CVPR
#3250

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

Anonymous CVPR submission

Paper ID 3250

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, a high-fidelity gaze redirection method that uses a two-stream 3DGS model to represent the face and eye regions separately. By leveraging the unstructured nature of 3DGS, we develop a novel eye representation for rigid eye rotation based on the target gaze direction. To enhance synthesis generalization across various subjects, we integrate an expression-conditional module to guide the neural renderer. Comprehensive experiments show that GazeGaussian outperforms existing methods in rendering speed, gaze redirection accuracy, and facial synthesis across multiple datasets. We also demonstrate that existing gaze estimation methods can leverage GazeGaussian to improve their generalization performance. The code will be released.

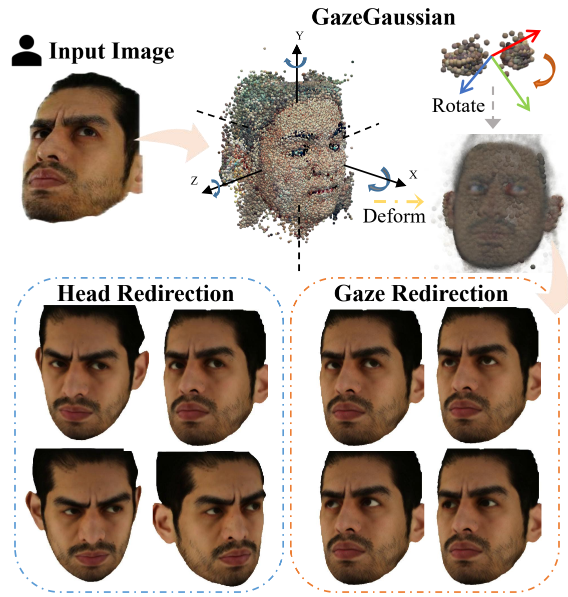

Figure 1. GazeGaussian for gaze redirection: Given an input image, GazeGaussian deforms face and eye Gaussians from canonical space to generate high-fidelity head images with accurate gaze redirection.

1. Introduction

Gaze estimation is a fundamental component across various applications [1, 25, 27], yet current estimators [3, 4, 48] often struggle to generalize effectively to out-of-distribution data. To address this, recent approaches [34, 40, 51] 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, 52, 53, 56] formulate gaze redirection as a 2D image manipulation task, relying on deep learning techniques to warp eye regions of the image toward 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) [26] and its variants [42, 44], several methods [12, 16, 59, 61] have achieved 3D dynamic head representation and high-fidelity avatar synthesis. Meanwhile, to enable precise control of gaze direction, recent research [34, 40, 51] 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, 24, 43] achieve impressive rendering quality with significantly faster training speeds. Recent research [31, 47, 50] has applied these methods to 3D head animation, typically using face-tracking [39, 60] 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 the best of 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 (GERR). 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 enhance synthesis generalization across different subjects and preserve facial details, 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 high-fidelity 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 enhance the synthesis generalization of 3DGS, we design an expression-guided neural renderer (EGNR) to retain facial details across various subjects.
- 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, 21], eye-replacement [32, 36], and warping-based methods [10, 19, 45]. However, these methods are limited by person-specific data requirements, restricted redirection range, and artifact introduction. To further improve gaze redirection, recent studies [14, 28, 46, 58] have employed neural network-based generative models. STED [58], building on the FAZE [28], introduces a self-transforming encoder-decoder that generates full-face images with high-fidelity control over gaze direction and head pose. Effective gaze redirection should account for both the 3D nature of eyeball rotation and the deformation of surrounding facial regions. With advancements in Neural Radiance Fields (NeRF) [26], several studies [22, 34, 40, 51] have aimed to model the complex rotation of the eyeball. GazeNeRF [34] employs a two-stream MLP architecture to separately model the face only and eye regions, achieving improved gaze redirection performance.

However, these methods are hindered by substantial computational demands and limited rendering efficiency. Additionally, gaze manipulation occurs at the feature map level 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 [23] is a parameterized 3D head model that maps parameters of shape, expression, and pose onto 3D facial geometry, allowing for realistic and controllable head avatar generation. Many subsequent works [2, 6, 8, 29, 30, 33] 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, 59, 61] 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], serveral approaches [7, 31, 47, 50] have explored its application in head avatar modeling. Gaussian Head Avatar [49] initializes Gaussians with a neutral mesh head and incorporates MLPs to deform complex facial expressions.

While these methods produce impressive results in creating 3D head avatars, they overlook precise gaze control and do not generalize well across different subjects. In contrast, GazeGaussian emphasizes precise gaze direction control by decoupling facial animations and gaze movement within a two-stream model. Furthermore, we introduce an expression-guided neural renderer designed to improve the quality of synthesis.

