InViTe: Individual Virtual Transfer for Personalized 3D Face Generation System

Mingyu Jang, **Kyungjune Lee**, **Seongmin Lee**, **Hoseok Tong**, **Juwan Chung**, **Yusung Ro** and **Sanghoon Lee**
Yonsei University
{jmg1002, naive2kj90, lseong721, hstong09, wndhks0506, maple3067, slee}@yonsei.ac.kr

Abstract

With the expansion of the virtual communication industry using VR/AR, it has attracted increasing attention to enable users to represent their personalities in a 3D avatar. As the face of 3D avatars plays a crucial role in conveying human personality, a system that generates and manipulates 3D faces is desired. However, establishing the system is challenging due to the need for human effort and specialized knowledge. To fill this void, we present the **In**dividual **Vi**rtual **T**ransf**e**r (InViTe), which enables the creation and customization of a 3D face according to the user's preference. Our proposed system is featured for 1) 3D face reconstruction with high fidelity texture map, 2) 3D face personalization, 3) realistic rendering results, and 4) real-time mobile virtual applications. We conduct an experiment to demonstrate that the proposed system can achieve sufficient individual personalization of 3D faces. Furthermore, we evaluate the system's data transmission protocol and demonstrate its efficiency. The demonstration video is available at https://www.youtube.com/watch?v=D_4pXZvGUWU.

1 Introduction

3D avatar personalization enables users to represent their personalities in the virtual space. It is continuously emerging due to the growth of the virtual communication industry because users have a desire to express their personality [Kim *et al.*, 2016; Ahn *et al.*, 2018; Wang *et al.*, 2020; Kim *et al.*, 2021]. The personality can be affected by the human face [Kachur *et al.*, 2020], so it is necessary to develop a system that personalizes 3D faces for avatars to represent one's personality.

The 3D face personalization system can be conducted through two steps: 3D face reconstruction and 3D face manipulation. These two steps have been achieved through various computer graphics or computer vision approaches [Paysan *et al.*, 2009; Li *et al.*, 2017]. However, they heavily rely on human efforts and require professional knowledge. To reconstruct a 3D face, sophisticated hardware like 3D scanners or multi-view camera systems is used to obtain a face mesh, which also needs refinement for high-quality geometry details [Jang *et al.*, 2022; Lee *et al.*, 2023; Li *et al.*, 2021]. For 3D face manipulation, a texture map is employed, obtained either with specialized equipment or designed manually by a 3D artist.

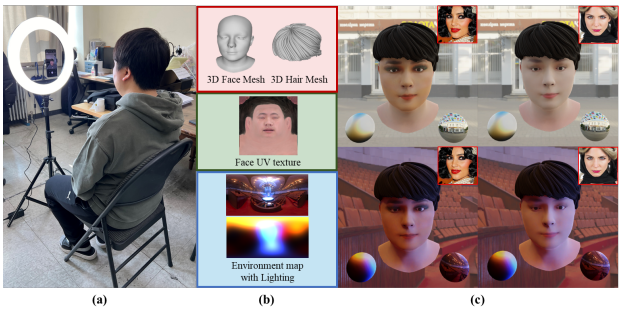

Figure 1: Summary of InViTe: (a) Capturing face to generate personalized 3D faces. (b) Intermediate outputs of InViTe for rendering. (c) 3D face manipulation involves changing makeup styles on a mobile device.

To alleviate this problem, in this paper, we propose a system called Individual Virtual Transfer (InViTe), designed to create and customize the 3D virtual face for personalization. Our proposed system has four desirable properties. **First, our system performs accurate 3D face reconstruction with a high fidelity texture map.** To do this, we utilize a 3D Morphable Model (3DMM) [Li *et al.*, 2017], which enables us to obtain both a 3D face geometry and textures through a linear combination of coefficients. However, since the texture computed from the coefficients can result in coarse results, our system additionally produces high-quality texture maps using a generative model-based face normalization [Bai *et al.*, 2023]. **Second, our system can personalize the obtained 3D face avatars.** The personalization preserves the identity while editing its appearance. To represent the detail of identity, our proposed system additionally recovers the hair from the input image. For editing appearance, we stylize the face UV texture by giving a 2D style image to manipulate the face appearance. **Third, our system provides realistic rendered results for a more immersive experience.** We include a Physically Based Rendering (PBR) process, which renders an image by considering the illumination effects. Specifically,

