Supplementary Material
RTGaze: Real-Time 3D-Aware Gaze Redirection from a Single Image

Hengfei Wang[1], **Zhongqun Zhang**[1,2], **Yihua Cheng**[1,†], **Hyung Jin Chang**[1]

[1]University of Birmingham
[2]College of Software, Nankai University
hengfei_wang@163.com, zhangzhongqun@nankai.edu.cn, {y.cheng.2, h.j.chang}@bham.ac.uk

Details of Data Pre-processing and Training

The resolution of raw images in ETH-XGaze (Zhang et al. 2020) is 6Kx4K. We first normalize the raw images using the method in (Zhang et al. 2020) and get the normalized head poses and gaze directions. The normalized distance between the camera and the center of the face is fixed at 950mm and the focal length in the normalized camera projection matrice is set to 1600. To align the data format with Live3D (Trevithick et al. 2023) and EG3D (Chan et al. 2022), we resize the normalized images to 512x512 and estimate camera poses using the model in (Deng et al. 2019). To apply our mask-guided 2D constraint, we use the face parsing model (Yu et al. 2018) to segment the whole and the eye region. We use the detected landmarks (Bulat and Tzimiropoulos 2017) to do the segmentation when the face parsing model does not work when processing some challenging images.

The personalized test set in ETH-XGaze includes 200 labeled images for each subject. We split the personalized test set into an input group and a target group following GazeNeRF (Ruzzi et al. 2023). The input group contains 100 images for each subject and the target group includes the other 100 images. We train our model with 80 subjects in the train set of ETH-XGaze first and then finetune the model with images from the input group for 10 epochs. We generate the images in the target group during evaluation.

Analysis of Redirection Accuracy in Pitch and Yaw Directions

To further analyze the performance of our model in different directions, we calculated gaze errors in pitch and yaw directions and present the error maps in degrees in Fig. 1. The mean errors in pitch and yaw directions are $5.085°$ and $5.989°$, respectively. Pitch redirection error increases at larger gaze angles due to decreasing data density in the ETH-XGaze dataset. Yaw redirection error rises as the pitch angle shifts from positive to negative, as downward gaze reduces eyeball visibility, making yaw redirection more challenging. Regarding the influence of input pitch angle, blurring is observed in the generated eye region at extreme pitch angles, due to limited training data with large pitch angles.

†Corresponding author.

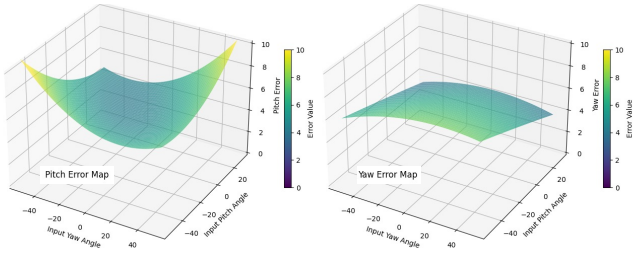

Figure 1: Error maps in pitch and yaw directions.

Improve Gaze Estimation

We conducted an evaluation of our gaze estimation model using the ETH-XGaze dataset. Initially, we trained the gaze estimator on the training set provided by ETH-XGaze. Subsequently, we performed an evaluation of the model by testing it on the person-specific test set in the ETH-XGaze dataset. We generated 100 images for each subject in the person-specific set. Then we implemented a fine-tuning process utilizing the synthesized images. The gaze error rate was significantly reduced from $5.975°$ to $4.758°$. This notable improvement in accuracy underscores the efficacy of our model and highlights its potential for precise gaze estimation.

Ablations on Cross-Attention Fusion and Triplane Decoder

To demonstrate the effectiveness of our network design, we conducted extra ablation studies on two key components: the cross-attention fusion and the triplane decoder. For the ablation, we replaced the cross-attention module with a simple concatenation operation and the triplane decoder with conventional convolutional layers. For each study, we retrained the model and evaluated its performance on the person-specific set of the ETH-XGaze dataset. The results, presented in Table 1, show that ablating the cross-attention module significantly decreases performance. This suggests that simple concatenation is insufficient for effectively fusing gaze embeddings and facial features, which is crucial for accurate gaze redirection based on the input gaze prompt. Furthermore, removing the triplane decoder caused the model to fail to converge, highlighting the importance of the triplane decoder for successfully decoding 3D facial information from the triplane

