Neural Parameterization for Dynamic Human Head Editing *Supplementary Material*

1 COMPARISON WITH REGISTERED MESH

In order to further evaluate the performance of our method, we compared the novel view synthesis quality and the texture temporal consistency with registered meshes on two multi-view face datasets. We first compare on Facescape [Yang et al. 2020]. The dataset contains multi-view data of 359 subjects making 20 expressions. It also provides topologically uniformed textured meshes for each subject. We treat the multi-view sequence of each subject as a video of 20 frames, with each frame indicating one expression. We compare our method with the provided textured mesh on the first 100 subjects, and report PSNR, SSIM, LPIPS and ASTD in Tab. 1. The result shows that our method surpasses the registered mesh representation from the dataset in terms of PSNR, SSIM and ASTD for both novel view synthesis and temporal alignment.

We then compare our method on the provided sample dataset in Beeler et al. [2011], which contains one multi-view sequence of 347 frames, as well as densely tracked face mesh. We again compare our method with the mesh and report quantitative results in Tab. 2. Our method performs worse than the tracked mesh in terms of novel view synthesis. One possible reason is that this dataset only contains 7 views as input images, while NeRF-like representations that we adopt typically require more input views to get good novel view synthesis results. However, note the NeRF-like representations have benefits such as modeling hairs and mouth interior, which the mesh-based method fails to model.

2 TEMPORAL CONSISTENCY

Our method can generally obtain temporally consistent results. However, sometimes temporal flickering artifacts may arise in some cases due to the per-frame optimization and reconstruction errors during training. To encourage more temporally consistent results, we could optionally apply a 2^{nd} order temporal smooth loss. Specifically, for each element x of the geometry parameters (global rigid transformation $\mathbf{H}$, control points $\mathbf{s}_i$, influence radius r_i of each control point), we define the temporal smooth loss as:

$$\mathcal{L}_{temporal} = \sum_{x \in \{\mathbf{H}, \mathbf{s}_i, r_i\}} |x^{(t)} + x^{(t-2)} - 2x^{(t-1)}|. \qquad (1)$$

We show the reconstruction results with and without the temporal smooth loss in Fig. 1. We visualize the temporal consistency by flattening the time axis along the image row marked as green in Fig. 1. The results show that the temporal smooth loss could stabilize the head and produce fewer jittering artifacts.

3 CANONICAL SPACE VISUALIZATION

We visualize the canonical space by setting the geometry parameters of all the frames to be the same as the first frame. The result is shown in Fig. 3. It can be seen that the head is stabilized in the canonical space, which reduces the complexity of a dynamic 3D volume.

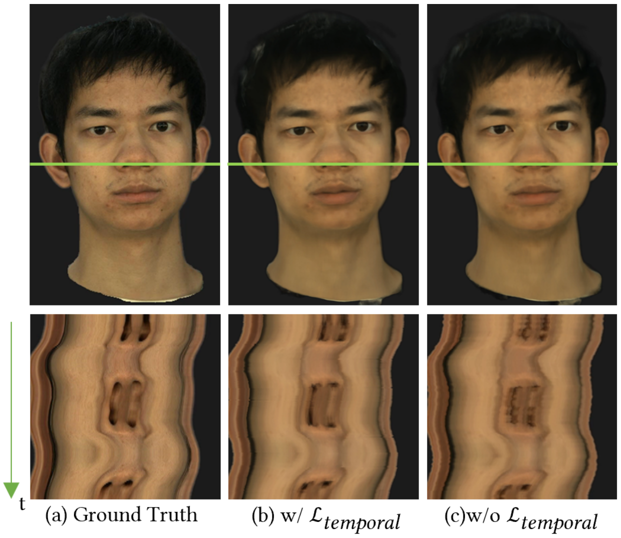

t
(a) Ground Truth
(b) w/ $\mathcal{L}_{temporal}$
(c)w/o $\mathcal{L}_{temporal}$

Fig. 1. Visualization of temporal consistency with and without temporal smooth loss. In the first row we visualize the novel view synthesis results of one frame, and in the second row we flatten the time axis along the green row. The t indicates the time axis.

4 MODEL OBJECTS ON HEAD

We tested our method on more challenging cases where the head is wearing accessories like earphones, glasses and hats. We show the reconstruction and editing results in Fig. 2. Thanks to the NeRF-like representation, our method successfully reconstruct the objects on head. The texture mapping is mostly plausible, but may suffer from mild distortion due to the complex topology.

