Few-Shot 3D Face Generation via a Controllable Diffusion Model Guided by Text and Images –Supplemental Material–

Anonymous ICME submission

In this material, we supplemented details of related techniques and the texture argumentation module we use to further improve the quality and alignment of generated PBR materials. Furthermore, we evaluated the generation diversity of ControlFace and compared other alternatives of geometry proxy such as landmarks in UV space.

I. Related techniques

A. Personalized Style Adaptation: LoRA

Addressing the challenge of integrating personalized styles, often characterized by a limited set of reference design paintings, into generic stable diffusion models presents significant difficulties. There is a large gap between the amount of a typical style reference set and the data used in generic diffusion model training, which leads to overfitting and implausible generation with naive fine-tuning strategies.

To enable personalized stylized generation, we introduce a style control module based on LoRA [1]. This module injects a compact trainable module into each transformer layer while freezing the main part of the pre-trained stable diffusion models. For a pre-trained weight matrix W_0 which is frozen and does not receive gradient updates, LoRA updates the low-rank decomposition by a trainable A and B matrix, where $B \in R^{d\times r}$, $A \in R^{r\times d}$. Here, r is the rank and d is the dimension. Given the input x, the forward pass yields:

$$h = W_0 x + \Delta W x = W_0 x + BAx \tag{1}$$

For a given personalized style, we train the LoRA module on a dataset consisting of around 20 reference images with the desired target style. The trained style control module is responsible for extracting the character style information and collaborates with our 3D-aware generative model to produce high-fidelity 3D characters with the specified style. Notably, the style reference images are not limited to portrait images but can encompass a broader range of general images, such as paintings. This flexibility further expands the potential applications of our approach.

B. Image-guided Generation: IP-Adapter

While text descriptions can guide our generative model to produce diverse and imaginative results, their capability to provide precise control over identity-specific or spatial features is limited. We introduce an image guidance module that leverages input portrait images to guide our generative model in reproducing digital humans with similar characteristics to the 2D input image. Note that the image and text prompts are independent and can be used together to jointly influence the digital human generation process.

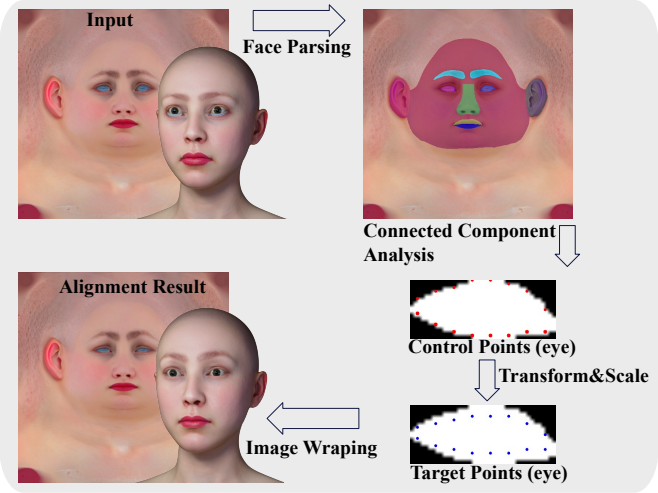

Fig. 1. Texture Alignment Pipeline. "Face Parsing" is implemented by a face parsing network to obtain a face segmentation. "Connected Component Analysis" is to find control points(red points) and target points(blue points). Finally, "Image Wrapping" is used to execute non-rigid transformations based on control and target points. Employing two-step texture alignment, we achieved precise alignment of the texture at the corners of the eyes.

We adopt the IP-Adapter architecture [2] to handle the image prompt. For an input image prompt $p_i \in \mathbb{R}^{H\times W\times C}$, we first extract the image features $e_{p_i} = \mathcal{E}_I(p_i)$ using the CLIP image encoder $\mathcal{E}_I$. We then project the image features e_{p_i} into the style features $f_{p_i} \in \mathbb{R}^{N\times 4}$ via a linear layer with layer normalization. These style features are incorporated into the frozen denoising Unet model through a decoupled cross-attention module.

