Supplementary material for: High-Fidelity Eye Animatable Neural Radiance Fields for Human Face

Hengfei Wang
hxw080@student.bham.ac.uk
Zhongqun Zhang
zxz064@student.bham.ac.uk
Yihua Cheng✉
y.cheng.2@bham.ac.uk
Hyung Jin Chang
h.j.chang@bham.ac.uk

School of Computer Science
University of Birmingham
Birmingham, UK

In our paper, we propose Dynamic Eye-aware NeRF (DeNeRF) which is high-fidelity eye animatable neural radiance fields for human face. DeNeRF converts 3D points from various perspectives into a standard space to facilitate the learning of a cohesive facial NeRF model. We have devised an eye deformation field for this transformation, which encompasses both rigid transformations, such as eyeball rotation, and non-rigid transformations. Because of the limitation of pages, we show the details of the **multi-view face tracking loss** and its **visualization** in this supplementary material.

1 Multi-view Face Tracking Loss

We build face appearance loss $\mathcal{L}_{appear}$, facial landmark loss $\mathcal{L}_{face}$ and pupil center loss $\mathcal{L}_{pupil}$ for multi-view face tracking. All the losses are mean absolute errors. The losses are defined as follows:

$$\mathcal{L}_{appear} = \left(\sum_{i=1}^{N} \left| \hat{C}_i - C_i \right| \right) / N, \tag{1}$$

where N is the number of multi-view images from one frame and N is 13 in our setting, C_i is the colors of all pixels from true images in one frame. $\hat{C}_i$ is the colors of all pixels from rendered images in the same frame.

$$\mathcal{L}_{face} = \left(\sum_{i=1}^{N} \left| \hat{X}_i - X_i \right| \right) / N, \tag{2}$$

where $X_i \in \mathbb{R}^{68\times 2}$ is the 2D positions of 68 facial landmarks of true faces from one frame, $\hat{X}_i \in \mathbb{R}^{68\times 2}$ is the 2D positions of 68 facial landmarks of rendered faces from the same frame.

$$\mathcal{L}_{pupil} = \left(\sum_{i=1}^{N} \left| \hat{P}_i - P_i \right| \right) / N, \tag{3}$$

where $P_i \in \mathbb{R}^{2\times 2}$ is the two pupil center positions of true faces from one frame. $\hat{P}_i \in \mathbb{R}^{2\times 2}$ is the two pupil center positions of rendered faces from the same frame.

2 Multi-view Face Tracking Visualization

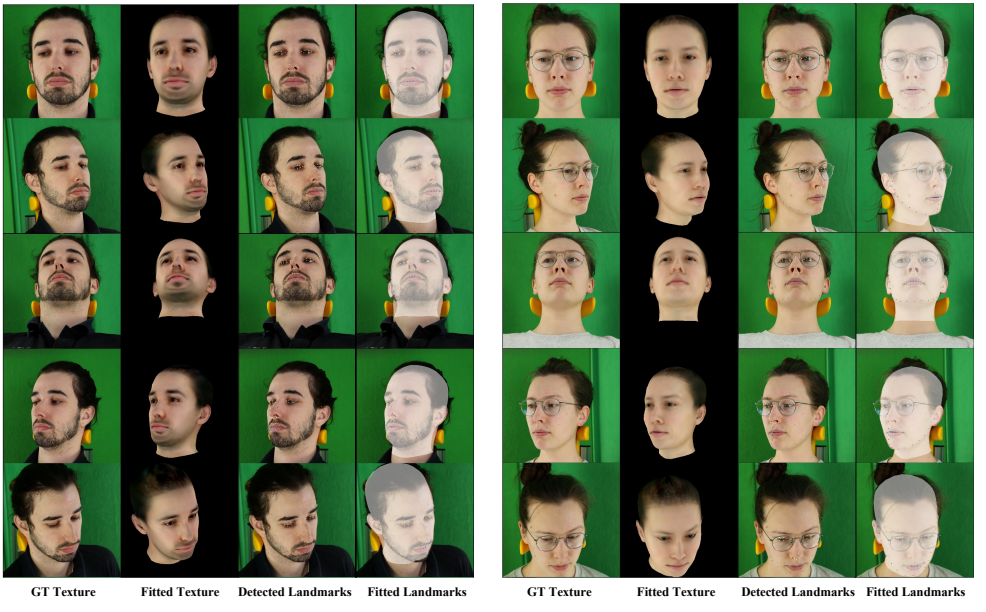

Figure 1: Fitting results of multi-view face tracking.

We show the visualization of face tracking in Fig. 1. The detected landmarks used for supervision are estimated using [1]. Our multi-view face tracker takes multi-view images and corresponding camera poses as inputs. We initialize the textured FLAME face model with zero parameters and project the face model to all views. We render face images and obtain face landmarks and pupil centers from the fitted FLAME model. Then we perform pixel-wise alignment in the face appearance and position alignment for facial landmarks and pupil centers. The result shows that our multi-view face tracking can get accurate facial parameters while keeping multi-view consistency.

3 Novel Gaze Rendering

We show some novel gaze rendering results in the attached video. It shows that our model can realize continuous eye animation while trained on frames with only 9 different gaze directions. This can be attributed to the precise positioning of the eyeball, achieved through

multi-view face tracking using FLAME [2], as well as the deformation strategy employed in canonical space. Besides, our model can reconstruct glasses and the light reflection on them accurately. It shows the good potential of our model as a facial generative method.

References

[1] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[2] 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.