GazeGaussian: High-Fidelity Gaze Redirection with 3D Gaussian Splatting

Supplementary Material

7. Overview

The supplementary material encompasses the subsequent components.

- Video for continuous gaze redirection
- Implementation details
- Dataset and pre-processing details
- Supplementary experiments
  - Ablation study on cross-dataset
  - Comparison with the FLAME-based method
  - GazeGaussian vs baseline + expression-guided
- Additional visualization results
  - Visualization for transformed Gaussians
  - Visualization for identity morphing
  - Visualization for ablation study
  - Visualization for cross-dataset comparison
- Ethical considerations and limitations

8. Video for continuous gaze redirection

Please refer to the video **"continuous gaze redirection.mp4"** in the supplementary material for continuous gaze redirection results on the ETH-Xgaze. The side-by-side visualization showcases smooth transitions and high-quality novel gaze synthesis produced by GazeGaussian.

9. Implementation details

We use the Adam optimizer [19], with a learning rate that follows an exponential decay schedule, starting at 1×10^{-4}. We use the VGG-based network pre-trained on ImageNet, as provided by the GazeNeRF [36] implementation, and fine-tune it on the ETH-XGaze training set for the functional loss $\mathcal{L}_{\mathcal{G}}$ as the pre-trained gaze estimator. Additionally, we utilize the ResNet50 backbone from the GazeNeRF [36] framework, trained on the ETH-XGaze training set, to output gaze and head pose for evaluation purposes. All experiments are conducted on an NVIDIA 4090 GPU. We first train an SDF network to extract the neutral mesh and initialize the two-stream Gaussian parameters in 10 epochs. The full pipeline is then trained for an additional 20 epochs until convergence. The loss weights are described in the method section.

10. Dataset and pre-processing details

Following the baseline GazeNeRF [36], all experiments are conducted on four widely used datasets.

ETH-XGaze [58] is a large-scale gaze estimation dataset featuring high-resolution images across a wide range of head poses and gaze directions. Captured with a multi-view camera setup under varying lighting conditions, it includes 756,000 frames from 80 subjects for training. Each frame contains images from 18 different camera perspectives. Additionally, a person-specific test set includes 15 subjects, each with 200 images provided with ground-truth gaze data.

ColumbiaGaze [39] contains 5,880 high-resolution images from 56 subjects. For each subject, images were taken in five distinct head poses, with each pose covering 21 preset gaze directions, allowing for detailed gaze estimation in controlled conditions.

MPIIFaceGaze [55, 56] is tailored for appearance-based gaze prediction. MPIIFaceGaze offers 3,000 face images for each of 15 subjects, paired with two-dimensional gaze labels to facilitate gaze estimation research.

GazeCapture [21] is a large-scale dataset collected through crowd-sourcing, featuring images captured across different poses and lighting conditions. For cross-dataset comparison, we use only the test portion, which includes data from 150 distinct subjects.

Pre-processing. We follow the preprocessing steps in GazeNeRF [36] and Gaussian Head Avatar [51]. The original resolution of ETH-XGaze [58] images is 6K × 4K, while images from other datasets vary in resolution. To standardize, we preprocess all images using the normalization method, aligning the rotation and translation between the camera and face coordinate systems. The normalized distance from the camera to the face center is fixed at 680mm. To extract 3DMM parameters and generate masks for the eyes and face-only regions, we utilize the face parsing model from [63]. GazeGaussian is trained on a single NVIDIA 4090 GPU for 20 epochs on the train set from ETH-XGaze. During inference, GazeGaussian fine-tunes on a single input image, taking approximately 30 seconds for fine-tuning and 0.2 seconds per image for generation.