Given the style features f_p extracted from the reference image, the text features f_t and the query features q, the output of decoupled cross-attention module O is as follows:

$$O = Softmax(\frac{QK^T}{\sqrt{d}})V + Softmax(\frac{Q(K^{'})^T}{\sqrt{d}})V^{'} \tag{2}$$

where $Q = qW_q$, $K = f_tW_k$, $V = f_tW_v$ are the query, key, and values matrices of the freezen UNet attention operation respectively, $K^{'} = f_pW_k^{'}$, $V^{'} = f_pW_v^{'}$ are the query, key, and values matrices of the new cross-attention layer which are trainable in the decoupled cross-attention module. In the

Fig. 2. Samples with same text prompts but different seeds. We sample several appearances of common persons and Orcs for the male face geometry and Na’vi and aliens for the female face geometry. ControlFace shows high diversity under the same text guidance.

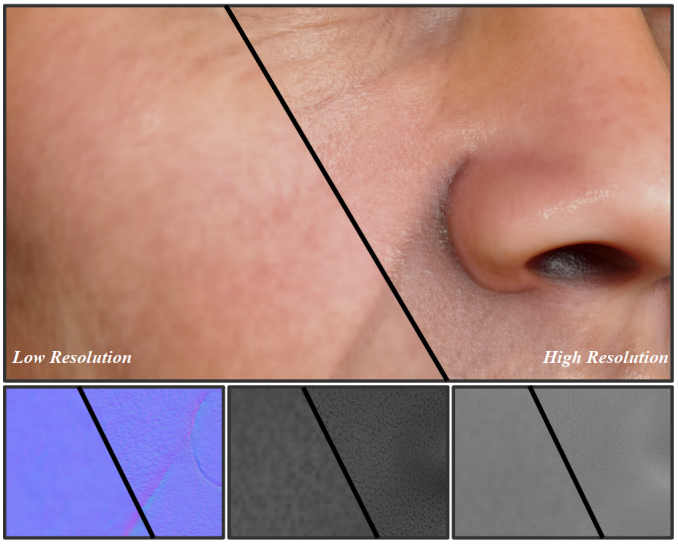

Fig. 3. Visualization of super-resolution PBR Assets. The first row illustrates a comparison of rendering results between the utilization of low-resolution PBR materials and high-resolution PBR materials. The second row presents a comparison of normal, specular, and roughness maps between low and high resolution.

decoupled cross-attention module, the same query is utilized for both image and text cross-attention.

II. Texture Augmentation

A. Texture Alignment: TPS

Although our 3D-aware control condition provides strong 3D-consistent priors, we observe minor pixel-level misalignments in the texture space, particularly in areas such as the eyes and mouth corners. To address this, we integrate a two-step process involving facial parsing and non-rigid transformation techniques to enhance the alignment. First, we employ a facial parsing algorithm [3] to accurately segment and identify distinct facial features within the texture space. Subsequently, along the boundaries of the connected regions, we place control points at intervals of every 5 pixels in the horizontal direction. Then, we apply a proportional scaling towards the center based on preset height and width thresholds to obtain target points. Following segmentation, we apply the Thin Plate Spline (TPS) algorithm [4] to execute non-rigid transformations based on the control and target points. TPS allows us to deform and adjust the segmented facial features, aligning them perfectly within pre-annotated regions in the texture space. The main pipeline of Texture Alignment is illustrated in Fig. 1.

B. Two-Stage Texture Super-Resolution Module

To further enhance the realism of the rendering, we enhance the material maps from 512×512 resolution to $4,096 \times 4,096$ resolution. We observed that training an $8\times$ super-resolution network to produce pore-level details is time-consuming and challenging. Therefore, we introduce a two-stage super-resolution strategy. Firstly, we upsample the resolution from 512×512 to 1024×1024 to restore the main facial features. Secondly, we further employ a $4\times$ upsampling to produce pore-level details.

