synthesis quality. While some methods use semantic maps to guide synthesis, region-controllable synthesis remains challenging. We propose 3DFaceController, a framework for decompositional and recompositional generative radiance fields enabling region-controllable face synthesis. 3DFaceController decomposes the global face field into local fields via signed distance functions (SDF), allowing independent rendering of local components and explicit generation of physically valid 3D structures. A style-based generator with a Spatial-Semantic-Recomposition (SSR) module then synthesizes high-resolution images by combining global and local features without additional optimization. Experiments show 3DFaceController achieves state-of-the-art performance in photorealism and disentanglement.

Keywords: Facial Region Control Face Synthesis Neural Radiance Field Decomposition Recomposition.

1. Introduction

Nowadays, generation of facial images in 2D space has achieved great success. With the generation quality of pre-trained StyleGAN series [9, 10, 8] and BigGAN [2], semantic manipulating methods [13, 6, 14, 22, 19, 18, 17] can tackle many attributes, such as expression, age and hair. While 3D pose controlling is still hard to tackle since 2D methods lack 3D understanding, which makes the synthesized image lose 3D consistency.

Recently, neural radiance fields (NeRF) have made new advances in multi-view image synthesis. [4, 5, 23, 12, 20] synthesize 3D-aware images without multi-view supervi-

*Corresponding author

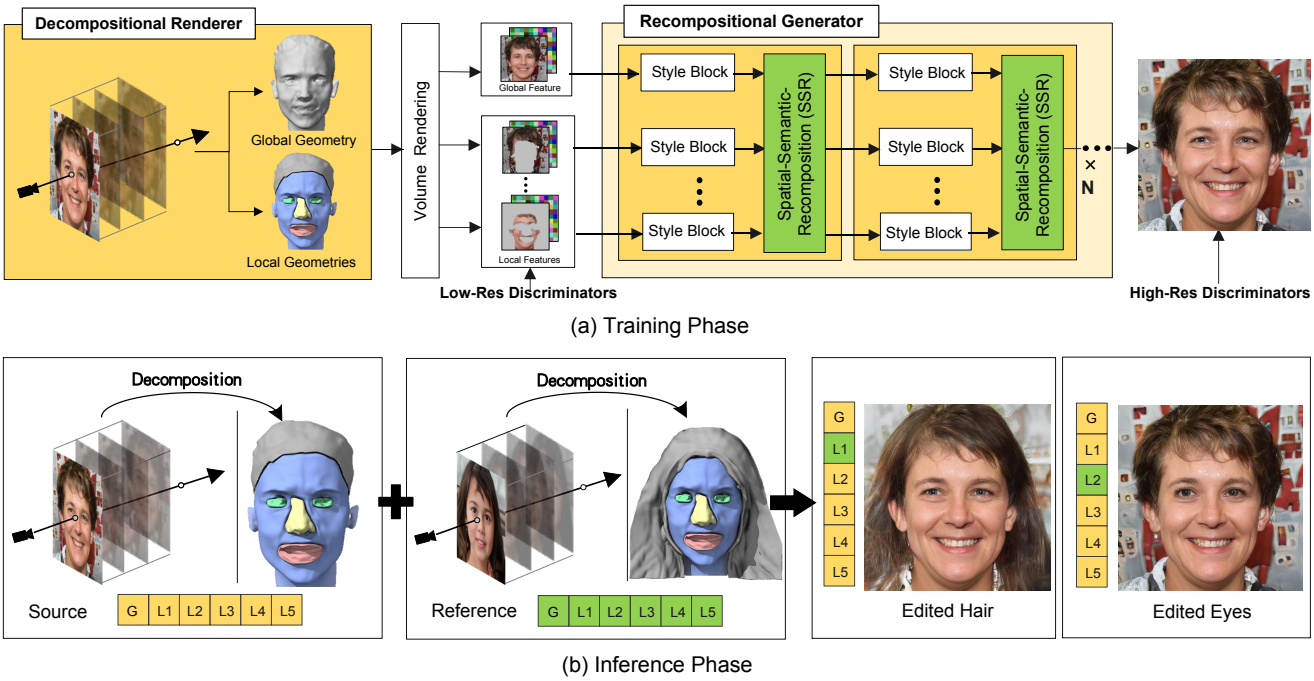

Figure 2: Framework of our proposed 3DFaceController. **(a) Training phase**: the decompositional renderer decomposes the global face into local components (e.g., the hair and hollow out face) and their corresponding features are fed into recompositional generator to synthesize high-resolution image. **(b) Inference phase**: we can conduct region-wise editing by recomposing features from different individuals.