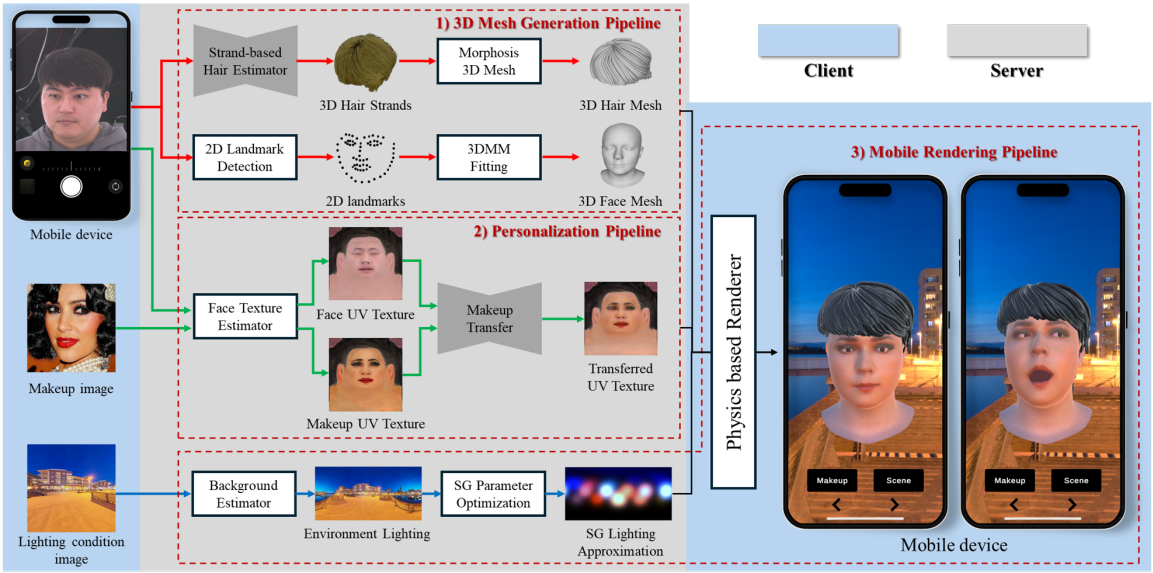

Figure 2: The InViTe framework is divided into client and server components. A single portrait image, a makeup image, and a lighting condition image are transmitted to the server through a mobile device. Subsequently, three pipelines are executed to render and manipulate the 3D face.

our system takes an image for the illumination condition as input and automatically estimates surrounding lighting information. **Finally, our system supports real-time mobile virtual applications.** However, the computation of the personalization process is complex to perform on a mobile device. Therefore, we employ a client-server system framework to reduce both computation costs and data transmission for real-time processing.

To demonstrate the effectiveness of the proposed system, we conduct a personalization experiment. Moreover, we additionally experiment the latency efficiency of our proposed system. Experimental results demonstrate that the proposed system is not only well-suited for 3D avatar personalization but also enables real-time rendering processes for immersive virtual applications.

2 System Framework

In this section, we describe our framework for creating and manipulating personalized 3D virtual faces as shown in Fig. 2. InViTe is broadly divided into client and server components. On the client side, the system receives a single portrait image, a makeup image, and a lighting condition image through a mobile device, which is then transmitted to the server. After that, on the server side, the system runs three pipelines as follows: 1) 3D Mesh Generation Pipeline, 2) Personalization Pipeline, and 3) Mobile Rendering Pipeline. The system's server generates a 3D face mesh and 3D hair mesh in the 3D mesh generation pipeline. Simultaneously, a face UV texture is generated in the Personalization Pipeline. Specifically, we employ a makeup style transfer approach to personalize a face UV texture. In the third pipeline, accounting for lighting effects, our system generates lighting information as Spherical Gaussian parameters. Finally, our system transmits all generated data to the mobile device for mobile rendering. The mobile device uses the front camera to track the 3D landmarks of the face [Kartynnik *et al.*, 2019] and deform the rendered 3D face mesh following the user's movement.

2.1 3D Mesh Generation Pipeline

To generate a 3D face mesh, we employ 3D face mesh estimation [Daněček *et al.*, 2022], which fits the FLAME model [Li *et al.*, 2017] from a single 2D image. However, it is insufficient to represent the user's identity since the Flame model does not support hair. The most effective method for processing hair is to generate each 3D hair strand, but it is generally a challenging task due to the extreme computing power required for physics calculations. Also, hair material possesses an anisotropic reflection characteristic which makes it difficult to achieve realistic rendering. Therefore, to accomplish plausible of rendering 3D hair mesh in the mobile device, we handle hair as a single mesh while the mesh is generated from separated strands. To generate a 3D hair mesh, we first employ a strand-based hair estimator [Zheng *et al.*, 2023]. Subsequently, we generate a solid 3D hair mesh along each strand by calculating directions within each point on every strand and consolidate all strands into a single mesh, as depicted in the morphosis 3D mesh in Fig. 2. As a result, it is possible to deliver realism induced by detailed geometry, while maintaining moderate computational burden.

