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

Xiaobao Wei[1,2]   Peng Chen[1,2]   Guangyu Li[1]   Ming Lu[3]   Hui Chen[1,2,†]   Feng Tian[1,2]

[1]Institute of Software, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences   [3]Intel Labs China
weixiaobao0210@gmail.com

Abstract

Gaze estimation encounters generalization challenges when dealing with out-of-distribution data. To address this problem, recent methods use neural radiance fields (NeRF) to generate augmented data. However, existing methods based on NeRF are computationally expensive and lack facial details. 3D Gaussian Splatting (3DGS) has become the prevailing representation of neural fields. While 3DGS has been extensively examined in head avatars, it faces challenges with accurate gaze control and generalization across different subjects. In this work, we propose GazeGaussian, the first high-fidelity gaze redirection method that uses a two-stream 3DGS model to represent the face and eye regions separately. Leveraging the unstructured nature of 3DGS, we develop a novel representation of the eye for rigid eye rotation based on the target gaze direction. To enable synthesis generalization across various subjects, we integrate an expression-guided module to inject subject-specific information into the neural renderer. Comprehensive experiments show that GazeGaussian outperforms existing methods in rendering speed, gaze redirection accuracy, and facial synthesis across multiple datasets. The code is available at: https://ucwxb.github.io/GazeGaussian.

1. Introduction

Gaze estimation is a fundamental component across various applications [1, 27, 29], yet current estimators [3, 4, 50] often struggle to generalize effectively to out-of-distribution data. To address this, recent approaches [36, 42, 52] have started exploring gaze redirection, which manipulates the gaze in an input image toward a target direction. This process generates augmented data to enhance the generalization capabilities of gaze estimators.

Earlier methods [10, 53, 54, 57] formulate gaze redirection as a 2D image manipulation task, relying on deep learning techniques to warp eye regions of the image toward
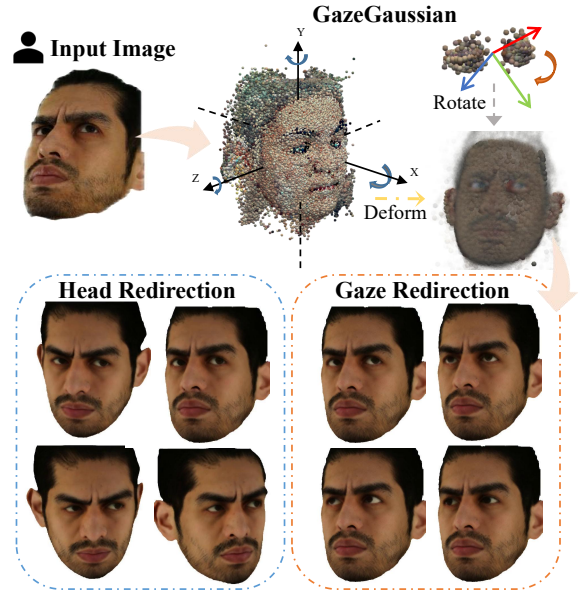

Figure 1. Gaze redirection: Given an input image, GazeGaussian deforms face and eye Gaussians from canonical space to generate high-fidelity head images with accurate gaze redirection.

the target gaze direction. However, these 2D approaches overlook the inherently 3D nature of head and gaze manipulation, often resulting in poor spatial consistency and limited synthesis fidelity. With advancements in Neural Radiance Fields (NeRF) [28] and its variants [44, 46], several methods [12, 16, 60, 62] have achieved 3D dynamic head representation and high-fidelity avatar synthesis. Meanwhile, to enable precise control of gaze direction, recent research [36, 42, 52] has introduced approaches that decouple the face and eye regions, modeling each with separate neural fields to achieve accurate gaze redirection.

As NeRF-based methods are hindered by high computational demands, 3D Gaussian Splatting [18] and its variants [17, 25, 45] achieve impressive rendering quality with significantly faster training speeds. Recent research [33, 49, 51] has applied these methods to 3D head

---

†Corresponding author.

animation, typically using face-tracking [41, 61] parameters to model dynamic 3D head representations. However, existing 3DGS-based approaches neglect the accurate control of gaze direction and struggle to generalize across different subjects, limiting their effectiveness for gaze redirection.

To address the above issues, we propose GazeGaussian, a high-fidelity gaze redirection method that leverages a two-stream 3D Gaussian Splatting (3DGS) model to represent the face and eye regions, respectively. To our knowledge, this is the first integration of 3DGS into gaze redirection tasks. An overview is shown in Fig. 1.

GazeGaussian begins by initializing the two-stream 3DGS model using a pre-trained neutral mesh on the training dataset. This mesh is divided into distinct regions for the face and eyes. By employing gaze direction and face tracking codes, we optimize a deformation field for the face and a rotation field for the eyes, allowing us to adjust the neutral Gaussians accordingly. To achieve precise eye rotation aligned with the target gaze, we present a novel Gaussian Eye Rotation Representation. In contrast to methods like GazeNeRF that implicitly alter feature maps, GazeGaussian explicitly adjusts the position of Gaussians in the eye branch according to the desired gaze direction, utilizing the controllable nature of 3DGS. To address possible errors in gaze direction, GazeGaussian develops an eye rotation field to enhance redirection accuracy. The two-stream Gaussians are rasterized into high-level features and sent to the neural renderer. Finally, to enable the generalization ability across different subjects and inject subject-specific information, we employ an expression-guided neural renderer (EGNR) to synthesize the final gaze-redirection images.

Our main contributions are summarized as follows:

- We introduce GazeGaussian, the first 3DGS-based gaze redirection pipeline, achieving precise gaze manipulation and generalizable head avatar synthesis.
- To enable rigid and accurate eye rotation based on the target gaze direction, we propose a novel two-stream 3DGS framework to decouple face and eye deformations, featuring a specialized Gaussian eye rotation for explicit control over eye movement.
- To equip 3DGS with generalization ability, we design an expression-guided neural renderer (EGNR) to integrate subject-specific characteristics.
- We conduct comprehensive experiments on ETH-XGaze, ColumbiaGaze, MPIIFaceGaze, and GazeCapture datasets, where GazeGaussian achieves state-of-the-art gaze redirection accuracy and facial synthesis quality with competitive rendering speed.

2. Related Work

Gaze Redirection. Gaze redirection is the task of manipulating the gaze direction of a face image to a target direction while preserving the subject's identity and other facial details.

