GazeNeRF: 3D-Aware Gaze Redirection with Neural Radiance Fields
Supplementary Material

1. Overview

In this supplementary material, we first provide more details of the data pre-processing and training procedure. We also show another ablation study results on loss components and a comparison between GazeNeRF trained from scratch and the pre-trained HeadNeRF [2] model. We then show additional qualitative results on different datasets. Furthermore, we show the results of the few-shot personal calibration experiments. We encourage the readers to also watch the supplementary video that contains more animated results of the proposed method.

2. Details of data pre-processing and training precedure

The original resolution of images from ETH-XGaze [6] is 6K×4K, and the resolutions of the images from other datasets are different from each other. To unify them, we pre-process the images with the data normalization method in [7], where the rotation and translation between the camera and face coordinate systems are standardized. We fix the normalized distance between the camera and the center of the face to 680mm. To centralize the faces in the normalized images, we use different values for the focal lengths for the normalized camera projection matrices, which are 1600, 1400, 1600 and 1200 for ETH-XGaze [6], MPIIFaceGaze [8], ColumbiaGaze [5] and GazeCapture [3], respectively.

To obtain the 3DMM parameters and the masks of the eyes and the face only regions, we use the face parsing model in [10] to segment the whole face. For some images, we also use the face parsing model in [4] and facial landmarks [1] to determine the eye masks only when the face paring model [10] returns empty results for the eyes.

GazeNeRF is trained with a single NVIDIA A40 GPU for one week. During inference, we fine-tune GazeNeRF and update four learable latent codes using a single image. Fine-tuning takes around one minute, and generating new image in one second.

3. Ablations on loss components

In this section, we show another ablation study on the contributions of different loss components. We train another three baseline GazeNeRF with different loss components. Here the baseline GazeNeRF represents the structure of GazeNeRF, which is *Two-stream+rotation*. We take the reconstruction loss as the base and verify the power of different loss components in an additive way. The results are listed in Tab. 1 and evaluated on ETH-XGaze dataset.

The results show that only using reconstruction loss achieves the worst performance regarding all evaluation metrics. Adding the perceptual loss boosts the performance in all metrics, especially gaze and head pose angular errors. Moreover, adding the disentanglement loss achieves the best performance in the most of evaluation metrics. Utilizing the functional loss helps to drop the gaze angular error of GazeNeRF at the cost of image quality (*e.g.* FID) and person identity.

4. Comparison between GazeNeRF trained from scratch and the pre-trained model

Tab. 2 shows the evaluation results between GazeNeRF trained from scratch and the pre-trained HeadNeRF model [2]. We can find that training with the pre-trained model helps improve the head pose error at the cost of the gaze angular error. Regarding the image quality and identity similarity, both models conduct the similar performance.

5. Personal calibration for gaze estimation

In this section, we demonstrate how GazeNeRF is beneficial for the downstream task of person-specific gaze estimation in a few-shot setting. Specifically, given a few calibration samples from person-specific test sets, we augment these real samples with gaze redirected samples generated by GazeNeRF. We then fine-

*These two authors contributed equally to this work.

	Gaze↓	Head Pose↓	SSIM↑	PSNR↑	LPIPS↓	FID↓	Identity Similarity↑
Baseline GazeNeRF $+\mathcal{L}_{\mathcal{R}}$	28.122	21.489	0.683	14.150	0.406	146.943	20.499
Baseline GazeNeRF $+\mathcal{L}_{\mathcal{R}}+\mathcal{L}_{\mathcal{P}}$	8.861	3.456	0.729	15.370	0.290	72.133	**49.265**
Baseline GazeNeRF $+\mathcal{L}_{\mathcal{R}}+\mathcal{L}_{\mathcal{P}}+\mathcal{L}_{\mathcal{D}}$	8.460	**3.386**	0.729	**15.461**	**0.288**	**72.044**	48.705
GazeNeRF (Baseline GazeNeRF $+\mathcal{L}_{\mathcal{R}}+\mathcal{L}_{\mathcal{P}}+\mathcal{L}_{\mathcal{D}}+L_{\mathcal{F}}$)	**6.944**	3.470	**0.733**	15.453	0.291	81.816	45.207

Table 1. Ablation study on different loss components.