11. Supplementary experiments

11.1. Ablation study on cross-dataset

To further validate the effectiveness of each proposed component, we conduct an ablation study on the cross-dataset evaluation to assess the generalization capability of our full pipeline. As shown in Tab. 4, the results are consistent with the ablation study in the main text. The proposed Gaussian eye rotation representation significantly improves eye redirection accuracy while ensuring robust redirection across cross-domain datasets. Additionally, the expression-guided neural renderer preserves the identity characteristics of the input image, enabling generalization ability across different subjects. From the ablation study on cross-dataset, we can further validate the importance of each component.

Table 4. Component-wise ablation study of GazeGaussian on the ColumbiaGaze, MPIIFaceGaze and GazeCapture datasets.

Two-stream	Gaus. Eye Rep.	Exp. Guided	ColumbiaGaze Gaze↓	Head↓	LPIPS↓	ID↑	MPIIFaceGaze Gaze↓	Head↓	LPIPS↓	ID↑	GazeCapture Gaze↓	Head↓	LPIPS↓	ID↑
✓			8.996	4.494	0.325	49.286	19.787	8.491	0.321	34.483	15.697	13.740	0.260	33.393
✓		✓	9.143	4.509	0.324	52.805	16.689	8.578	0.303	35.194	15.926	14.869	0.261	36.004
✓	✓		7.799	3.754	0.284	57.252	11.938	6.860	0.257	35.614	10.339	8.208	0.216	40.458
	✓	✓	7.710	3.899	0.280	58.969	12.559	6.188	0.246	37.444	11.296	8.460	0.224	42.294
✓	✓	✓	**7.415**	**3.332**	**0.273**	**59.788**	**10.943**	**5.685**	**0.224**	**41.505**	**9.752**	**7.061**	**0.209**	**44.007**

Table 5. Comparison between baselines + expression-guided neural renderer and GazeGaussian on ETH-xgaze

Method	Gaze↓	Head Pose↓	SSIM↑	PSNR↑	LPIPS↓	FID↓	Identity Similarity↑	FPS↑
GazeNeRF	6.944	3.470	0.733	15.453	0.291	81.816	45.207	46
GazeNeRF + EGNR	6.854	3.025	0.764	16.147	0.258	67.219	50.268	44
GHA	30.963	8.498	0.638	12.108	0.359	74.560	27.272	**91**
GHA + EGNR	28.374	6.533	0.714	14.213	0.305	69.101	41.332	90
GazeGaussian (Ours)	**6.622**	**2.128**	**0.823**	**18.734**	**0.216**	**41.972**	**67.749**	74

11.2. Comparison with the FLAME-based method

The FLAME-based baseline by Wang et al. [42] is not open source and lacks metrics in gaze redirection in its published materials. Nonetheless, we have cited the reported results on the ETH-Xgaze dataset in Wang's paper and provided a comparison in Tab 6. The results demonstrate that our GazeGaussian still achieves better synthesis quality, especially for perceptual metrics.

Table 6. Image quality comparison with the FLAME-based method.

Methods	SSIM ↑	PSNR ↑	LPIPS ↓
Wang et al.	0.732	19.144	0.265
Ours	**0.823**	18.734	**0.216**

11.3. GazeGaussian vs baseline + expression-guided

We make a comparison between GazeGaussian and baselines (GazeNeRF, Gaussian Head Avatar) enhanced with the expression-guided neural renderer (EGNR) on the ETH-XGaze dataset. As shown in Tab. 5, integrating EGNR into GazeNeRF and Gaussian Head Avatar (GHA) leads to noticeable improvements in gaze redirection accuracy and image quality. This demonstrates the versatility of the proposed expression-guided neural renderer in enhancing facial synthesis and better capturing identity-specific expressions. Although GHA is restricted to animating single head avatar, it can benefit from enhanced generalization through expression-guided neural renderer. However, even with the added EGNR, the performance of GazeNeRF and GHA remains limited compared to GazeGaussian. The fundamental constraint lies in GazeNeRF's representation, which lacks the explicit modeling of gaze and facial expression dynamics offered by GazeGaussian's two-stream Gaussian structure. GHA restricts to full head animation, missing two-stream modeling for face and eye disentanglement, leading to decreased performance.