III. Training Details

A. Data Pre-process

High-quality 3D faces are scarce, especially the exquisite geometry and corresponding PBR textures. These are not satisfied by the public dataset. Therefore, we obtain our dataset

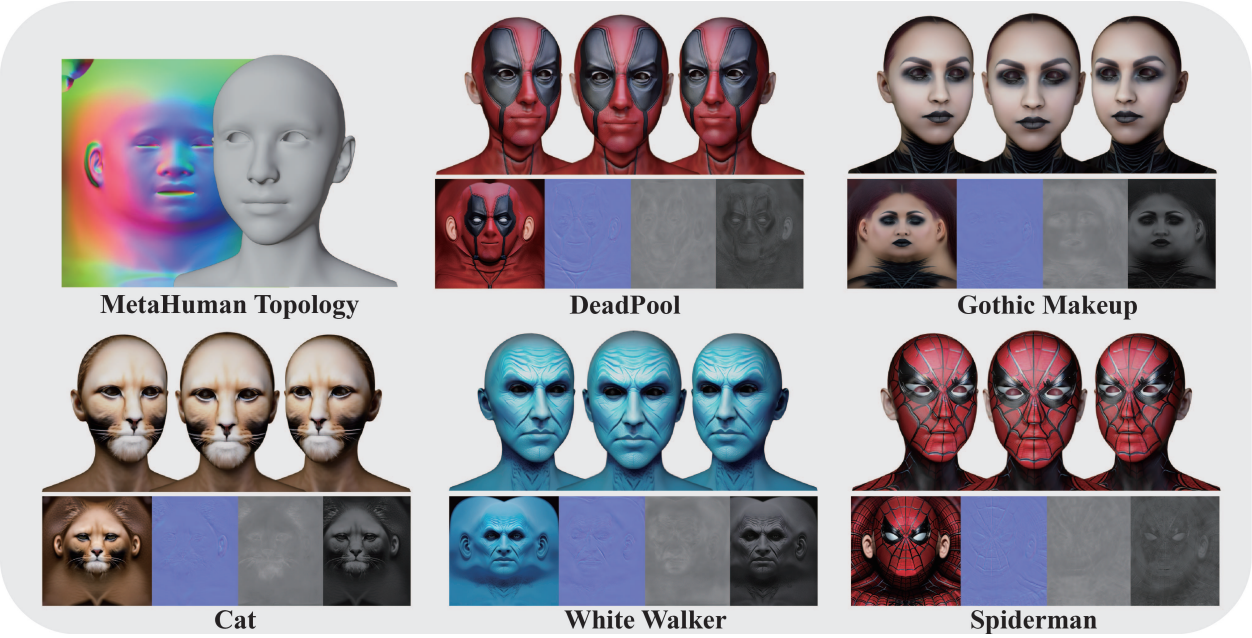

Fig. 4. Extra result of Zero-Shot topology control. We take the geometry normal map of MetaHuman topology as input(upper-left) and generate different faces in the corresponding topology, which can be verified by the textures under each case.