5 IMPLICIT AND EXPLICIT TEXTURE DECOMPOSITION

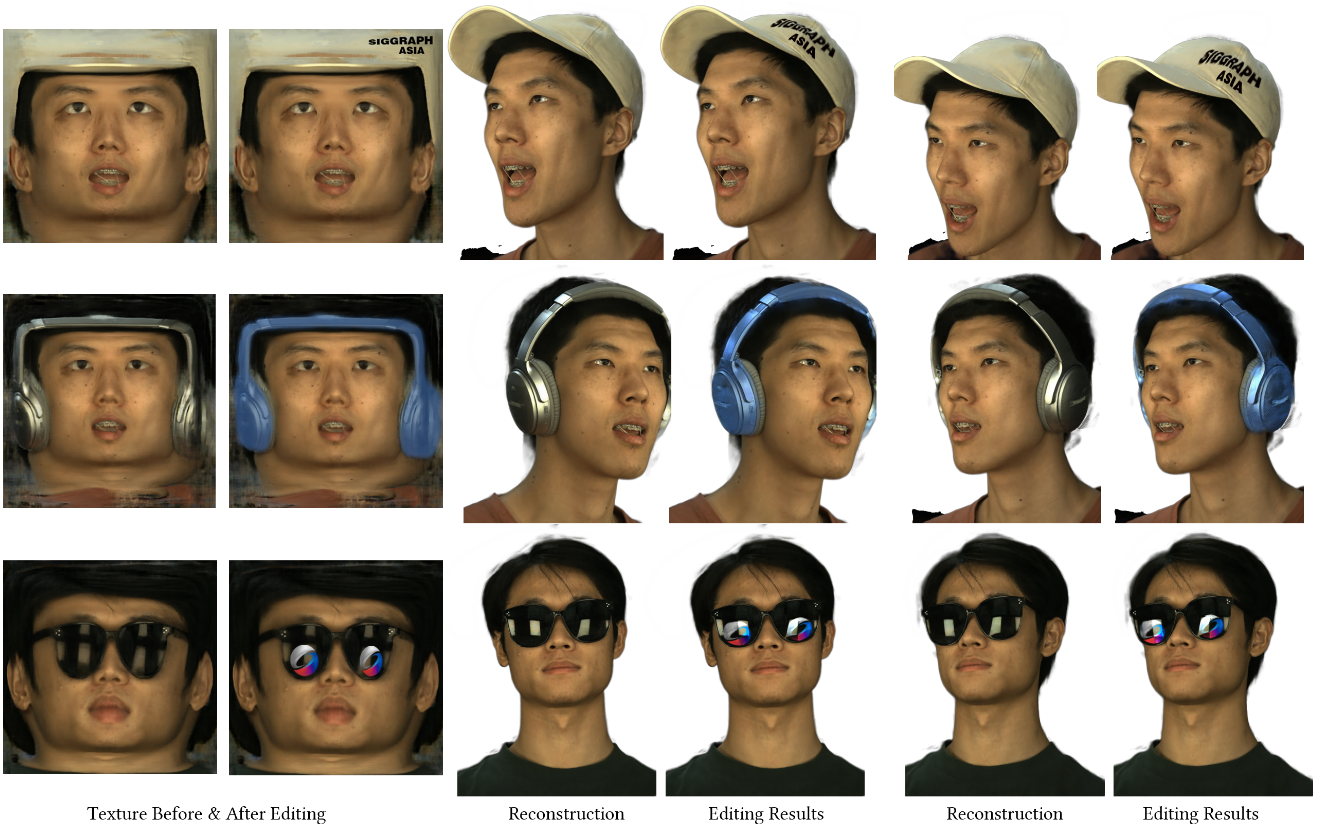
Ideally, the explicit texture should contain only the base color, while the implicit texture is supposed to model the residuals caused by time and view variations. However, it can be seen from Fig. 8 in the

Table 1. Quantitative results on the Facescape [Yang et al. 2020] dataset. We compare with the topologically uniformed (TU) meshes provided by the dataset.

	PSNR↑	SSIM↑	LPIPS↓	ASTD↓
Ours	23.61	0.6460	0.09677	5.535
TU Mesh	21.72	0.5759	0.05785	7.707

Table 2. Quantitative results on the Beeler [Beeler et al. 2011] dataset.

	PSNR↑	SSIM↑	LPIPS↓	ASTD↓
Ours	30.65	0.8128	0.2608	0.004617
Tracked Mesh	31.09	0.8794	0.05778	0.01372



Texture Before & After Editing

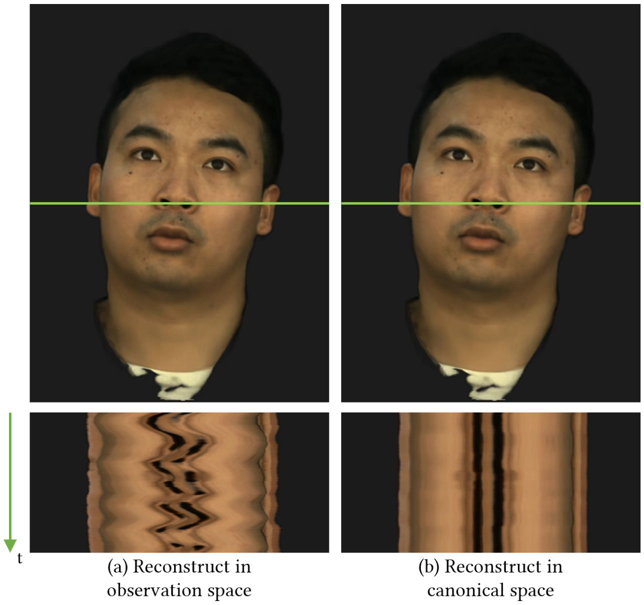
Reconstruction

Editing Results

Reconstruction

Editing Results

Fig. 2. Reconstruction and editing results when there are objects (earphones, sunglass, and hats) on heads.



(a) Reconstruct in observation space

(b) Reconstruct in canonical space

Fig. 3. Visualization of the dynamic human head in canonical space. The head is stabilized in the canonical space. The t indicates the time axis.

paper that even if we pick the correct value $\lambda_{sparsity}$ that achieves the best trade-off between texture decomposition and reconstruction quality, the implicit texture still contains some amount of base color, especially in regions like eyes. This is because it is really challenging to track a precisely aligned texture mapping, leading to temporal variations modeled by implicit texture. However, in our experiments we find this does not greatly affect the appearance editing results in most cases.

REFERENCES

Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W. Sumner, and Markus Gross. 2011. High-quality passive facial performance capture using anchor frames. *ACM Trans. Graph.* 30, Article 75 (August 2011), 10 pages. Issue 4. https://doi.org/10.1145/2010324.1964970

Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. 2020. FaceScape: a Large-scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*