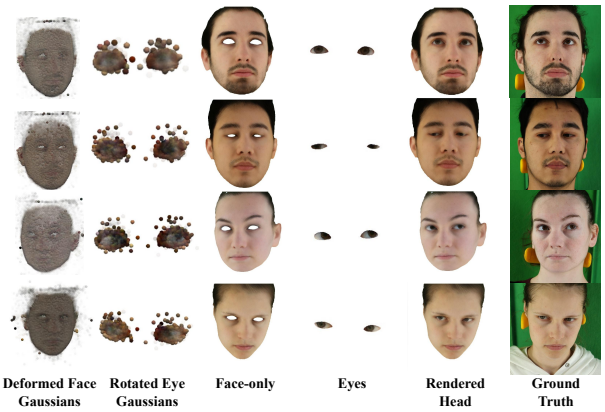

Figure 7. Visualization of transformed two-stream Gaussians after deformation from the canonical space.

12. Supplementary visualization

12.1. Visualization for transformed Gaussians

To demonstrate the advantages of GazeGaussian's explicit control of head pose and gaze direction for head and eye regions, we visualize the Gaussians after deformation from the canonical space. As shown in Fig. 7, the explicit support for rotation and translation in GazeGaussian allows the deformed Gaussians to form a reasonable spatial distribution and accurate color representation. This capability enables precise geometric control and high-fidelity image rendering. In contrast, GazeNeRF performs rotations only on the feature map level, failing to fully deform in 3D space, which limits its performance compared to our method.

12.2. Visualization for identity morphing

Fig. 8 showcases identity morphing results on the ETH-XGaze dataset. We randomly select two subjects with identical gaze directions and head poses. By interpolating their latent codes, we generate a smooth transition between the

Figure 8. Face morphing results on the ETH-XGaze dataset.

two identities while keeping the gaze direction and head pose consistent. This visualization demonstrates the capability of GazeGaussian to preserve gaze alignment and head orientation during synthesis, even as the facial features gradually change according to the interpolated latent codes.

12.3. Visualization for ablation study

Fig. 10 presents additional qualitative results from our ablation study conducted on the ETH-XGaze dataset. These visualizations highlight the importance of each proposed component in GazeGaussian.

Without the Gaussian eye rotation representation, the model struggles to achieve accurate eye control, resulting in noticeable deviations in gaze direction and reduced realism in the eye region. This demonstrates the critical role of the Gaussian eye rotation representation in enabling precise and realistic gaze redirection. Additionally, the absence of the expression-guided neural renderer leads to a significant loss in facial detail and expression fidelity. With the renderer included, the synthesized images exhibit finer facial details and improved consistency with the target identity, showcasing the renderer's effectiveness in enhancing the overall quality of face synthesis. These results confirm that both components contribute significantly to the superior performance and visual fidelity of GazeGaussian.

12.4. Visualization for cross-dataset comparison

We provide additional cross-dataset comparison visualizations for MPIIFaceGaze (Fig. 11), ColumbiaGaze (Fig. 12) and GazeCapture (Fig. 13). Compared to the baseline, GazeGaussian achieves high-fidelity gaze redirection with superior image synthesis quality.

13. Ethical considerations and limitations

Our approach allows for the creation of lifelike portrait videos that may be exploited to spread misinformation, sway public opinion, and erode trust in media, with grave societal impacts. Thus, developing trustworthy techniques to discern real from fake content is crucial. We firmly oppose any unauthorized or harmful use of this technology and highlight the need to address ethical issues in its implementation.

Figure 9. Example of a failure case.

While GazeGaussian represents a significant advancement in gaze redirection quality, there is still one unresolved issue. Due to limitations in facial tracking models such as FLAME, it remains challenging to accurately model accessories such as glasses, earrings, and even hair details as shown in Fig. 9. An existing method [26] has attempted to use cylindrical Gaussian representations to model hair. To further enhance GazeGaussian, improving the 3DGS facial representation will be a key focus of our future work.