Earlier approaches for gaze redirection include novel view synthesis [5, 11, 22], eye-replacement [34, 38], and warping-based methods [10, 20, 47]. However, these methods are limited by person-specific data, restricted redirection range, and artifact introduction. To further improve gaze redirection, recent studies [14, 30, 48, 59] have employed neural network-based generative models. STED [59], building on FAZE [30], introduces a self-transforming encoder-decoder that generates full-face gaze redirection images. With advancements in Neural Radiance Fields (NeRF) [28], several studies [23, 36, 42, 52] have aimed to model the complex rotation of the eyeball. GazeNeRF [36] employs a two-stream MLP architecture to separately model the face and eye regions, achieving improved gaze redirection performance.

However, these methods are hindered by substantial computational demands and limited rendering efficiency. Gaze manipulation occurs at the feature maps and remains an implicit approach. In contrast, GazeGaussian allows for explicit control over eye rotations, improving gaze redirection accuracy and accelerating the synthesis process.

Head Avatar Synthesis. The synthesis of head avatars has garnered considerable attention in recent years. FLAME [24] is a parameterized 3D head model that maps parameters of shape, expression, and pose onto 3D facial geometry, allowing for controllable head avatar generation. Many subsequent works [2, 6, 8, 31, 32, 35] focus on using the FLAME model for speech-driven head avatar animation. Recent head animation techniques can be categorized into two main approaches: NeRF-based methods and 3DGS-based methods. NeRF-based approaches [9, 16, 60, 62] leverage neural radiance fields to deform facial movements from a canonical space. HeadNeRF [16] introduces a parametric head model that controls facial shape, expression, and albedo under different lighting conditions. With the emergence of 3D Gaussian Splatting (3DGS) [18], several approaches [7, 33, 49, 51] have explored its application in single head avatar modeling. Gaussian Head Avatar [51] initializes Gaussians with a neutral mesh head and incorporates MLPs to deform complex facial expressions.

However, these methods ignore precise gaze control and are confined to the single avatar animation. In contrast, GazeGaussian achieves precise gaze direction control by decoupling facial animations and gaze movement within a two-stream model. Furthermore, an expression-guided neural renderer is designed to inject person-specific information for generalization ability across different subjects.

3. Overview

The pipeline of GazeGaussian is illustrated in Fig. 2, including the two-stream Gaussians and the proposed expression-guided neural renderer. Before the beginning of the pipeline, we follow the data preprocessing in GazeNeRF [36] and Gaussian Head Avatar [51], which include background re-
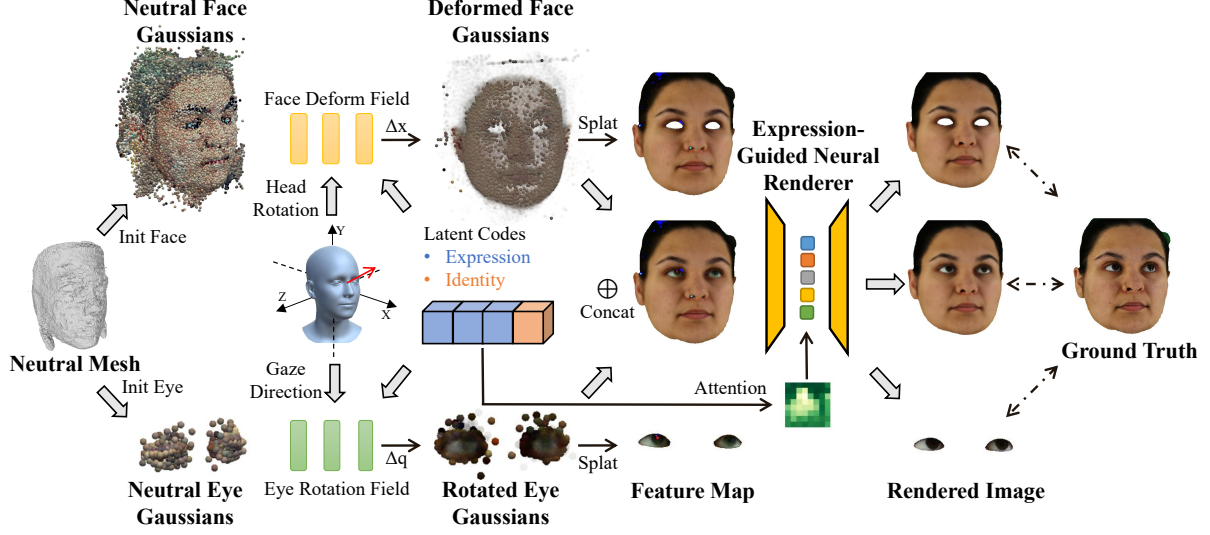
Figure 2. Pipeline of GazeGaussian. We initialize face-only and eye regions from a pre-trained neutral mesh. Using target expression codes, head rotation, and gaze direction, GazeGaussian optimizes face deformation and eye rotation fields to transform the neutral Gaussians. The transformed Gaussians are splatted into feature maps. The expression codes guide the neural renderer through cross-attention, enabling the rendering of feature maps into high-fidelity gaze redirection images, which are then supervised by multi-view ground truths.

moval, gaze direction normalization, and facial tracking for each frame. To obtain a neutral mesh for Gaussian initialization, we first reconstruct a Sign Distance Function (SDF) based neutral geometry and then optimize a face deformation field and an eye rotation field from the training data. A neutral mesh representing a coarse geometry across different subjects can be extracted using DMTet [37]. We then partition the neutral mesh into face-only and eye regions using 3D landmarks, initializing the two-stream Gaussians. Based on these neutral Gaussians, GazeGaussian optimizes a face deformation field and an eye rotation field to transform the Gaussians according to the target expression codes, gaze direction, and head rotation. Next, we concatenate the two-stream Gaussians and rasterize them into a high-dimensional feature map representing the head, face-only, and eye regions. Finally, these feature maps are fed into the expression-guided neural renderer to generate high-fidelity gaze redirection images. The ground truth image is used to supervise the rendered face-only, head, and eye images.

