3DGH: 3D Head Generation with Composable Hair and Face – Supplementary Material –

A DATASETS

Multi-view Captures. We fit linear blend shapes for our deformable hair geometry from 283 hairstyles captured in a studio environment. The studio capture setup is similar to the capture system presented by Cao et al. [2022] and Saito et al. [2024], where we obtain calibrated and synchronized multi-view images at a resolution of 4096×2668 through the use of 110 cameras. For each RGB image, it's preprocessed to produce the corresponding segmentation maps, which are used as supervision for our hair geometry fitting algorithm.

Single-view Images. We use the official checkpoint of PanoHead [An et al. 2023] as the portrait image generator to generate training images for our generative model. To generate images from PanoHead, we sample the input latent code $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, and camera pose Π with the yaw angle sampled from $\mathcal{U}(0, 2\pi)$, and pitch angle sampled from $\mathcal{U}(\pi/2 - 0.5, \pi/2 + 0.5)$. They are passed to PanoHead to generate and render RGB images at a resolution of 512×512, from which we use [Lin et al. 2021] to obtain the foreground mask at resolution 512×512, and finally apply [Zheng et al. 2022] to parse the image into a hair-face segmentation map at resolution 512×512. For each iteration, we perform these operations on the fly to obtain the training images, resulting in $25m$ images in total used for training.

B TRAINING DETAILS

Our network architecture follows the official implementation of StyleGAN2 [Karras et al. 2020], where each mapping network consists of 2 hidden layers. For the geometry mapping network, it consists of a single hidden layer of 512 hidden units and uses the softplus activation function. We modify the output convolution layers such that they produce a feature map of shape $256 \times 256 \times 32$. Subsequently, a series of lightweight MLP decoders are applied to map the output features at each texel to different Gaussian parameters, including position, rotation, scale, color, and opacity. The MLP decoder shares the same architecture as the geometry mapping network, but reduces the number of hidden units to 64.

Our model is trained from scratch using the Adam optimizer [Kingma and Ba 2014]. We use a learning rate of 0.0025 for the generator and 0.002 for the discriminator, leading to a stable training configuration in our case. Our model is trained for $25M$ images following Chan et al. [2022] with an effective batch size of 64, which takes around 4 days to train on 32 NVIDIA A100 GPUs with 80G of VRAM each. To fit hair geometries w.r.t. multi-view segmentation maps, we utilize the Adam optimizer as well, with a learning rate of 0.01 for the Jacobians and 0.2 for the centroid translation. On a single NVIDIA A100 GPU, the optimization process of 500 iterations for a single mesh finishes within 1 minute.

C QUALITATIVE ABLATION

In Fig. 1, we present qualitative comparisons of RGB images and hair-face segmentation maps rendered from models trained with different segmentation supervision configurations. *Seg. in* $\mathcal{D}$ refers to concatenating the rendered segmentation maps with the discriminator input, allowing the discriminator to adversarially evaluate whether the segmentation is realistic. However, our experiment reveals that this configuration is unstable during training and prone to model collapse in the early stages of GAN training, yielding meaningless outputs as shown in the left column. By contrast, simply removing the segmentation input and training the model without segmentation loss (*w/o Seg. loss*) surprisingly produces renderings with acceptable quality and reasonable hair-face segmentation, as illustrated in the middle column. We hypothesize that it is because our template mesh already provides a good initialization for hair and face Gaussians. The adversarial training then guides the model to refine these Gaussians with as minimal displacement as possible, leading to a certain level of separation even in the absence of explicit segmentation loss. Finally, incorporating our segmentation loss (*w/ Seg. loss*) produces results with a significantly clearer separation between hair and face Gaussians, as demonstrated in the right column.