sion but have face editability ignored. Existing NeRF-based generation methods explore editability in implicit fields which often learn semantic map field along with conventional density field and color field. FENeRF [16] enables face editing via manipulating semantic map using optimization-based inversion. MaTe3D [21] achieves mask-guided and text-based portrait generation with diffusion prior. However, existing methods still struggle with region-controllable synthesis, leading to incorrect results.

To address this challenge, we propose **3DFaceController**, enabling region-controllable face synthesis by decomposing the neural radiance field into semantic parts and recomposing them. We learn a global face representation and independent local representations (e.g., nose, eyes, mouth) using signed distance functions (SDF). By incorporating SDF consistency, density consistency (from MaTe3D [21]), and color consistency losses, we ensure physically plausible 3D surfaces and 3D-consistent thumbnails. For high-resolution synthesis, a multi-style-based 2D upsampler with a **Spatial-Semantic-Recomposition (SSR)** module is developed to recompose global and local features at different layers, enabling diverse and precise face manipulations. During inference, region-controllable synthesis is achieved by swapping specific decompositional features (see Fig. 1). The framework of our proposed 3DFaceController is shown in Fig. 2.

The main contributions of 3DFaceController are summarized as follows:

- We develop a 3DFaceController, a compositional generative framework enabling region-controllable synthesis, producing photorealistic images and accurate radiance fields.
- We propose a SSR Module to recompose local features into the global network, supporting common editing and synthesis methods.

2. Methodology

2.1. Decompositional Volume Renderer

With the 3D point coordinates $\mathbf{x} = (x, y, z)$, viewing direction $\mathbf{v} = (\theta, \phi)$ and conditional latent code $\mathbf{w}$, volume renderer produces feature vector $\mathbf{f}(\mathbf{x}, \mathbf{v})$, view-dependent color value $\mathbf{c}(\mathbf{x}, \mathbf{v})$ and a series of SDF values $\mathbf{d} = \{d_g, d_{l1}, d_{l2}, ..., d_{ln}\}$. The SDF values are not just about whole face $d_g(\mathbf{x})$, but about corresponding local facial components (ie., nose, eye, mouth)$\{d_{l1}(\mathbf{x}), d_{l2}(\mathbf{x}), ..., d_{ln}(\mathbf{x})\}$. n represents the number of semantic part, see Fig. 2. Our volume renderer is a multi-head network and represents each facial component individually. The formulation of the

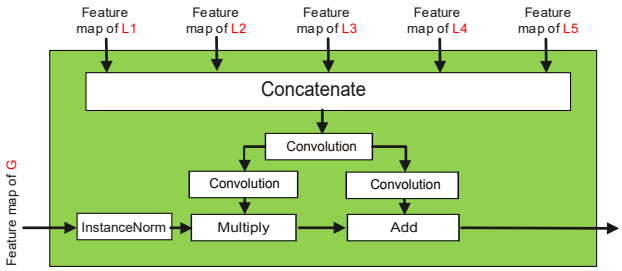

Figure 3: Diagram of our Spatial-Semantic-Recomposition (SSR). SSR generates spatially varying parameters from semantic regions and separates the local components for synthesis, enabling 3DFaceController to produce highly decoupled photorealistic images.

volume renderer is:

$$(\mathbf{x}, \mathbf{v}, \mathbf{w}) \mapsto (\mathbf{d}, \mathbf{f}, \mathbf{c}) \mapsto (I_g^{thumb}, I_{li}^{thumb}). \quad (1)$$

Once the decompositional volume renderer is trained, thumbnails of the global face I_g^{thumb}, local facial components I_{li}^{thumb} can be rendered via volume rendering. Specifically, with ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{v}$ emanated from camera position o in direction d, we query N sample points along each ray and yield density values of global face σ_g and local components $\{\sigma_{l1}, \sigma_{l2}, ..., \sigma_{ln}\}$ (converted by signed distance values), color $\mathbf{c}(\mathbf{x}, \mathbf{v})$ and feature vector $\mathbf{f}(\mathbf{x}, \mathbf{v})$. Then we obtain the the pixels $C(\mathbf{r})$ and feature maps $F(\mathbf{r})$ of global face and local components via volume rendering.