4. Method

4.1. Preliminaries

The vanilla 3D Gaussians [18] with n points are represented by their center $\mu \in \mathbb{R}^3$, the color computed by k degrees of spherical harmonics $c \in \mathbb{R}^{3(k+1)^2}$, the quaternion rotation $R \in \mathbb{R}^4$, scale $S \in \mathbb{R}^3$ and opacity $\alpha \in \mathbb{R}$. The covariance matrix Σ can be obtained by $\Sigma = RSS^T R^T$, which then formulates the Gaussian function $G(x) = e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$. These Gaussians are then rasterized into an image I, in which the pixel color C is computed by blending N ordered

Gaussians overlapping the pixel:

$$C = \sum_{i=1}^{N} c_i \alpha_i \prod_{j=1}^{i-1}(1 - \alpha_j), \quad (1)$$

4.2. Two-stream GazeGaussian Representation

Our task is to synthesize a head avatar conditioned on gaze direction, head rotation, and expression latent codes. To decouple the complex movements in the face and eyes, we introduce a two-stream Gaussian model consisting of a face-only deformation branch and an eye rotation branch in the following section.

4.2.1. Face Deformation

For the face-only branch, inspired by Gaussian Head Avatar [51], we first construct canonical neutral face Gaussians with trainable attributes: $\{\mu_0^f, z_0^f, R_0^f, S_0^f, \alpha_0^f\}$. All attributes are identical to a standard Gaussian, except for $z_0^f \in \mathbb{R}^{128}$, which signifies the point-wise feature vectors. Rather than directly modeling complex color c^f, we decode z_0^f into c^f and affect other attributes' deformation to capture intricate facial details.

To transform canonical face Gaussians into the target space, we first model the influence of latent codes τ and head pose γ through respective weights λ_τ and λ_γ. We compute the minimum distance d from each Gaussian center μ to the 3D facial landmarks (excluding eye regions). Based on the distance, we assign influence weights λ_τ and λ_γ to the expressive latent codes τ and head pose γ. For Gaussians near landmarks ($d < d_1$), they are primarily influenced by

expressive latent codes and set $\lambda_\tau = 1$. For distant Gaussians $(d > d_2)$, they are predominantly affected by head pose, with $\lambda_\tau = 0$. To ensure a smooth transition between these regions $(d_2 \geq d \geq d_1)$, we define $\lambda_\tau = (d_2 - d)/(d_2 - d_1)$. Consequently, the pose weight is determined as $\lambda_\gamma = 1 - \lambda_\tau$. The distance thresholds are empirically set to $d_1 = 0.15$ and $d_2 = 0.25$.

After that, we develop MLPs $\{E_\mu^f, E_c^f, E_\kappa^f\} \in E^f$ for latent codes-conditioned expression and $\{P_\mu^f, P_c^f, P_\kappa^f\} \in P^f$ for head pose-conditioned deformation:

$$\mu^f = \mu_0^f + \lambda_\tau E_\mu^f(\mu_0^f, \tau) + \lambda_\gamma P_\mu^f(\mu_0^f, \gamma),$$
$$c^f = \lambda_\tau E_c^f(z_0^f, \tau) + \lambda_\gamma P_c^f(z_0^f, \gamma), \qquad (2)$$
$$\kappa^f = \kappa_0^f + \lambda_\tau E_\kappa^f(z_0^f, \tau) + \lambda_\gamma P_\kappa^f(z_0^f, \gamma),$$

where $\kappa^f = \{R^f, S^f, \alpha^f\}$ and $\kappa_0^f = \{R_0^f, S_0^f, \alpha_0^f\}$. Finally, we apply rigid rotations and translations to transform Gaussians in the canonical space to the world space. Then, these Gaussians are rasterized into the feature maps:

$$\mathcal{M}_f = \mathcal{R}(\{\mu^f, c^f, R^f, S^f, \alpha^f\}), \qquad (3)$$

where $\mathcal{R}$ represents the rasterizer and $\mathcal{M}_f$ indicates the feature map from the face-only branch.

4.2.2. Eye Rotation

Different from Gaussian Head Avatar [51], we design an eye branch for accurate eye modeling and gaze redirection. For the eye branch, we construct canonical neutral eye Gaussians with attributes $\{\mu_0^e, z_0^e, R_0^e, S_0^e, \alpha_0^e\}$. These attributes share the same dimension as those in the face-only branch, except that $S_0^e \in \mathbb{R}^{N \times 1}$ is constrained to be spherical, aligning with the rotational properties of the eyeball. Since eyes are relatively small and mainly influenced by the gaze direction, λ used in the face is omitted here.

Directly applying the same face deformation strategy fails to fully leverage the unique characteristics of eyeball rotational motion, resulting in inaccurate gaze redirection. Therefore, we first rotate the eye Gaussians in the canonical space and then incorporate the expressive latent codes τ to generate deformation offsets. Since the gaze labels contain noise, directly using the normalized gaze direction φ to rotate the Gaussians would lead to numerical optimization errors. To address this, we optimize two separate MLPs: $E_\mu^e \in E^e$ and $G_\mu^e \in G^e$ to predict the biases for Gaussian eye rotation.

$$\mu^e = E_\mu^e(\mu_0^e, \tau) + G_\mu^e(\mu_0^e, \varphi)\mu_0^e, \qquad (4)$$

Other attributes are deformed in a similar way with $\{E_c^e, E_a^e\} \in E^e$ and $\{G_c^e, G_a^e\} \in G^e$:

$$c^e = E_c^e(z_0^e, \tau) + G_c^e(z_0^e, \varphi),$$
$$\kappa^e = \kappa_0^e + E_\kappa^e(z_0^e, \tau) + G_\kappa^e(z_0^e, \varphi), \qquad (5)$$

where $\kappa^e = \{R^e, S^e, \alpha^e\}$ and $\kappa_0^e = \{R_0^e, S_0^e, \alpha_0^e\}$. By the proposed explicit eye rotation, GazeGaussian succeeds in

face and eye disentanglement, which is ignored in the Gaussian Head Avatar [51]. Finally, eye Gaussians are transformed into the world space. Then these eye Gaussians are rasterized into the feature maps:

$$\mathcal{M}_e = \mathcal{R}(\{\mu^e, c^e, R^e, S^e, \alpha^e\}), \qquad (6)$$

To obtain the full head rendering, we simply concat the two-stream Gaussians $\{\mu^h, c^h, R^h, S^h, \alpha^h\} = \text{Concat}(\{\mu^f, c^f, R^f, S^f, \alpha^f\}, \{\mu^e, c^e, R^e, S^e, \alpha^e\})$ and rasterized them into head feature maps:

$$\mathcal{M}_h = \mathcal{R}(\{\mu^h, c^h, R^h, S^h, \alpha^h\}), \qquad (7)$$

4.3. Expression-Guided Neural Renderer

To filter noise caused by uneven lighting or camera chromatic aberration [51], a UNet-like neural renderer $\mathcal{U}$ opts to rasterize feature maps into the final face-only, eyes, and head images $\{\mathcal{I}_f, \mathcal{I}_e, \mathcal{I}_h\}$:

$$\{\mathcal{I}_f, \mathcal{I}_e, \mathcal{I}_h\} = \mathcal{U}(\{\mathcal{M}_f, \mathcal{M}_e, \mathcal{M}_h\}), \qquad (8)$$