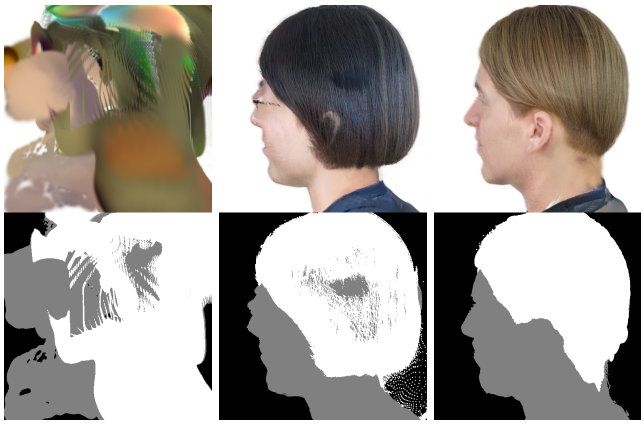

Fig. 1. Samples rendered from models trained with different segmentation supervision configurations.

In Fig. 2, we present qualitative comparisons of samples rendered from models trained with different hair geometry configurations. When the hair geometry is fixed, hair Gaussians tend to require larger deviations to represent varying hairstyles, leading to floating Gaussians being placed in random positions, which is particularly noticeable in the rendered segmentation maps shown in the middle column. When the hair geometry is deformable, shape variations are streamlined to the mesh itself, effectively constraining the spatial distribution of hair Gaussians to a tighter layer around the mesh surface. Consequently, the rendered segmentation maps exhibit fewer floating artifacts, and the hair details in the rendered RGB images are improved, as demonstrated in the left column.

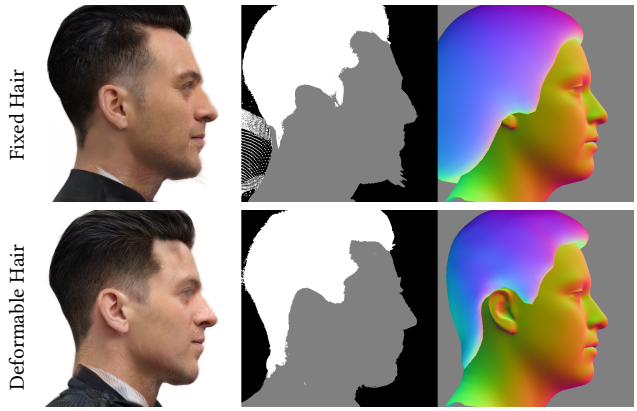

Fig. 2. Samples rendered from models trained with different hair geometry configurations.

In Fig. 3, we present qualitative comparisons of different hair-face correlation modules. When employing the concatenation mechanism, the composition results exhibit a strong dependency on face information, leading to minimal changes in the hairstyle when the face is fixed. In contrast, the cross-attention mechanism achieves a better balance between hair and face information, enabling the composition to more accurately preserve the reference hairstyle while maintaining the plausibility of the generated result.

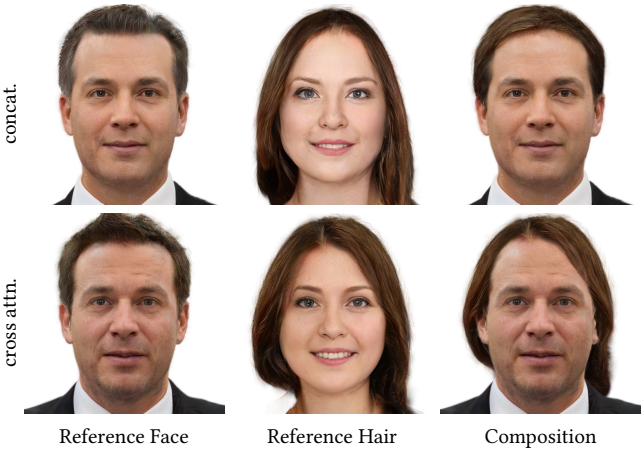

Fig. 3. Hair-face compositions from models trained with different hair-face correlation modules.