2.2 Personalization Pipeline

In our proposed system, the personalization pipeline consists of the texture estimator and the makeup transfer. The texture estimator extracts the face texture from a single image. The makeup transfer takes two face textures as content and style respectively, and stylizes the makeup attributes of the makeup UV texture into the content face UV texture. Specifically, the personalized results may rely on the quality of the face UV texture, but it is still a challenging task due to occlusion or

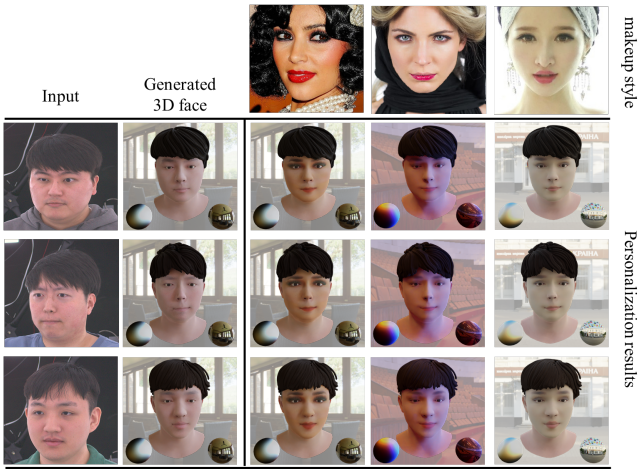

Figure 3: Qualitative results of InViTe. All Results show that personalization is sufficient even with changes in the subjects' makeup style and environment lighting.

various head poses/expressions in the single face image. To alleviate this problem, we adopt the face normalization approach used in previous work [Bai *et al.*, 2023], which produces a completed face UV texture from normalized multi-view faces using styleGAN-based facial image editing. After obtaining the face UV texture, we design the makeup transfer network that produces an alpha blending map and makeup layer similar to [Yang *et al.*, 2023]. In particular, for more high-fidelity results, we employ the styleGAN [Abdal *et al.*, 2021] based architecture. We train the network using the MT dataset following the training strategy in the previous work [Yang *et al.*, 2023]. Finally, to transfer the makeup style, we blend the makeup layer and the face UV texture using the alpha blending map to obtain the transferred texture.

2.3 Mobile Rendering Pipeline

In our framework, the rendering pipeline with all of the 3D face data in Sec. 2.1 and 2.2 is processed on mobile client through PBR. To express lighting to the user's movement described in Sec. 2, environment lighting is applied to the rendering process. We employ a GAN-based model similar to [Dastjerdi *et al.*, 2023] which generates 360 high dynamic range image (HDRI) from a single image selection of the user as background estimator. The environment lighting enables superior rendering results, but high resolution 360 HDRI with intense computation is usually required. However, considering a lightweight mobile environment, we utilize a Spherical Gaussian (SG) representation. The SG approximation of HDRI is compact and efficient, where any HDRI can be compressed into a small vector with simple irradiance calculation coming from the nature of the Gaussian function. The HDRI is approximated into 19 by 3 vectors by optimizing the RGB strengths of 19 lobes with fixed location and width. Therefore, fast and efficient rendering of faces with user selection-based environment lighting on a mobile device is possible.

Data type	Ours	Video stream	Mesh stream
Latency(s)	**0.03**	0.17	0.31
Relative ratio	**1.0**	5.7	10.3

Table 1: The comparison of latency between server and client. Video stream continuously transmits frames rendered on the server to the client. Mesh stream transmits raw mesh to the client.

3 System Evaluation

We first show qualitative results as shown in Fig. 3. As depicted in the figure, our system supports various personalizations through makeup transfer and produces rendering results by introducing PBR. To verify the effectiveness of our system, we conduct two quantitative experiments: personalization and data transmission.

(1) Personalization. To evaluate personalization performance, we generate 10 fitted 3DMM parameters and 3D hair from 10 volunteers years old to 32 years old. Then, we randomly select 15 makeup images from the MT test set [Li *et al.*, 2018] and perform makeup transfer to each face texture. We render each 3D face with the transferred texture using selected makeup images. After that, we measure the accuracy of face verification using ArcFace [Deng *et al.*, 2019] to determine whether the face in the rendered view matches the face in the portrait image. Furthermore, we also conduct a subjective test in which participants were asked to answer a question; *How well does this result reflect personalization without altering the identity?*. The participants can score the scene from 1 (worst) to 5 (best). The results achieved a **verification task accuracy of 91.4%** and received a **subjective score of 4.3**. Therefore, the results show that our proposed system achieves proper personalization by only manipulating the face appearance.