However, the vanilla renderer used in Gaussian Head Avatar [51] and GazeNeRF [36] focuses solely on image post-processing, resulting in a loss of subject-specific facial details. To further improve generalization ability by providing subject-specific information, we inject the expressive latent codes τ into the neural renderer through a slice cross-attention module. Let z_b represent the bottleneck feature obtained from the encoder of $\mathcal{U}$. We utilize the latent codes to query this bottleneck feature, using it as a conditional signal to guide the renderer's synthesis process. The guiding process can be formulated as:

$$z_b' = z_b + z_b \cdot \text{Attn}(q = \tau, k = z_b, v = z_b)), \qquad (9)$$

where $\text{Attn}(\cdot)$ denotes the cross-attention operation that fuses the latent codes with the bottleneck feature. Then the refined feature z_b' is decoded as final images.

4.4. Training

GazeGaussian Initialization. Initialization for the 3D Gaussians (3DGS) is crucial for stable optimization. Following Gaussian Head Avatar [51], we initialize the two-stream Gaussians using the neutral mesh extracted from an SDF field. This neutral mesh provides a coarse geometry and texture, which are used to initialize the positions and features of the Gaussians. To decouple the face-only and eye regions, we compute the 3D neutral landmarks and use learnable parameters to define the vertices near the eyes as the initial Gaussians for the eye region, while the rest of the head is used to initialize the face-only Gaussians. Additionally, we transfer the parameters of all deformation and color MLPs while the MLPs for attribute prediction and the expression-guided neural renderer are randomly initialized.

Image Synthesis Loss. The masked ground truth image $\mathcal{I}_{gt}$ is used to supervise the rendered images $\mathcal{I}_f, \mathcal{I}_e, \mathcal{I}_h$, corresponding to the face-only, eyes, and head regions, respectively. Additionally, we enforce the first three channels of the feature maps $\mathcal{M}_f, \mathcal{M}_e, \mathcal{M}_h$ to learn RGB colors. For rendered images and the first three channels of feature maps, we apply the same loss functions. For brevity, we take the loss for the rendered eye image $\mathcal{I}_e$ as an example. We mask the ground truth image using an eye mask and then apply L1 loss, SSIM loss, and LPIPS loss on the masked image:

$$\mathcal{L}_{\mathcal{I}}^e = ||\mathcal{I}_{gt} - \mathcal{I}_e||_1 + \lambda_{SSIM}(1 - SSIM(\mathcal{I}_{gt}, \mathcal{I}_e)) \\ + \lambda_{VGG} VGG(\mathcal{I}_{gt}, \mathcal{I}_e), \quad (10)$$

where $\lambda_{SSIM} = \lambda_{VGG} = 0.1$ is the weight of loss. The image synthesis loss is the sum of the three rendered images and three feature maps:

$$\mathcal{L}_{\mathcal{I}} = \mathcal{L}_{\mathcal{I}}^f + \mathcal{L}_{\mathcal{I}}^e + \mathcal{L}_{\mathcal{I}}^h + \mathcal{L}_{\mathcal{M}}^f + \mathcal{L}_{\mathcal{M}}^e + \mathcal{L}_{\mathcal{M}}^h, \quad (11)$$

where $\mathcal{L}_{\mathcal{I}}^f, \mathcal{L}_{\mathcal{I}}^e, \mathcal{L}_{\mathcal{I}}^h$ represent the losses for the rendered face-only, eye, and head images, respectively. $\mathcal{L}_{\mathcal{M}}^f, \mathcal{L}_{\mathcal{M}}^e, \mathcal{L}_{\mathcal{M}}^h$ represent the losses for the feature maps corresponding to the face-only, eye, and head regions, respectively. The image synthesis loss ensures the full disentanglement of the eye and the rest of the face.

Gaze Redirection Loss. To improve task-specific performance and eliminate task-relevant inconsistencies between the target image $\mathcal{I}_{gt}$ and the reconstructed head image $\mathcal{I}_h$, we adopt the functional loss used in STED [59] and GazeNeRF [36]. The gaze redirection loss can be formulated as:

$$\mathcal{L}_{\mathcal{G}}(\mathcal{I}_h, \mathcal{I}_{gt}) = \mathcal{E}_{\text{ang}}(\psi^g(\mathcal{I}_h), \psi^g(\mathcal{I}_{gt})), \\ \mathcal{E}_{\text{ang}}(\varphi, \hat{\varphi}) = \arccos \frac{\varphi \cdot \hat{\varphi}}{\|\varphi\| \|\hat{\varphi}\|}, \quad (12)$$

where $\psi^g(\cdot)$ represents the gaze direction estimated by a pre-trained gaze estimator network, and $\mathcal{E}_{\text{ang}}(\cdot, \cdot)$ represents the angular error function. Our final loss function is:

$$\mathcal{L} = \lambda_{\mathcal{I}} \mathcal{L}_{\mathcal{I}} + \lambda_{\mathcal{G}} \mathcal{L}_{\mathcal{G}}, \quad (13)$$

where $\lambda_{\mathcal{I}} = 1.0$ and $\lambda_{\mathcal{G}} = 0.1$. GazeGaussian is trained with the final loss until convergence.