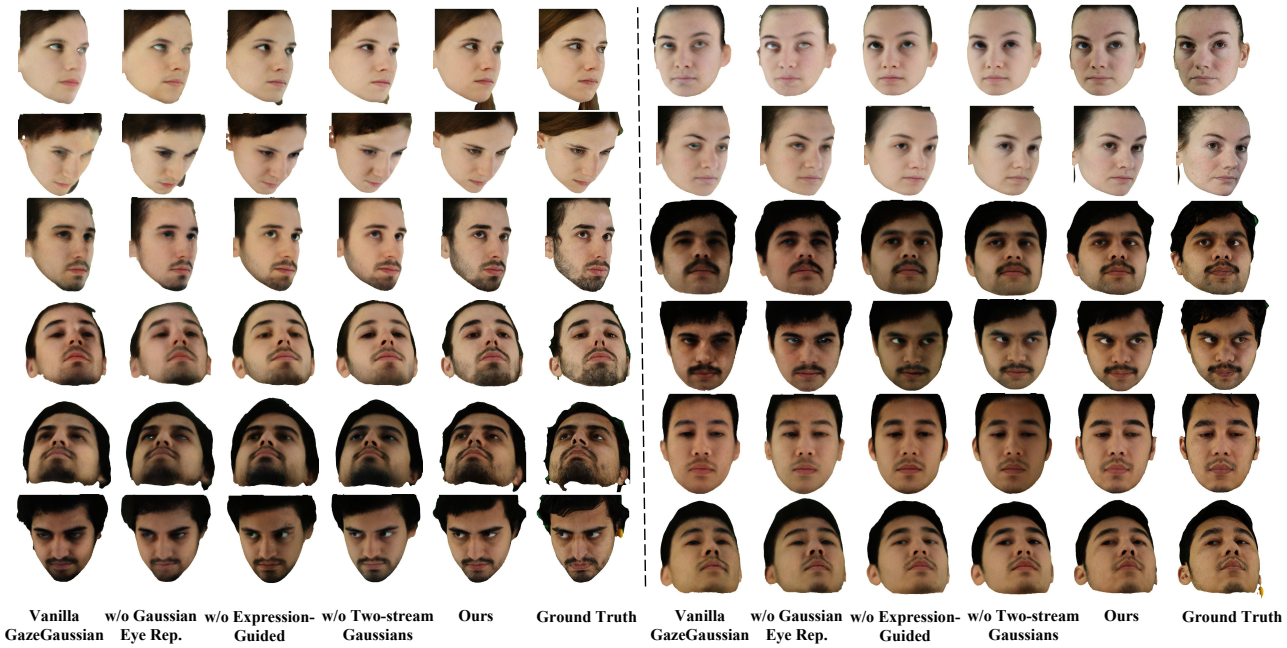

Figure 10. Additional qualitative ablation study on the ETH-XGaze dataset.