(2) Data transmission. To verify the transmission efficiency of our system architecture, we conduct experiments on the data transmission protocol. We categorize data transmission in our system into three cases: Our approach, video stream, and mesh stream. The video stream renders the 3D face on the server using 3D landmarks and continuously transmits the rendered view to the mobile device. In contrast, for the mesh stream, the server directly transmits the 3D face mesh, and the client is responsible for rendering. The results are shown in Table 1. Compared to other methods, Ours achieves a transmission efficiency of 82.4% compared to video stream and 90.3% compared to mesh stream, which proves that InViTe can serve real-time rendering to users.

4 Conclusion

We propose a 3D face personalization system, called InViTe, which enables both the generation of 3D faces and customization of their textures for personalization. Furthermore, we incorporate a PBR process into our system to achieve realistic rendered results. Based on a server-client architecture, our proposed system supports real-time mobile virtual applications. Evaluation results demonstrate the effectiveness of the system's data transmission and personalization.

Acknowledgements

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name: Global Talent for Generative AI Copyright Infringement and Copyright Theft, Project Number: RS-2024-00398413, Contribution Rate: 100%).

Contribution Statement

Mingyu Jang and Kyungjune Lee are equally contributed to this work. Sanghoon Lee is the corresponding author of this work.

References

[Abdal *et al.*, 2021] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021.

[Ahn *et al.*, 2018] Sewoong Ahn, Woojae Kim, Jinwoo Kim, Jaekyung Kim, and Sanghoon Lee. Visual preference prediction for enhanced images on ultra-high-definition display. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3548–3552. IEEE, 2018.

[Bai *et al.*, 2023] Haoran Bai, Di Kang, Haoxian Zhang, Jinshan Pan, and Linchao Bao. Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 362–371, 2023.

[Daněček *et al.*, 2022] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022.

[Dastjerdi *et al.*, 2023] Mohammad Reza Karimi Dastjerdi, Jonathan Eisenmann, Yannick Hold-Geoffroy, and Jean-François Lalonde. Everlight: Indoor-outdoor editable hdr lighting estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7420–7429, 2023.

[Deng *et al.*, 2019] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[Jang *et al.*, 2022] Mingyu Jang, Seongmin Lee, Jiwoo Kang, and Sanghoon Lee. Technical consideration towards robust 3d reconstruction with multi-view active stereo sensors. *Sensors*, 22(11):4142, 2022.

[Kachur *et al.*, 2020] Alexander Kachur, Evgeny Osin, Denis Davydov, Konstantin Shutilov, and Alexey Novokshonov. Assessing the big five personality traits using real-life static facial images. *Scientific Reports*, 10(1):8487, 2020.

[Kartynnik *et al.*, 2019] Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile gpus. *In Third Workshop on Computer Vision for AR/VR*, 2019.

[Kim *et al.*, 2016] Haksub Kim, Sewoong Ahn, Woojae Kim, and Sanghoon Lee. Visual preference assessment on ultra-high-definition images. *IEEE Transactions on Broadcasting*, 62(4):757–769, 2016.

[Kim *et al.*, 2021] Woojae Kim, Anh-Duc Nguyen, Jinwoo Kim, Jongyoo Kim, Heeseok Oh, and Sanghoon Lee. Diverse and adjustable versatile image enhancer. *IEEE Access*, 9:80883–80896, 2021.

[Lee *et al.*, 2023] Kyungjune Lee, Jeonghaeng Lee, Hyucksang Lee, Mingyu Jang, Seongmin Lee, and Sanghoon Lee. Faceclone: Interactive facial shape and motion cloning system using multi-view images. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 512–513. IEEE, 2023.

[Li *et al.*, 2017] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

[Li *et al.*, 2018] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 645–653, 2018.

[Li *et al.*, 2021] Tianye Li, Shichen Liu, Timo Bolkart, Jiayi Liu, Hao Li, and Yajie Zhao. Topologically consistent multi-view face inference using volumetric sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3824–3834, 2021.

[Paysan *et al.*, 2009] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009.

[Wang *et al.*, 2020] Wencheng Wang, Xiaojin Wu, Xiaohui Yuan, and Zairui Gao. An experiment-based review of low-light image enhancement methods. *Ieee Access*, 8:87884–87917, 2020.

[Yang *et al.*, 2023] Xingchao Yang, Takafumi Taketomi, and Yoshihiro Kanamori. Makeup extraction of 3d representation via illumination-aware image decomposition. In *Computer Graphics Forum*, volume 42, pages 293–307. Wiley Online Library, 2023.

[Zheng *et al.*, 2023] Yujian Zheng, Zirong Jin, Moran Li, Haibin Huang, Chongyang Ma, Shuguang Cui, and Xiaoguang Han. Hairstep: Transfer synthetic to real using strand and depth maps for single-view 3d hair modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12726–12735, 2023.