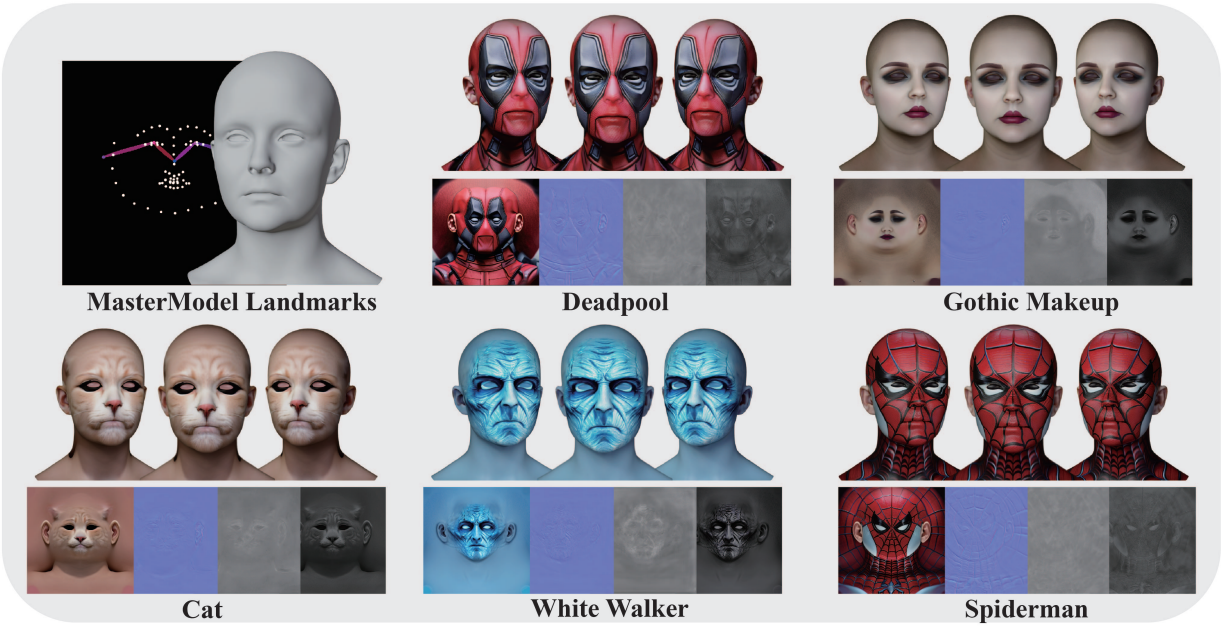

Fig. 5. Extra result of landmarks control. Following the experiments for zero-shot topology control, we sample faces under the same prompts. Under the control of Master Model topology landmarks (upper-left), ControlFace is capable of generating out-of-domain faces like the geometry normal map does.

from commercial 3DScanStore. With 36 faces, we unified all these samples into our facial topology including 20,971 vertices and 41,836 faces, and carried out manual annotations. To increase data diversity and mitigate mode collapse, we implemented pairwise alpha blending, yielding 272 training samples. For geometry proxy, we utilize the FaRL framework [3] to detect 68 facial landmarks and employ a rasterization-based renderer to render the normalized geometry normal of the geometry proxy into UV texture space. The XYZ channels of the rendered texture represent the orientations of the left, upper, and frontal directions, respectively.

To further enhance the training data, we employ the following data augmentation techniques. Firstly, we use facial landmarks as control points and group them based on facial semantics. Then, we utilize two different translation strategies for data augmentation. One entails shifting all landmarks

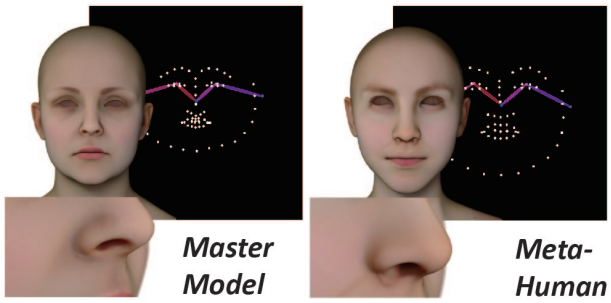

Fig. 6. Comparison of out-of-domain topology results under other geometry proxy control. Our new model trained on Master Model landmarks fails in synthesizing out-of-domain topology (MetaHuman).

horizontally or vertically by 0-20 pixels. The second strategy is to apply translations to each semantic group. Each semantic group is randomly translated horizontally or vertically by 0-20 pixels, ensuring no intersection between semantic groups. We consider the translated control points as target points. Subsequently, we use the TPS algorithm to warp both the diffuse albedo and the 3D-aware control condition based on paired control points and target points. This data augmentation strategy allows our method to adapt to various geometry topologies without the need for retraining.