5. Experiments

To demonstrate the effectiveness of GazeGaussian, we first conduct a within-dataset comparison on the ETH-XGaze dataset [58], testing GazeGaussian alongside state-of-the-art gaze redirection and head generation methods. Next, we perform a cross-dataset comparison on ColumbiaGaze [39], MPIIFaceGaze [55, 56], and GazeCapture [21] to assess

generalization. We also conduct an ablation study to analyze the contributions of each component in GazeGaussian. All experimental settings strictly follow GazeNeRF [36], including dataset preprocessing, the ETH-XGaze split (14.4K images for efficient training), image pairs for evaluation, and the same pre-trained gaze estimator. Additionally, we validate the impact of synthesized data on gaze estimator performance in Sec. 5.5. **Due to space limitations, please refer to Sec. 11 and Sec. 12 in the supplementary for more details on the experiment and visualization results.**

5.1. Experimental Settings

Dataset Pre-processing. Following preprocessing in GazeNeRF [36], we normalize raw images [40, 57] and resize them into a resolution 512×512. To enable separate rendering of the face and eyes regions, we generate masks using face parsing models [63]. We also use the 3D face tracking method from [51] to produce identity and expression codes and camera poses for the input of our method. For normalization, gaze labels provided by datasets are converted to pitch-yaw angles in the head coordinate system across all datasets. More details are provided in the supplementary materials.

Baselines. We compare our method with the self-supervised gaze redirection approach STED [59], along with NeRF-based models such as HeadNeRF [15] and the state-of-the-art method GazeNeRF [36], as well as the latest 3DGS-based head synthesis method, Gaussian Head Avatar [51]. As the NeRF-based methods, NeRF-Gaze [52] and Wang *et al.* [42] are not yet open-sourced, they are not available for inclusion in our comparisons.

Metrics. We evaluate all models using four categories: redirection accuracy, image quality, identity preservation (ID), and rendering speed. Redirection accuracy is measured by gaze and head poses angular errors, using the same ResNet50 [13]-based estimator in GazeNeRF [36]. Image quality is assessed with SSIM, PSNR, LPIPS, and FID. Identity preservation (ID) is evaluated with FaceX-Zoo [43], comparing identity consistency between redirected and ground-truth images. Rendering speed is reported as average FPS.

5.2. Within-dataset Comparison

Following the experimental setup of GazeNeRF, we perform a within-dataset evaluation to compare the performance of GazeGaussian with other state-of-the-art methods. All models are trained using 14.4K images derived from 10 frames per subject, with 18 camera view images per frame, covering 80 subjects in the ETH-XGaze training set. For a fair comparison, the training set is strictly the same as GazeNeRF. Given that Gaussian Head Avatar is not specifically designed for gaze redirection and does not support gaze as a condition, only expressive latent codes are provided for its training and evaluation. The evaluation is conducted on the person-specific test set of the ETH-XGaze dataset. This test

Table 1. Within-dataset comparison: Quantitative results of the GazeGaussian to other SOTA methods on the ETH-XGaze dataset in terms of gaze and head redirection errors in degree, rendered image quality (SSIM, PSNR, LPIPS, FID), identity similarity and rendering FPS.

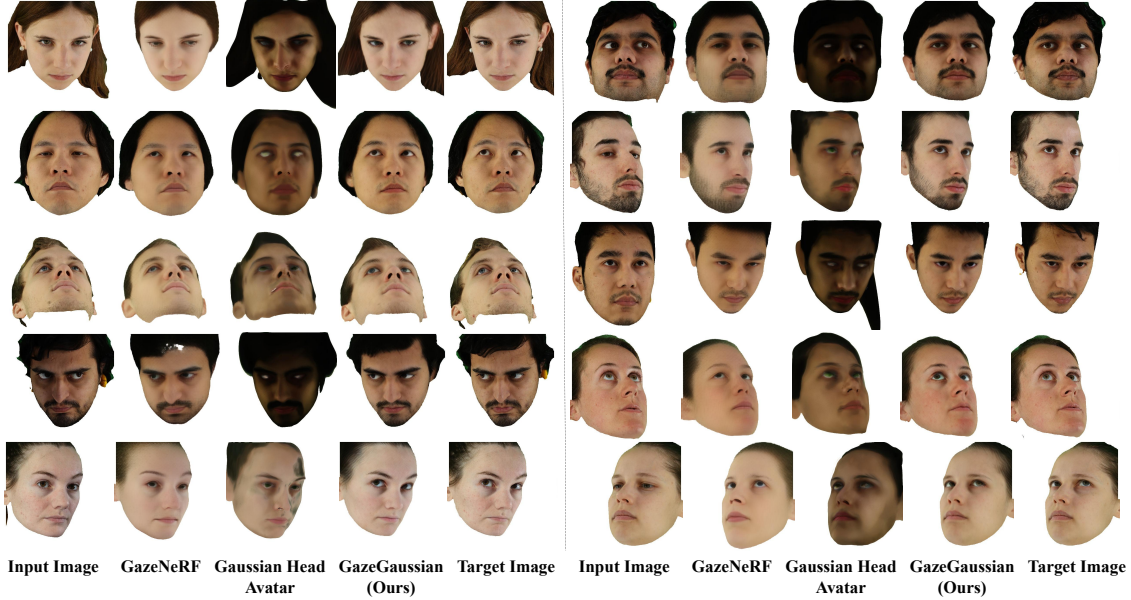| Method	Gaze↓	Head Pose↓	SSIM↑	PSNR↑	LPIPS↓	FID↓	Identity Similarity↑	FPS↑
STED	16.217	13.153	0.726	17.530	0.300	115.020	24.347	18
HeadNeRF	12.117	4.275	0.720	15.298	0.294	69.487	46.126	35
GazeNeRF	6.944	3.470	0.733	15.453	0.291	81.816	45.207	46
Gaussian Head Avatar	30.963	13.563	0.638	12.108	0.359	74.560	27.272	**91**
GazeGaussian (Ours)	**6.622**	**2.128**	**0.823**	**18.734**	**0.216**	**41.972**	**67.749**	74

Figure 3. Within-dataset visualization: head images are generated from the ETH-XGaze test set using our GazeGaussian, GazeNeRF, and Gaussian Head Avatar. GazeGaussian generates photo-realistic images with the target gaze direction, preserving identity and facial details. In contrast, GazeNeRF loses identity information and facial details, while Gaussian Head Avatar fails to manipulate the gaze direction.

set consists of 15 subjects, each with 200 images annotated with gaze and head pose labels. We follow the pairing setting in GazeNeRF, which pairs these 200 labeled images per subject as input and target samples, and the same pairings are used across all models to ensure fairness.