D ADDITIONAL RESULTS

In Figs. 4 to 7 we show uncurated samples generated from our method, including the rendered RGB images, hair-face segmentation maps, mesh normal maps, and 3D Gaussian visualization.

REFERENCES

Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. 2023. PanoHead: Geometry-Aware 3D Full-Head Synthesis in 360°. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20950–20959.

Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason Saragih. 2022. Authentic Volumetric Avatars from a Phone Scan. *ACM Trans. Graph.* 41, 4, Article 163 (July 2022), 19 pages.

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. 2022. Efficient Geometry-aware 3D Generative Adversarial Networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16123–16133.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* (2014).

Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2021. Real-Time High-Resolution Background Matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8762–8771.

Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. 2024. Relightable Gaussian Codec Avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 130–141.

Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. 2022. General Facial Representation Learning in a Visual-linguistic Manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18697–18709.

Fig. 4. Uncurated samples (RGB).

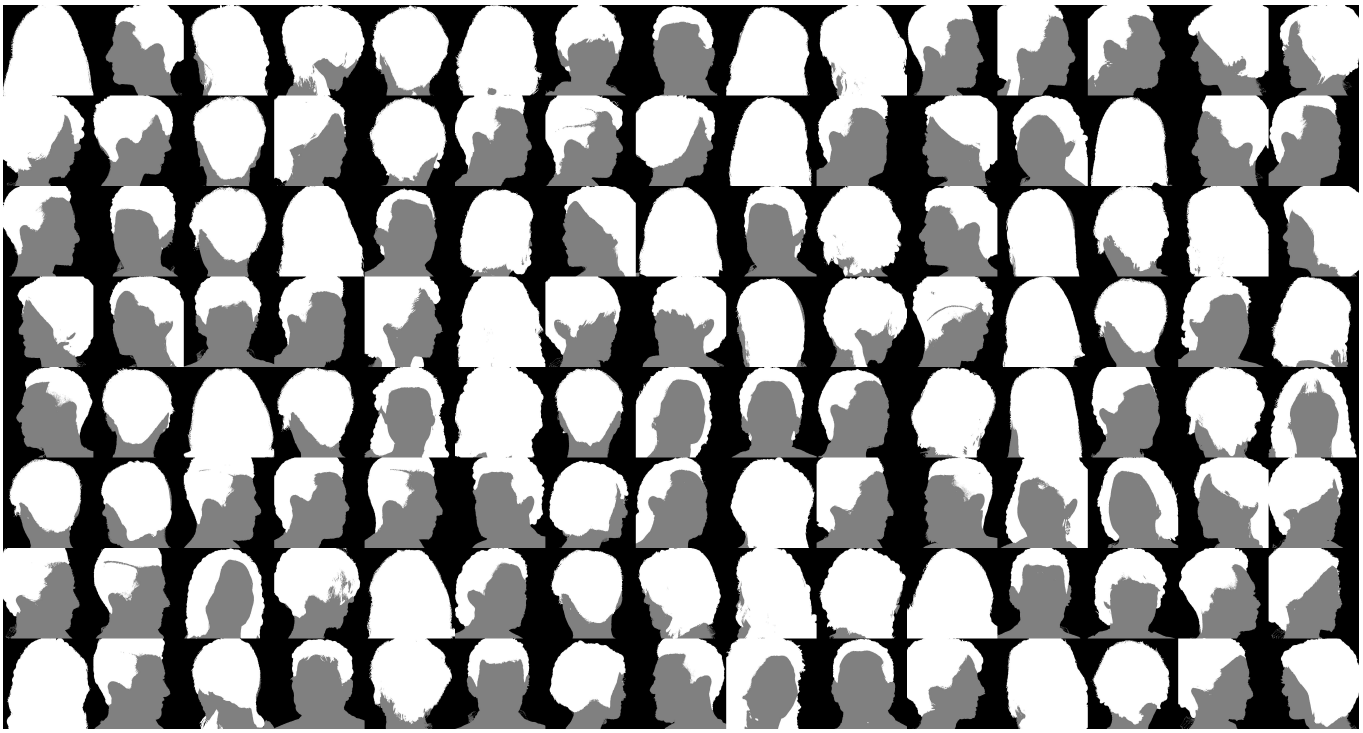

Fig. 5. Uncurated samples (Hair-face segmentation maps).