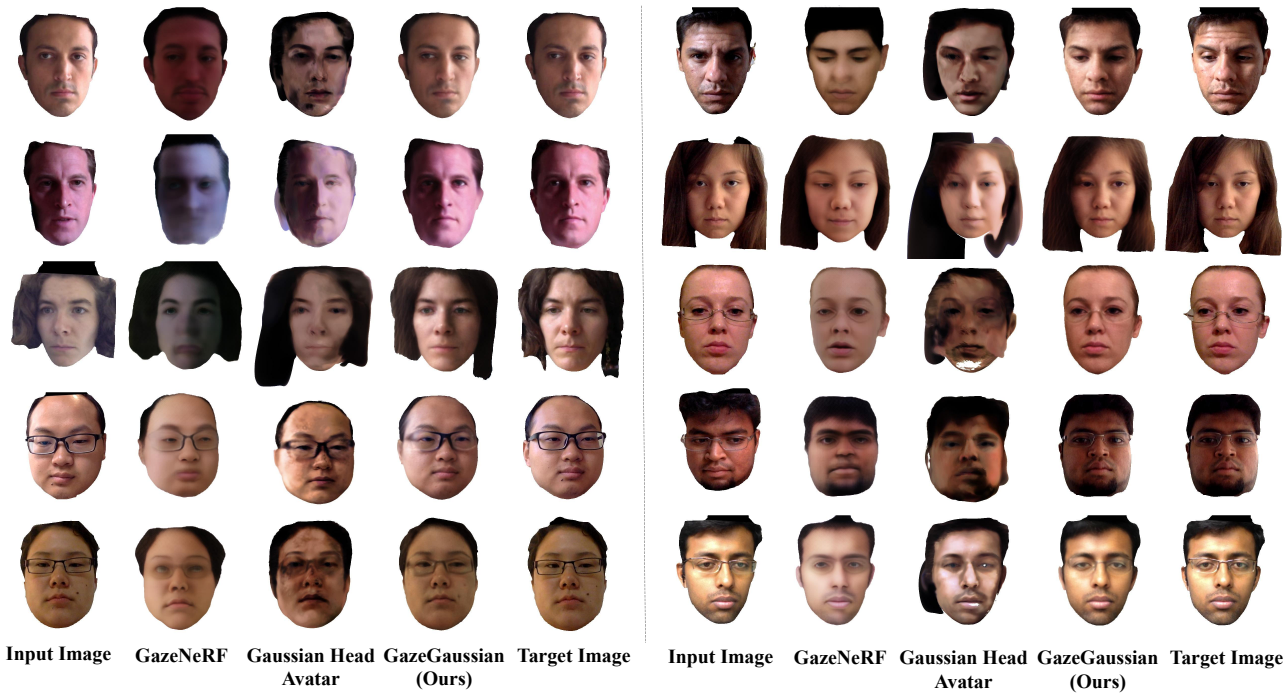

Figure 11. Cross-dataset visualization on MPIIFaceGaze.