2.2. Recompositional Image Generator

Similar to mainstream upsampler-based baselines [3, 20, 12], we adopt a style-based generator to lift the outputs of volume renderer to high-resolution images I_g. Instead of the feature map of global face used in traditional lift module, we have extra branches for processing the feature maps of each facial component. This allows us to recompose the global feature and local features as flexible as possible. Fig. 2 shows the details of our recompositional generator. The generator has a global branch and several local branches. The global branch (blue blocks in Fig. 2) is designed to learn the main representation of the whole identity and to project the feature to image space via ToRGB layer. The local branch (coral blocks in Fig. 2) receives a series of local representations individually. Also, these representations should be disentangling from each other while synthesizing region-controllable results. The entangling representations would loss editability and make the former decomposition meaningless.

To effectively integrate the information from the global and local branches, we design a feature recomposition module Spatial-Semantic-Recomposition (SSR) to recompose the local and global features at each backbone layer, as shown in Fig. 3. The SSR captures the spatial and semantic constraints and makes the synthesized images more disentangling. The SSR block has two inputs: a global representation produced by the previous global branch and a concatenation of decompositional representations of local branches. The semantic relationships within the local features are learned by spatially-adaptive parameters. Hence the semantic relationships lead to controllable and disentangled face editing.

2.3. Optimization

2.3.1 Decompositional Renderer Training

For the global face and its associated local facial components, we leverage the classic non-saturating adversarial loss [11] with R1 regularization $\mathcal{L}_{adv}$, along with three additional regularization losses [12]: pose alignment loss $\mathcal{L}_{view}$, Eikonal loss $\mathcal{L}_{eik}$, and minimal surface loss $\mathcal{L}_{surf}$. These losses are utilized for both the global face and the local components. Furthermore, in addition to the SDF-consistent loss $\mathcal{L}_{sdf}$ and density-consistent loss $\mathcal{L}_{sigma}$ proposed in [21], we introduce an additional color-consistent loss to establish connections among multiple geometrical structures at the image level:

$$\mathcal{L}_{img} = ||I_g^{thumb}, \sum_{i=1}^{n} I_{li}^{thumb}||^2. \quad (2)$$

In summary, the loss function of our decompositional renderer is:

$$\begin{aligned} \mathcal{L}_{dr} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{view}\mathcal{L}_{view} + \lambda_{eik}\mathcal{L}_{eik} + \\ \lambda_{surf}\mathcal{L}_{surf} + \lambda_{img}\mathcal{L}_{img} + \\ \lambda_{sdf}\mathcal{L}_{sdf} + \lambda_{sigma}\mathcal{L}_{sigma}, \end{aligned} \quad (3)$$

where, $\lambda_{adv} = 1$, $\lambda_{view} = 15$, $\lambda_{eik} = 0.1$, $\lambda_{surf} = 0.05$, $\lambda_{img} = 1$, $\lambda_{sdf} = 5$, $\lambda_{sigma} = 10$.

2.3.2 Recompositional Generator Training

Similar to mainstream upsampler-based models, we train the recompositional generator via non-saturating GAN loss $\mathcal{L}_{adv}$ [11] and path regularization $\mathcal{L}_{path}$ [12]:

$$\mathcal{L}_{rg} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{path}\mathcal{L}_{path}, \quad (4)$$

where, $\lambda_{adv} = 1$ and $\lambda_{path} = 2$.

3. Experimental Results

3.1. Qualitative Analysis

In this section, we provide a qualitative analysis of our method's performance in terms of generator capabilities and disentangled editing.