Fig. 6. Uncurated samples (Mesh normal maps).

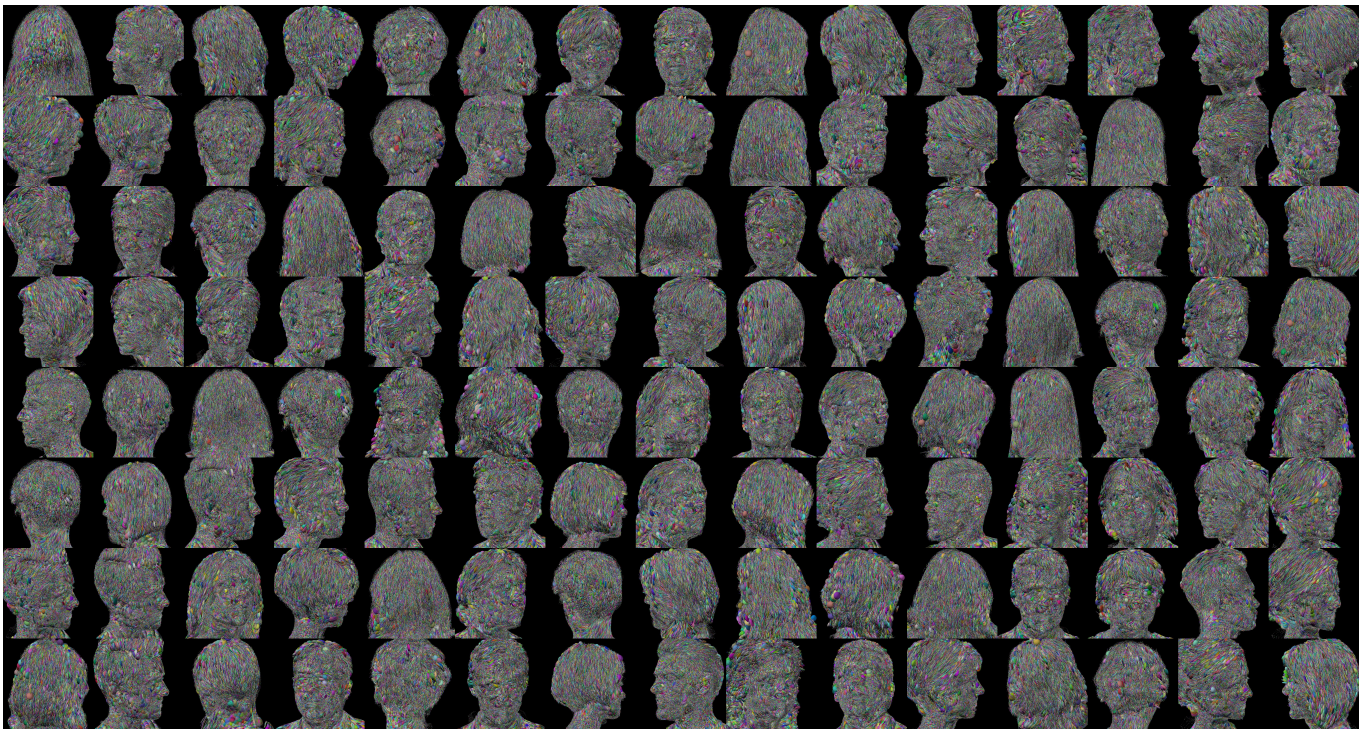

Fig. 7. Uncurated smaples (3D Gaussian visualization).