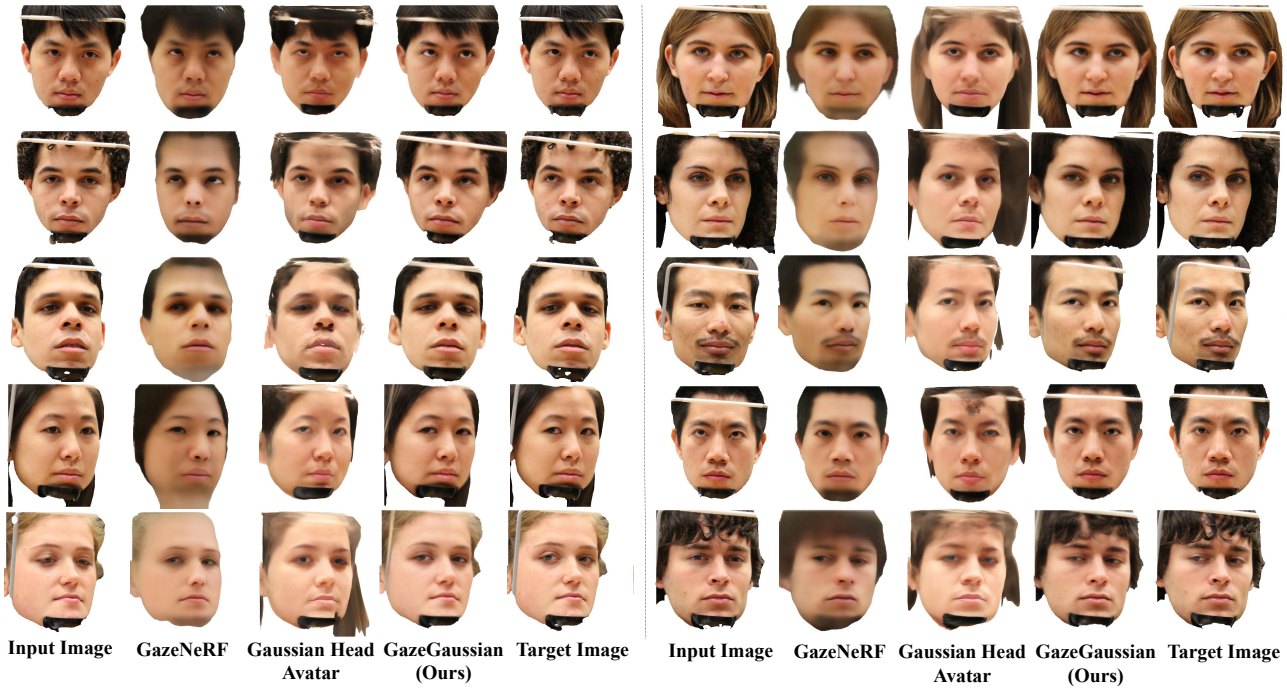

Figure 12. Cross-dataset visualization on ColumbiaGaze.

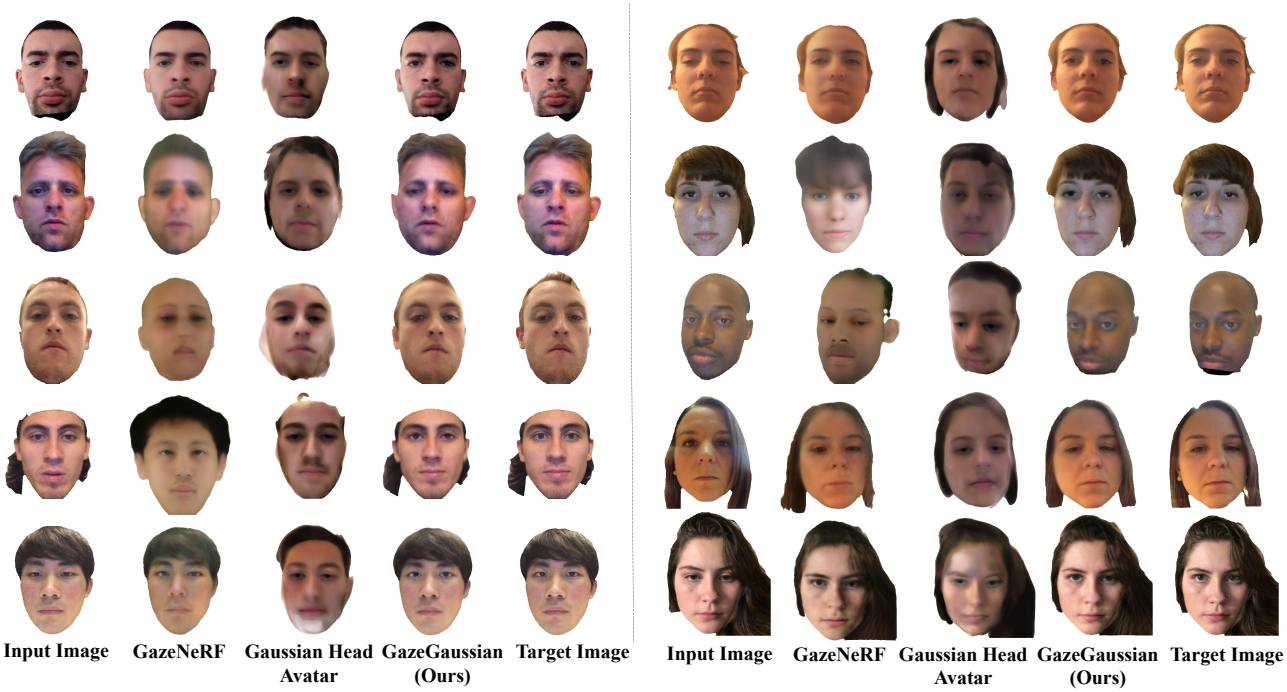

Figure 13. Cross-dataset visualization on GazeCapture.