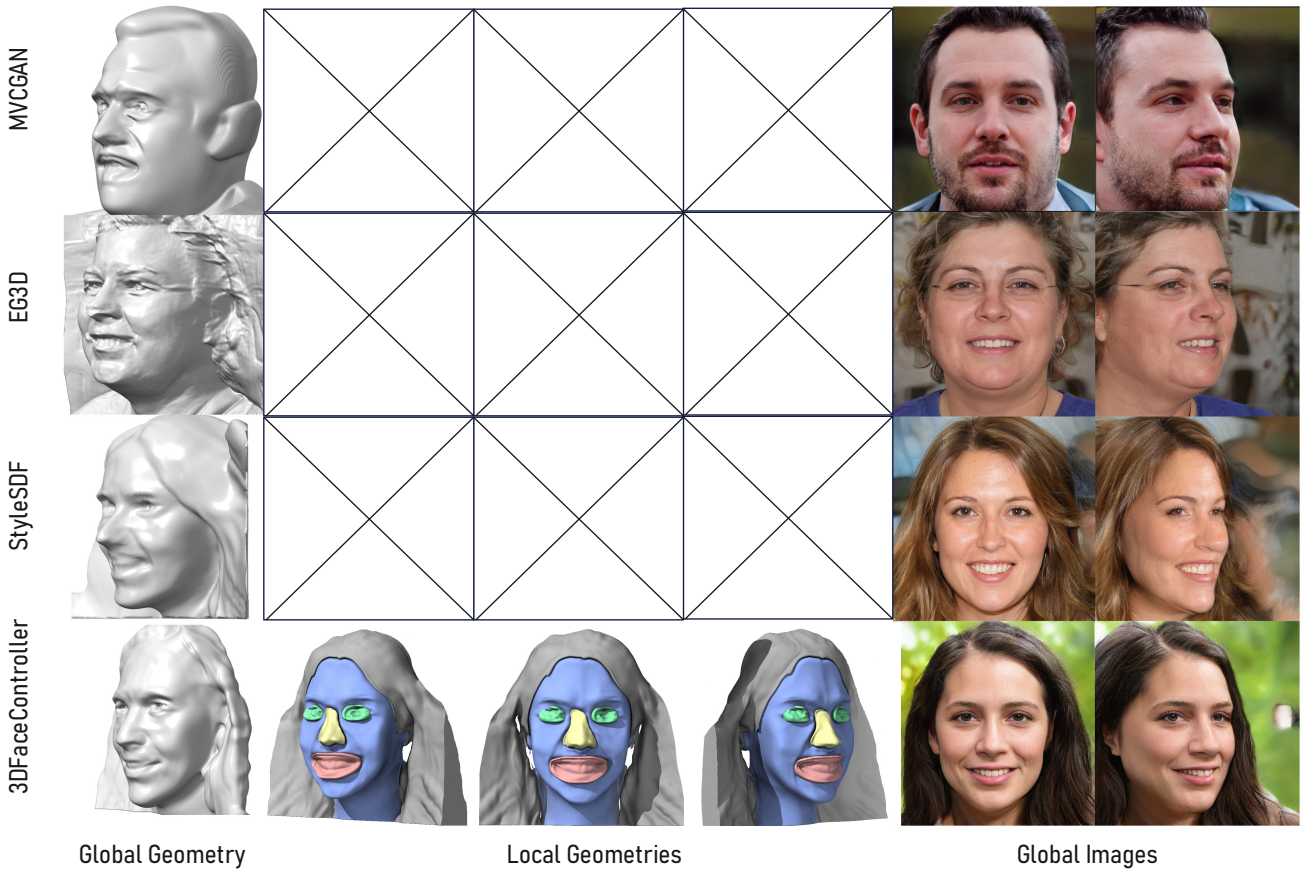

Figure 4: Qualitative comparison among upsampler-based baselines including MVCGAN [20], EG3D [3], StyleSDF [12], and our approach, conducted at 512^2 resolution on FFHQ dataset. Our method excels in learning precise geometry while maintaining comparable image quality, exhibiting minimal drawbacks such as staircasing artifacts (MVCGAN), oversmoothing tendencies (StyleSDF), and background density inconsistencies (EG3D). The symbol ☒ denotes the model's limitation in generating local geometries.