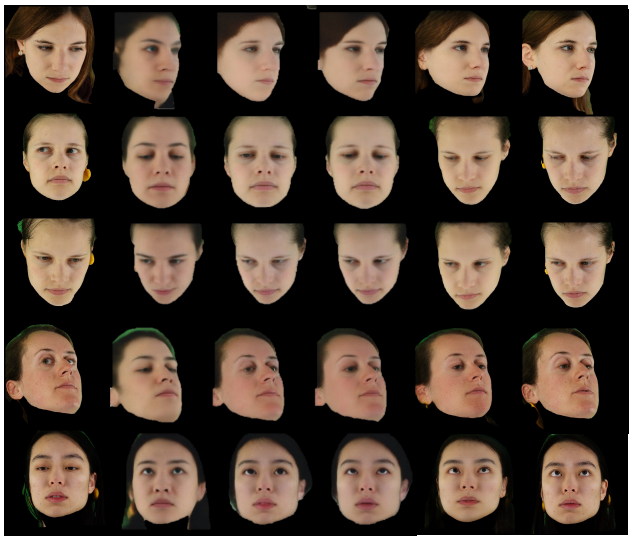

Figure 2: Additional visualization of generated images from ETH-XGaze with our RTGaze, ST-ED, HeadNeRF, and GazeNeRF. The background is eliminated using face masks. Our model can generate photo-realistic images with extensive details. In contrast, ST-ED struggles to preserve identity information. HeadNeRF and GazeNeRF face challenges in maintaining facial details.

representation.

	Gaze↓	FID↓	ID↑
w/o Cro.	21.188	48.574	56.566
w/o Tri.	Null	Null	Null
Ours	**9.047**	**38.346**	**60.708**

Table 1: Ablation on cross-attention fusion and triplane decoder.

Details of Gaze-Controllable Facial Representation

The low-frequency encoder $\mathcal{F}_l$ includes a pre-trained DeepLabV3 network and a vision transformer encoder. The vision transformer encoder consists of 2 transformer layers with 4 heads and a hidden size of 1024. This structure is able to capture the global context of the input image and extract low-frequency features (Bai et al. 2025). The high-frequency encoder $\mathcal{F}_h$ is a convolutional neural network with 6 convolutional layers and convolutional neural network has been verified to be effective in capturing high-frequency details (Wang et al. 2020; Bai et al. 2025). The fist convolutional layer has a stride of 2 and other layers have a stride of 1. All the layers have a kernel size of 3x3 with 128 channels.

Additional Qualitative Results

In Fig. 2, we show additional qualitative results comparing our model to the SOTA methods. All the models are evaluated on the personalized test set of the ETH-XGaze dataset.

Figure 3: Additional visualization of generated images under novel gazes and novel views. The upper part is the generated images under novel gazes which are from left to right. The lower part is the generated results under novel views. Our model generates 3D faces with controllable gazes from a single image, producing photorealistic results across diverse head poses and gaze directions while maintaining 3D and gaze consistency.

Our model generates images with superior quality while ST-ED (Zheng et al. 2020) encounters difficulties in preserving identity information and both HeadNeRF (Hong et al. 2022) and GazeNeRF (Ruzzi et al. 2023) struggle to retain facial details.

In Fig. 3, we show additional qualitative results of generation under novel gazes and novel views. Our model can generate 3D faces with controllable gazes from a single input image. It produces photorealistic face images across a wide range of head poses and gaze directions. The results under novel viewpoints demonstrate the model's strong 3D consistency throughout the generation process. Additionally, its capability to produce consistent gaze images is validated by the results under novel gaze directions.

References

Bai, Q.; Shi, Z.; Xu, Y.; Ouyang, H.; Wang, Q.; Yang, C.; Wang, X.; Wetzstein, G.; Shen, Y.; and Chen, Q. 2025. Real-time 3d-aware portrait editing from a single image. In *European Conference on Computer Vision*, 344–362. Springer.

Bulat, A.; and Tzimiropoulos, G. 2017. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, 1021–1030.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; 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.

Deng, Y.; Yang, J.; Xu, S.; Chen, D.; Jia, Y.; and Tong, X. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 0–0.

Hong, Y.; Peng, B.; Xiao, H.; Liu, L.; and Zhang, J. 2022. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20374–20384.

Ruzzi, A.; Shi, X.; Wang, X.; Li, G.; De Mello, S.; Chang, H. J.; Zhang, X.; and Hilliges, O. 2023. GazeNeRF: 3D-Aware Gaze Redirection with Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Trevithick, A.; Chan, M.; Stengel, M.; Chan, E.; Liu, C.; Yu, Z.; Khamis, S.; Ramamoorthi, R.; and Nagano, K. 2023. Real-time radiance fields for single-image portrait view synthesis.

Wang, H.; Wu, X.; Huang, Z.; and Xing, E. P. 2020. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8684–8694.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 325–341.

Zhang, X.; Park, S.; Beeler, T.; Bradley, D.; Tang, S.; and Hilliges, O. 2020. ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation. In *The European Conference on Computer Vision*.

Zheng, Y.; Park, S.; Zhang, X.; De Mello, S.; and Hilliges, O. 2020. Self-Learning Transformations for Improving Gaze and Head Redirection. *Advances in Neural Information Processing Systems*.