	Gaze↓	Head Pose↓	SSIM↑	PSNR↑	LPIPS↓	FID↓	Identity Similarity↑
GazeNeRF (Two-stream + rotation + $L_{\mathcal{F}}$)	**6.944**	3.470	**0.733**	**15.453**	0.291	81.816	**45.207**
GazeNeRF + pre-trained HeadNeRF	7.134	**3.080**	0.732	14.761	**0.285**	**76.293**	43.443

Table 2. Comparison of GazeNeRF trained from scratch to pre-trained HeadNeRF model.

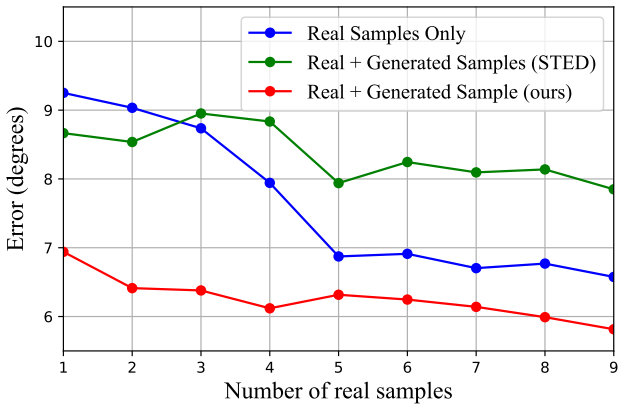

Figure 1. Downstream personal gaze estimation task in a few-shot setting. The x-axis is the number of real samples used, and the y-axis is the gaze estimation error in degree. The results are calculated by averaging the angular error between the 15 subjects of the ETH-XGaze person-specific set. We show the result of only using real samples (blue), using real plus generated samples from STED (green), and using real plus generated samples from our GazeNeRF (red) to fine-tune the pre-trained gaze estimator.

tune the gaze estimator pre-trained on ETH-XGaze's training set with these augmented samples and compare the performance with the baseline model that is fine-tuned only with real samples. To eliminate the influence of the number of samples, the size of augmented samples is always 200 (real + generated samples). We change the number of real samples used for the fine-tuning during the evaluation.

The result is shown in Fig. 1, where the x-axis is the number of real samples used and the y-axis is the gaze estimation error in degree on the ETH-XGaze person-specific test set. We test up to nine real samples for the few-shot setting. Observe from the figure that fine-tuning the pre-trained gaze estimator with both real and generated samples from GazeNeRF brings a significant improvement in gaze error versus only fine-tuning with real samples. This trend is more evident when fewer real samples are available. It indicates that the generated sample from GazeNeRF is of high fidelity in terms of its gaze angle, such that it can be helpful to improve the downstream gaze estimation accuracy. In Fig. 1 we also compare the result of few-shot personal calibration when the generated samples come from STED [9]. ST-ED performs the worst in this case. It shows that the 2D generative model is less helpful for the downstream gaze estimation task, which is due to the lack of consideration of the 3D nature of the gaze redirection task.

6. Additional qualitative results

In Fig. 2 we show additional qualitative results of GazeNeRF and the SOTA baselines, evaluated on the person-specific test set of the ETH-XGaze dataset.