Generator Performance Fig. 4 provides a qualitative comparison of our method with upsampler-based 3D-aware GANs. The 3DFaceController exhibits accurate geometry learning and achieves comparable image quality, showcasing its robustness across diverse branches and optimization objectives. Furthermore, the figure illustrates the model's ability to generate physically plausible decompositional 3D surfaces.

Disentangled Editing In face editing tasks, our objective is to edit a specific local region in a manner that is disentangled from others. Specifically, we aim to modify one region while preserving the rest unchanged. As depicted in Fig. 5, we present two identities labeled in red and blue. These faces have been broken down into six components: one global feature and 5 local features. During the recomposition process, we replace the red features with the blue ones, resulting in disentangled and realistic region-wise editing. Notably, in instances where hair features are replaced, our method effectively captures styles from both identities, blending textures and shapes appropriately.

3.2. Quantitative Analysis

In this section, we summarize the insights gained from both the metrics evaluation and user study, shedding light on the quality and visual appeal of our generated images.

Metrics Evaluation We utilize two quantitative metrics to comprehensively evaluate the synthesis quality of the generated images: the Frechet Inception Distance (FID) [7] and Kernel Inception Distance (KID) [1]. These metrics offer insights into the fidelity and diversity of the produced images. A lower FID and KID value signifies higher image quality, indicating better alignment with real image distributions. As depicted in Table 1, our model, 3DFaceController, demonstrates competitive and promising performance in terms of both FID and KID, emphasizing its effectiveness in generating high-quality synthetic images.