B. Training

1) 3D-aware Albedo Diffusion: During training, we freeze the pre-trained parameters of the UNet in the stable diffusion model and fine-tune the trainable parameters in ControlNet. Specifically, we initialize our control module using the official pre-trained models $control\_v11p\_sd15\_normalbae$ [1]. Our experiments are based on SD v1.5 [2]. To enhance the details and realism of our generated results, we employed the pre-trained checkpoint epicrealism_pureEvolution [3], as the base model to train our ControlFace. In the training process, we randomly drop out part of the prompts with a probability of 50%. This approach directed the network to concentrate more on the input 3D-aware conditions.

2) Detailed Geometry&PBR Material Generation: We noticed that training the albedo-to-material mapping using relatively low-resolution images leads to a noticeable loss of fine-scale skin details. To mitigate this issue, we crop patches of size 512×512 from high-resolution material maps. These patches are generated with varying positions, scales, and flip orientations. Finally, we employ linear blending across multiple instances of the data to enhance the robustness and generalization capabilities of our model.

3) Super-Resolution module: In our implementation, we follow the network architecture of Real-ESRGAN [5] in both stages. Note that each material map is enhanced separately. The training loss is a combination of L_1 loss, perceptual loss based on VGG19, and a GAN loss, as introduced by [5]. We finetune the Real-ESRGAN $2\times$ with the Unet for 10 epochs. We finetune the Real-ESRGAN $4\times$ Super Resolution network for 20k training iterations separately. The learning rates of both are 1×10^{-4}.

[1]https://huggingface.co/lllyasviel/control_v11p_sd15_normalbae

[2]https://huggingface.co/runwayml/stable-diffusion-v1-5

[3]https://civitai.com/models/25694?modelVersionId=94744

IV. Extra Results

A. Diversity

In addition to generating diverse faces under different text prompts, we further evaluate the diversity from different seeds. Under the fixed geometry and prompt, we sampled different albedo textures using different seeds in Fig. 2. As a result, ControlFace shows high diversity in generating various faces under different seeds by giving a certain prompt.

B. Physically-Based Rendering Assets

Constrained by the low resolution, PBR textures of 512 resolution fail to provide a wealth of skin details, leading to over-smoothing rendering. The outcomes of our Albedo-To-PBR module are depicted in Fig. 3. The results demonstrate that our two-stage super-resolution strategy restores pore-level skin details from low-resolution PBR textures, significantly enhancing the realism of the rendering outcomes.

C. Generation Results for Out-of-domain Topology

We only train our generative model on Master Model topology, without MetaHuman topology geometry, and we demonstrate the ability to generate out-of-domain faces and topologies in the main paper. Here, we further evaluate the effectiveness of topology control for generating out-of-domain faces. As shown in Fig. 4, ControlFace can also generate diverse faces under the control of another topology.

D. Other Geometry Proxy

Similarly, landmarks in UV texture can introduce the priors of topology in UV texture space, in a sparse way. We further evaluated the effectiveness of landmarks control for generating out-of-domain identities(see Fig. 5) and topology(see Fig. 6). Using landmarks as a condition can preserve the generative diversity from the pre-trained Stable Diffusion model, while it struggles to generate the texture in out-out-domain topologies, such as MetaHuman, due to the sparse representation compared with geometry normals.

References

[1] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[2] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," 2023.

[3] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen, "General facial representation learning in a visual-linguistic manner," *arXiv preprint arXiv:2112.03109*, 2021.

[4] Jean Duchon, "Splines minimizing rotation-invariant semi-norms in sobolev spaces," in *Constructive Theory of Functions of Several Variables: Proceedings of a Conference Held at Oberwolfach April 25–May 1, 1976*. Springer, 1977, pp. 85–100.

[5] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1905–1914.