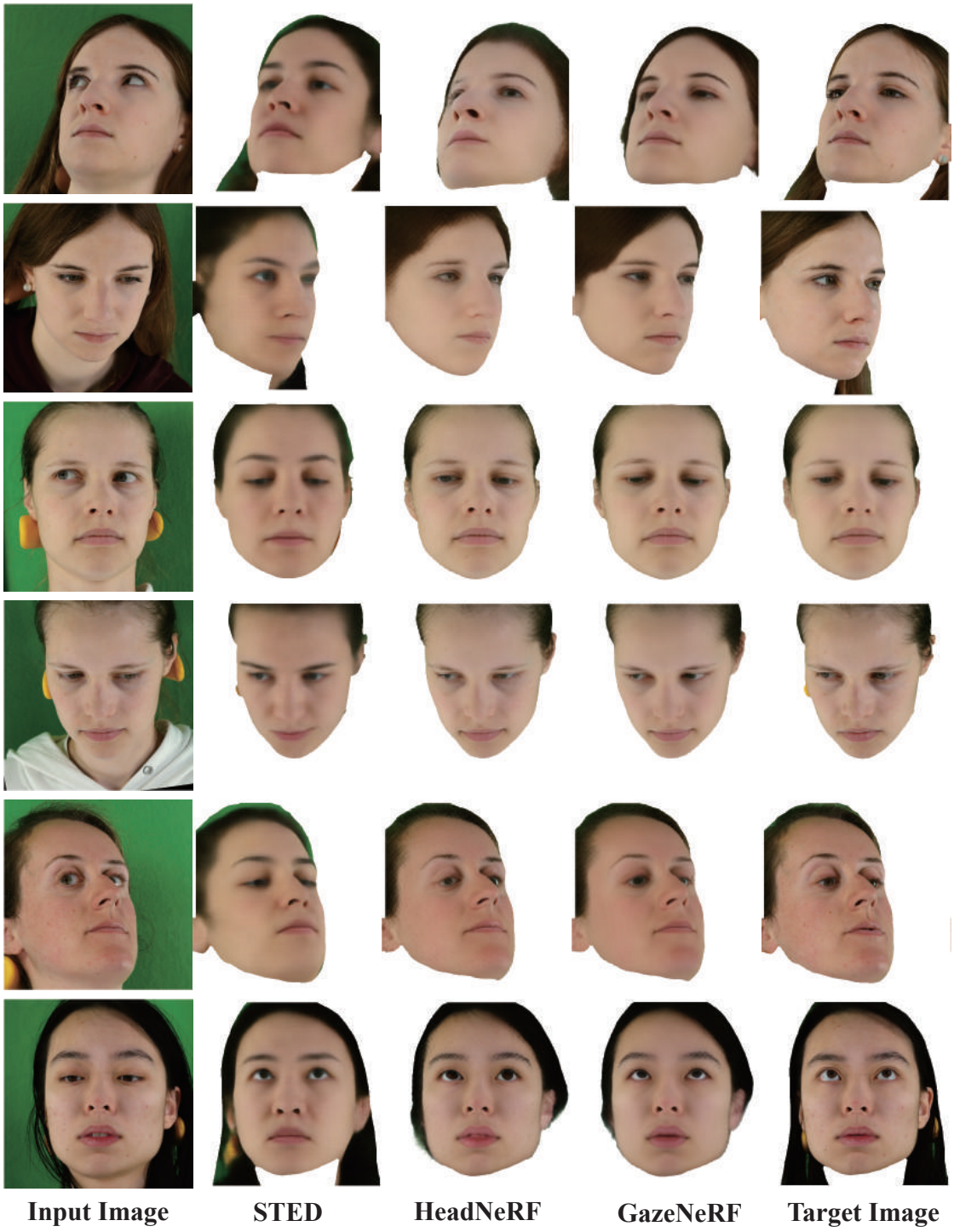

Figure 2. Additional visualization of generated images from ETH-XGaze with our GazeNeRF, STED and HeadNeRF. All faces are applied with face masks to remove the background. Our GazeNeRF can generate photo-realistic face images with different gaze directions and head poses. STED suffers from losing identity information, and HeadNeRF cannot generate fine-grained eyes.