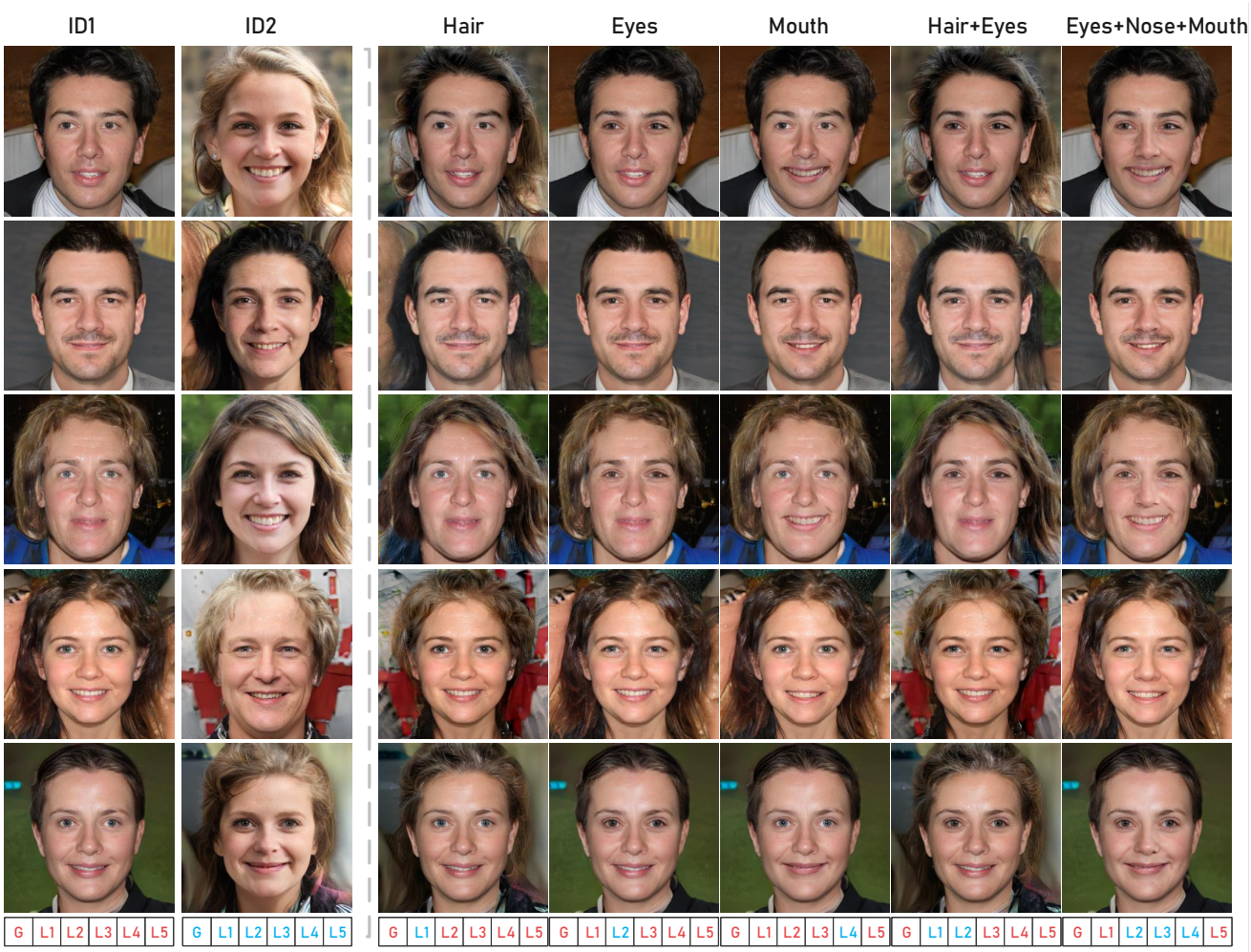

Figure 5: The editing results of 3DFaceController. Our approach supports region-wise editing on single class or multi classes. The figure shows editing results on **hair and background** (the 3rd volume), **eyes** (the 4th volume), **mouth** (the 5th volume), the union of **hair, background and eyes** (the 6th volume) and the union of **eyes, nose and mouth** (the 7th volume).

Method	FID ↓	KID ↓	User Study ↑
FENeRF [16]	39.5	5.889	5.7
IDE-3D [15]	**13.4**	**0.130**	19.7
3DFaceController	15.6	0.225	**74.6**

Table 1: Quantitative comparisons with the state-of-the-art methods. Note that our 3DFaceController excels in both FID and KID metrics, showcasing its effectiveness in generating high-quality synthetic images that are visually convincing and realistic, surpassing alternative methods.

User Study Additionally, we conduct a user study using the synthesized images to evaluate the photorealism of our results. During this study, we invite 100 participants, each of whom is randomly assigned 20 sets of images selected from $1,000$ groups. Participants will compare our method against several baselines through pairwise comparisons. The percentages obtained indicate the frequency with which participants preferred our approach over each baseline. Notably, our method consistently outperformed all baselines, even in scenarios where the baselines exhibited superior FID or KID scores. This underscores the capability of our approach to produce visually compelling and realistic images, surpassing technical metrics that may favor alternative methods.

4. Conclusion

This study demonstrates the feasibility of decomposing the global neural radiance field into distinct local generative radiance fields and subsequently recombining them to generate high-resolution images with physically plausible geometry. By employing explicit decomposition and composition strategies, 3DFaceController showcases superior performance compared to existing models in terms of producing realistic radiance fields, thereby exhibiting robust 3D-consistent properties. We consider this approach a promising direction for integrating 2D GANs with 3D awareness. Extensive experimentation validates the viability of decompositional radiance fields at both image and geometry levels, while the compositional generator excels in achieving realistic face editing on a region-wise basis. The results underscore the potential for advancing image synthesis and manipulation tasks in the realm of 3D-aware generative models.

Disclosure of Interests.

The authors have no competing interests to declare that are relevant to the content of this article.

Acknowledgments.

This work was supported in part by The Hong Kong Polytechnic University (PolyU) under Grant P0042740, the PolyU Research Institute for Sports Science and Technology under Grant P0044571, and the Macao Polytechnic University under Grant RP/FCA-03/2024. Thanks to OpenBayes.com for providing model training and computing capabilities.