Tab. 1 presents the quantitative results of GazeGaussian alongside baseline methods. It can be observed that Gaze-Gaussian consistently outperforms prior methods across all metrics. Compared to the previous SOTA method GazeN-eRF, which only applies rotation to feature maps for implicit gaze redirection, GazeGaussian adopts a Gaussian eye rotation to explicitly control eye movement. Such a technique improves redirection precision, resulting in high-quality head image synthesis. Additionally, GazeGaussian achieves a rendering speed of 74 FPS, nearly doubling the performance of GazeNeRF, underscoring its efficiency. In contrast, Gaussian Head Avatar (GHA), the latest 3DGS-based model, fails to achieve high performance in gaze redirection. The lack of dedicated mechanisms for gaze disentanglement and explicit

eye region modeling in GHA leads to poor gaze redirection performance. By decoupling the face and eye representation with two-stream Gaussians, GazeGaussian offers both higher accuracy and better visual quality, particularly in challenging scenarios involving extreme head poses or subtle gaze variations. Furthermore, GHA is limited to animating a single avatar, whereas GazeGaussian with the proposed neural-guided renderer achieves better generalization ability and identity preservation across different subjects.

We present a visualization comparison of different methods in Fig. 3. GHA fails to generalize to different subjects and struggles to preserve personal identity in the rendered images, which is quantitatively verified as the low 'identity similarity' in Tab. 1. Moreover, GHA produces blurred and unrealistic eye regions, significantly degrading the visual quality of gaze redirection. GazeNeRF, which implicitly rotates the feature map, fails to effectively control eye appearance under extreme gaze directions (as shown in the last row). Furthermore, GazeNeRF struggles with rendering
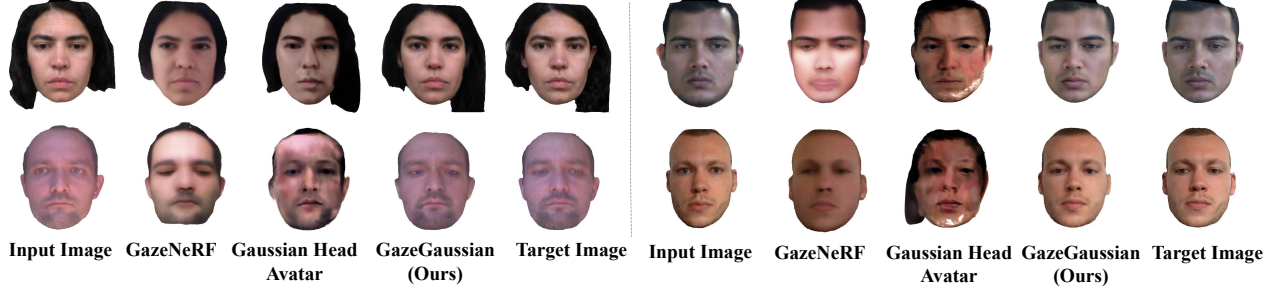
Figure 4. Cross-dataset visualization: head images are from the MPIIFaceGaze test set using our GazeGaussian, GazeNeRF, and Gaussian Head Avatar. Please refer to Sec. 12.4 in the supplementary for more visualization.
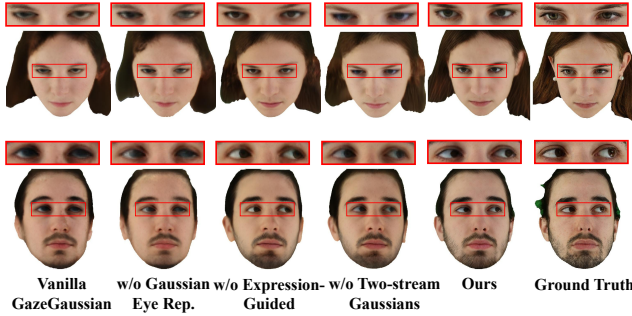

Figure 5. Qualitative ablation study on the ETH-XGaze dataset.

fine-grained facial details and exhibits notable artifacts in hair rendering. The failure to precisely render eye details in GHA and GazeNeRF restricts their capability in gaze redirection. In contrast, GazeGaussian consistently produces highly realistic results, even under challenging conditions, setting a new benchmark for gaze redirection tasks.

5.3. Cross-dataset Comparison

To access the generalization capability of GazeGaussian, we train all methods on the ETH-Xgaze dataset and evaluate them on three other datasets: ColumbiaGaze, MPIIFaceGaze, and the test set of GazeCapture. The training setup remains consistent with the within-dataset evaluation, using the same model configurations and trained parameters.

The results shown in Tab. 2 and Fig. 4 demonstrate that GazeGaussian consistently outperforms all other methods across the three datasets and all evaluation metrics. By introducing a novel expression-guided neural renderer, Gaze-Gaussian gains better generalization ability by integrating person-specific features, resulting in optimal identity preservation. On the other hand, GHA's performance is limited by its modeling strategy, showing poor adaptability to unseen datasets. It produces less clear eye regions and achieves significantly lower identity similarity scores compared to GazeGaussian. These results further validate the superiority of GazeGaussian, making it a more robust choice for handling diverse datasets and complex gaze redirection tasks.

5.4. Ablation Study

To validate the effectiveness of each component, we conduct a component-wise ablation study on the ETH-XGaze dataset. The results are shown in Tab. 3 and Fig. 5. Due to the space limitation, Please refer to Sec. 11.1 in the supplementary material for the ablation study on cross-dataset.

Vanilla-GazeGaussian. In this version, we omit the proposed Gaussian eye rotation representation and expression-guided neural renderer. The corresponding experimental results are shown in the first row of the table and the first column of the visualizations. The eye deformation is treated the same as the face, and the neural renderer remains unchanged from GazeNeRF. The results show that, due to the lack of control over eye rotation, gaze redirection errors are large, and the image synthesis quality is relatively low.

w/o Gaussian eye rotation representation. To verify the contribution of the proposed Gaussian eye rotation representation, we omit it in the GazeGaussian. The results are shown in the second row of the table and the second column of the figure. Compared to the full version of GazeGaussian, the introduction of a specialized representation for eye deformation significantly improves gaze redirection accuracy and enhances the detail in the eye region.

w/o Expression-Guided. We remove the expression-guided neural renderer and rely solely on the vanilla neural renderer in GazeNeRF for image synthesis. The results, shown in the third row of the table and the third column of the figure, indicate a noticeable decline in image quality. Without expression guidance, the model struggles to preserve identity across different subjects, leading to worse identity similarity. The rendered images exhibit lower fidelity in capturing facial details owing to the absence of person-specific features.

w/o Two-stream. Replacing the two-stream structure with a single-stream Gaussian model for both face and eye regions leads to performance degradation and loss of synthesis details, as shown in the fourth row of the table and the fourth column of the figure. Combining face and eye regions in a single stream fails to capture the eye region's complex dynamics, resulting in less accurate gaze redirection and lower image fidelity. The two-stream architecture, which

Table 2. Cross-dataset comparison: Quantitative results of GazeGaussian to other SOTA baselines on ColumbiaGaze, MPIIFaceGaze, and GazeCapture datasets in terms of gaze and head redirection errors in degree, LPIPS, and identity similarity (ID).

Method	ColumbiaGaze				MPIIFaceGaze				GazeCapture			
	Gaze↓	Head↓	LPIPS↓	ID↑	Gaze↓	Head↓	LPIPS↓	ID↑	Gaze↓	Head↓	LPIPS↓	ID↑
STED	17.887	14.693	0.413	6.384	14.796	11.893	0.288	10.677	15.478	16.533	0.271	6.807
HeadNeRF	15.250	6.255	0.349	23.579	14.320	9.372	0.288	31.877	12.955	10.366	0.232	20.981
GazeNeRF	9.464	3.811	0.352	23.157	14.933	7.118	0.272	30.981	10.463	9.064	0.232	19.025
Gaussian Head Avatar	10.939	3.953	0.347	46.183	12.021	8.530	0.295	30.932	11.571	7.664	0.295	22.236
GazeGaussian (Ours)	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 3. Component-wise ablation study of GazeGaussian on the ETH-XGaze dataset in terms of gaze and head redirection errors in degree, redirection image quality (SSIM, PSNR, LPIPS and FID), and identity similarity (ID).

Two-stream	Gaussian Eye Rep.	Expression-Guided	Gaze↓	Head Pose↓	SSIM↑	PSNR↑	LPIPS↓	FID↓	Identity Similarity↑
✓			13.651	2.981	0.753	16.376	0.272	55.481	38.941
✓		✓	13.489	3.149	0.751	16.365	0.274	54.327	41.521
✓	✓		8.883	2.635	0.766	16.692	0.254	48.891	45.013
	✓	✓	7.494	3.098	0.769	16.873	0.250	49.658	46.155
✓	✓	✓	6.622	2.128	0.823	18.734	0.216	41.972	67.749

decouples the face and eye regions, enables more precise modeling of each region's unique characteristics, improving gaze accuracy and image quality.

Among the ablation experiments, the full GazeGaussian achieves the best performance. This improvement results from the two-stream Gaussian structure, which decouples the face and eye regions, and the proposed Gaussian eye rotation representation, which enables accurate control of eye rotation. Additionally, the expression-guided neural renderer enhances the model's ability to generalize across subjects while preserving facial details.
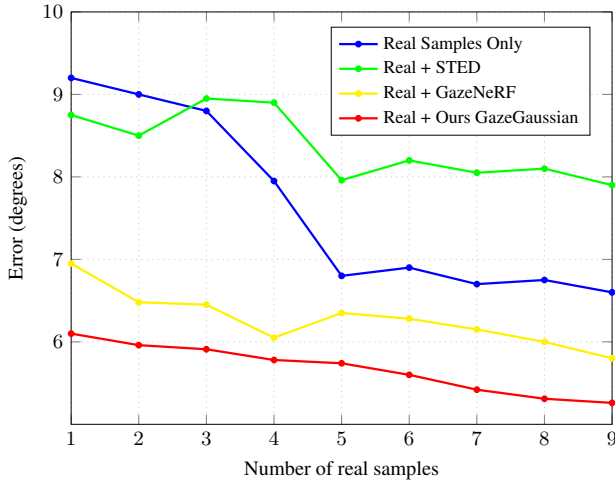

Figure 6. Improvement for gaze estimation accuracy.

5.5. Improving gaze estimation performance

Following GazeNeRF, we evaluate the improvement of each method for downstream gaze estimation tasks. Specifically,

following settings in GazeNeRF, given a few calibration samples from person-specific ETH-Xgaze test sets, we augment these real samples with gaze-redirected samples generated by GazeGaussian. We then fine-tune the gaze estimator (same in GazeNeRF) pre-trained on ETH-XGaze using these augmented samples and compare its performance with a baseline model fine-tuned only on real samples. To ensure a fair comparison, the total number of augmented samples is fixed at 200 (real + generated samples), and we vary the number of real samples used for fine-tuning during the evaluation.

As shown in Fig. 6, the x-axis represents the number of real samples used, and the y-axis shows the gaze estimation error in degrees on the ETH-XGaze person-specific test set. We evaluate up to nine real samples in the few-shot setting. Fine-tuning the pre-trained gaze estimator with real and generated samples from GazeGaussian achieves the lowest gaze estimation error across all settings. In contrast, samples generated by GazeNeRF and STED lead to higher gaze estimation errors. The results clearly demonstrate that the generated samples from GazeGaussian are of higher fidelity and more effective for improving gaze estimation accuracy.

6. Conclusion

We present GazeGaussian, the first high-fidelity gaze redirection pipeline that uses a two-stream 3DGS model for face and eye disentanglement, along with an expression-conditional neural renderer. Numerous experiments have demonstrated that GazeGaussian achieves state-of-the-art performance on the task of gaze direction, paving the way for more robust gaze estimation in real-world applications.

References

[1] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. Conversational gaze aversion for humanlike robots. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 25–32, 2014. 1

[2] Peng Chen, Xiaobao Wei, Ming Lu, Yitong Zhu, Naiming Yao, Xingyu Xiao, and Hui Chen. Diffusiontalker: Personalization and acceleration for speech-driven 3d face diffuser. *arXiv preprint arXiv:2311.16565*, 2023. 2

[3] Yihua Cheng and Feng Lu. Gaze estimation using transformer. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3341–3347. IEEE, 2022. 1

[4] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[5] Criminisi, Shotton, Blake, and Torr. Gaze manipulation for one-to-one teleconferencing. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 191–198 vol.1, 2003. 2

[6] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10101–10111, 2019. 2

[7] Helisa Dhamo, Yinyu Nie, Arthur Moreau, Jifei Song, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Headgas: Real-time animatable head avatars via 3d gaussian splatting. *arXiv preprint arXiv:2312.02902*, 2023. 2

[8] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18780, 2022. 2

[9] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 2

[10] Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *Computer Vision – ECCV 2016*, pages 311–326, Cham, 2016. Springer International Publishing. 1, 2

[11] Dominik Giger, Jean-Charles Bazin, Claudia Kuster, Tiberiu Popa, and Markus Gross. Gaze correction with a single webcam. *IEEE International Conference on Multimedia & Expo*, 2014. 2