References

[1] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying Mmd GANs. In *International Conference on Learning Representations*, 2018. 4

[2] A. Brock, J. Donahue, and K. Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*, 2018. 1

[3] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. D. Mello, O. Gallo, L. Guibas, J. Tremblay, S. Khamis, T. Karras, and G. Wetzstein. Efficient Geometry-aware 3D Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 3, 4

[4] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein. pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5799–5809, 2021. 1

[5] J. Gu, L. Liu, P. Wang, and C. Theobalt. StyleNeRF: A Style-based 3D Aware Generator for High-resolution Image Synthesis. In *International Conference on Learning Representations*, 2022. 1

[6] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris. GANSpace: Discovering Interpretable GAN Controls. In *Advances in Neural Information Processing Systems*, volume 33, pages 9841–9850, 2020. 1

[7] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs Trained by a Two Time-scale Update Rule Converge to a Local Nash Equilibrium. *Advances in neural information processing systems*, 30:6629–6640, 2017. 4

[8] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila. Alias-Free Generative Adversarial Networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 852–863, 2021. 1

[9] T. Karras, S. Laine, and T. Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1

[10] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and Improving the Image Quality of StyleGAN. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8107–8116, 2020. 1

[11] L. Mescheder, A. Geiger, and S. Nowozin. Which training methods for gans do actually converge? In *International Conference on Machine Learning*, pages 3481–3490, 2018. 3

[12] R. Or-El, X. Luo, M. Shan, E. Shechtman, J. J. Park, and I. Kemelmacher-Shlizerman. StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 1, 3, 4

[13] Y. Shen, J. Gu, X. Tang, and B. Zhou. Interpreting the Latent Space of GANs for Semantic Face Editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9240–9249, 2020. 1

[14] Y. Shen and B. Zhou. Closed-Form Factorization of Latent Semantics in GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1532–1540, 2021. 1

[15] J. Sun, X. Wang, Y. Shi, L. Wang, J. Wang, and Y. Liu. IDE-3D: Interactive Disentangled Editing for High-Resolution 3D-aware Portrait Synthesis. *ACM Transactions on Graphics*, 41(6):1–10, 2022. 5

[16] J. Sun, X. Wang, Y. Zhang, X. Li, Q. Zhang, Y. Liu, and J. Wang. FENeRF: Face Editing in Neural Radiance Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7672–7682, 2022. 2, 5

[17] J. Zhang, Y. Luximon, P. Shah, K. Zhou, and P. Li. Customize My Helmet: A Novel Algorithmic Approach Based on 3D Head Prediction. *Computer-Aided Design*, 150:103271:1–103271:10, 2022. 1

[18] J. Zhang, K. Zhou, Y. Luximon, , P. Li, and T. Lee. Mesh-WGAN: Mesh-to-Mesh Wasserstein GAN with Multi-Task Gradient Penalty for 3D Facial Geometric Age Transformation. *IEEE Transactions on Visualization and Computer Graphics*, 30(8):4927–4940, 2024. 1

[19] J. Zhang, K. Zhou, Y. Luximon, T.-Y. Lee, and P. Li. 3DCMM: 3D Comprehensive Morphable Models with UV-UNet for Accurate Head Creation. *IEEE Transaction on Multimedia*, pages 1–14, 2024. 1

[20] X. Zhang, Z. Zheng, D. Gao, B. Zhang, P. Pan, and Y. Yang. Multi-View Consistent Generative Adversarial Networks for 3D-aware Image Synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 18450–18459, 2022. 1, 3, 4

[21] K. Zhou, D. Gao, X. Wang, J. Zhang, P. Zhang, X. Sun, L. Zhang, S. Yang, B. Zhang, L. Bo, et al. MaTe3D: Mask-guided Text-based 3D-aware Portrait Editing. *arXiv preprint arXiv:2312.06947*, 2023. 2, 3

[22] K. Zhou, X. Zhu, D. Gao, K. Lee, X. Li, and X.-c. Yin. SD-GAN: Semantic Decomposition for Face Image Synthesis with Discrete Attribute. In *Proceedings of the ACM International Conference on Multimedia*, pages 2513–2524, 2022. 1

[23] P. Zhou, L. Xie, B. Ni, and Q. Tian. CIPS-3D: A 3D-aware Generator of GANs Based on Conditionally-independent Pixel Synthesis. *arXiv preprint arXiv:2110.09788*, pages 1–10, 2021. 1