[12] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[14] Zhe He, Adrian Spurr, Xucong Zhang, and Otmar Hilliges. Photo-realistic monocular gaze redirection using generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2019. 2

[15] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 5

[16] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 1, 2

[17] Nan Huang, Xiaobao Wei, Wenzhao Zheng, Pengju An, Ming Lu, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. S3gaussian: Self-supervised street gaussians for autonomous driving. *arXiv preprint arXiv:2405.20323*, 2024. 1

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 2, 3

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*, 2015. 1

[20] Daniil Kononenko and Victor Lempitsky. Learning to look up: Realtime monocular gaze correction using machine learning. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4667–4675, 2015. 2

[21] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016. 5, 1

[22] Claudia Kuster, Tiberiu Popa, Jean-Charles Bazin, Craig Gotsman, and Markus Gross. Gaze correction for home video conferencing. *ACM Trans. Graph.*, 31(6), 2012. 2

[23] Gengyan Li, Abhimitra Meka, Franziska Mueller, Marcel C Buehler, Otmar Hilliges, and Thabo Beeler. Eyenerf: a hybrid representation for photorealistic synthesis, animation and relighting of human eyes. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022. 2

[24] 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. 2

[25] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20654–20664, 2024. 1

[26] Haimin Luo, Min Ouyang, Zijun Zhao, Suyi Jiang, Longwen Zhang, Qixuan Zhang, Wei Yang, Lan Xu, and Jingyi Yu. Gaussianhair: Hair modeling and rendering with light-aware gaussians. *arXiv preprint arXiv:2402.10483*, 2024. 3

[27] Annu George Mavely, JE Judith, PA Sahal, and Steffy Ann Kuruvilla. Eye gaze tracking based driver monitoring system. In *2017 IEEE international conference on circuits and systems (ICCS)*, pages 364–367. IEEE, 2017. 1

[28] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020. 1, 2

[29] Nitish Padmanaban, Robert Konrad, Emily A Cooper, and Gordon Wetzstein. Optimizing vr for all users through adaptive focus displays. In *ACM SIGGRAPH 2017 Talks*, pages 1–2, 2017. 1

[30] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9368–9377, 2019. 2

[31] Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5292–5301, 2023. 2

[32] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20687–20697, 2023. 2

[33] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 1, 2

[34] Yalun Qin, Kuo-Chin Lien, Matthew Turk, and Tobias Höllerer. Eye gaze correction with a single webcam based on eye-replacement. In *Advances in Visual Computing*, pages 599–609, Cham, 2015. Springer International Publishing. 2

[35] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European conference on computer vision (ECCV)*, pages 704–720, 2018. 2

[36] Alessandro Ruzzi, Xiangwei Shi, Xi Wang, Gengyan Li, Shalini De Mello, Hyung Jin Chang, Xucong Zhang, and Otmar Hilliges. Gazenerf: 3d-aware gaze redirection with neural radiance fields. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9676–9685, 2023. 1, 2, 4, 5

[37] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021. 3

[38] Zhixin Shu, Eli Shechtman, Dimitris Samaras, and Sunil Hadap. Eyeopener: Editing eyes in the wild. *ACM Trans. Graph.*, 36(1), 2016. 2

[39] B.A. Smith, Q. Yin, S.K. Feiner, and S.K. Nayar. Gaze Locking: Passive Eye Contact Detection for Human?Object Interaction. In *ACM Symposium on User Interface Software and Technology (UIST)*, pages 271–280, 2013. 5, 1

[40] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1821–1828, 2014. 5

[41] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7346–7355, 2018. 2

[42] Hengfei Wang, Zhongqun Zhang, Yihua Cheng, and Hyung Jin Chang. High-fidelity eye animatable neural radiance fields for human face. *arXiv preprint arXiv:2308.00773*, 2023. 1, 2, 5

[43] Jun Wang, Yinglu Liu, Yibo Hu, Hailin Shi, and Tao Mei. Facex-zoo: A pytorch toolbox for face recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3779–3782, 2021. 5

[44] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1

[45] Yu Wang, Xiaobao Wei, Ming Lu, and Guoliang Kang. Plgs: Robust panoptic lifting with 3d gaussian splatting. *arXiv preprint arXiv:2410.17505*, 2024. 1

[46] Xiaobao Wei, Renrui Zhang, Jiarui Wu, Jiaming Liu, Ming Lu, Yandong Guo, and Shanghang Zhang. Nto3d: Neural target object 3d reconstruction with segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20352–20362, 2024. 1

[47] Erroll Wood, Tadas Baltrusaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. Gazedirector: Fully articulated eye gaze redirection in video, 2017. 2

[48] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Wensen Feng. Controllable continuous gaze redirection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1782–1790, 2020. 2

[49] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

[50] Mingjie Xu, Haofei Wang, and Feng Lu. Learning a generalized gaze estimator from gaze-consistent feature. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3027–3035, 2023. 1

[51] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3, 4, 5

[52] Pengwei Yin, Jingjing Wang, Jiawu Dai, and Xiaojun Wu. Nerf-gaze: A head-eye redirection parametric model for gaze estimation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2760–2764. IEEE, 2024. 1, 2, 5

[53] Yu Yu and Jean-Marc Odobez. Unsupervised representation learning for gaze estimation. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 7314–7324, 2020. 1

[54] Mingfang Zhang, Yunfei Liu, and Feng Lu. Gazeonce: Real-time multi-person gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4197–4206, 2022. 1

[55] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520, 2015. 5, 1

[56] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 2299–2308. IEEE, 2017. 5, 1

[57] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proc. International Symposium on Eye Tracking Research and Applications (ETRA)*, pages 12:1–12:9, 2018. 1, 5

[58] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision (ECCV)*, 2020. 5, 1

[59] Yufeng Zheng, Seonwook Park, Xucong Zhang, Shalini De Mello, and Otmar Hilliges. Self-learning transformations for improving gaze and head redirection. In *Neural Information Processing Systems (NeurIPS)*, 2020. 2, 5

[60] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21057–21067, 2023. 1, 2

[61] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pages 250–269. Springer, 2022. 2

[62] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4574–4584, 2023. 1, 2

[63] zllrunning. Using modified bisenet for face parsing in pytorch, 2019. https://github.com/zllrunning/face-parsing.PyTorch. 5, 1