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,

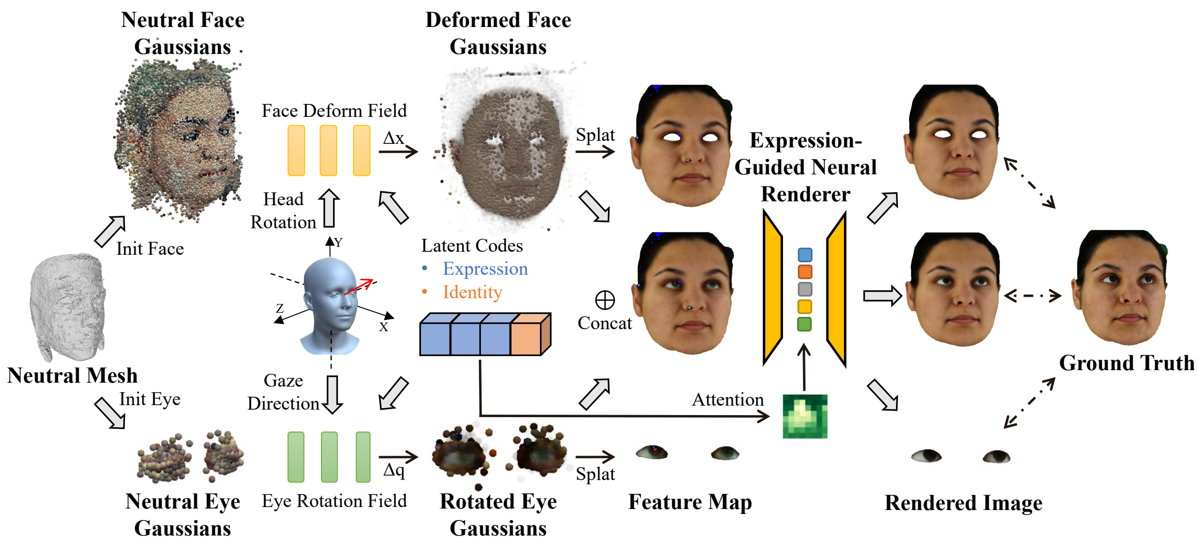

Figure 2. Pipeline of GazeGaussian. We initialize face-only and eye regions from a pre-trained neutral mesh. Using target expression codes, head rotation, and gaze direction, GazeGaussian optimizes face deformation and eye rotation fields to transform the neutral Gaussians. The transformed Gaussians are splatted into feature maps. The expression codes guide the neural renderer through cross-attention, enabling the rendering of feature maps into high-fidelity images, which are then supervised by multi-view RGB images.

we follow the data preprocessing in GazeNeRF [34] and Gaussian Head Avatar [50], which include background removal, 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 [35]. 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 positions X, the multi-channel color C, the rotation Q, scale S and opacity A. The color C is computed using spherical harmonics, and the rotation Q is represented as the quaternion. These Gaussians are then rasterized and rendered to a multi-channel image I based on the camera parameters μ. This rendering process can be expressed as:

$$I = \mathcal{R}(X, C, Q, S, A; \mu), \tag{1}$$

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 branch and an eye branch. In the following subsections, we will describe the face deformation and eye rotation processes, respectively.

4.2.1. Face Deformation

For the face-only branch, inspired by Gaussian Head Avatar, we first construct canonical neutral face Gaussians with attributes: $\{\boldsymbol{X}_0^f, \boldsymbol{F}_0^f, \boldsymbol{Q}_0^f, \boldsymbol{S}_0^f, \boldsymbol{A}_0^f\}$, which are fully optimizable. $\boldsymbol{X}_0^f \in \mathbb{R}^{N\times 3}$ represents the positions of the Gaussians in the canonical space. $\boldsymbol{F}_0^f \in \mathbb{R}^{N\times 128}$ denotes the point-wise feature vectors as their intrinsic properties. $\boldsymbol{Q}_0^f \in \mathbb{R}^{N\times 4}$, $\boldsymbol{S}_0^f \in \mathbb{R}^{N\times 3}$ and $\boldsymbol{A}_0^f \in \mathbb{R}^{N\times 1}$ denotes the neutral rotation, scale and opacity respectively. The neutral color is directly predicted from the point-wise feature vectors $\boldsymbol{F}_0^f$. Then we construct several MLPs, denoted as $\boldsymbol{\Phi}^f$, to serve as face deformation fields that transform the neutral face Gaussians. Next, we describe the process of applying offsets to each Gaussian attribute.

Positions X^f of the Gaussians. We predict the displacements respectively controlled by the latent codes and the head pose in the canonical space through two different MLPs:

CVPR #3250

${}^{exp}_{def}\mathcal{F}^f \in \boldsymbol{\Phi}^f$ and ${}^{pose}_{def}\mathcal{F}^f \in \boldsymbol{\Phi}^f$. Then, we add them to the neutral positions.

$$\begin{aligned} X^f = \boldsymbol{X}_0^f &+ \lambda_{exp}(\boldsymbol{X}_0^f){}^{exp}_{def}\mathcal{F}^f(\boldsymbol{X}_0^f, \theta) \\ &+ \lambda_{pose}(\boldsymbol{X}_0^f){}^{pose}_{def}\mathcal{F}^f(\boldsymbol{X}_0^f, \beta), \end{aligned} \tag{2}$$

θ denoting latent codes including expression and identity coefficients and β denoting the head pose. $\lambda_{exp}(\cdot)$ and $\lambda_{pose}(\cdot)$ represent the degree to which the point is influenced by the expression or head pose, respectively, which can be calculated as:

$$\lambda_{exp}(x) = \begin{cases} 1, & dist(x, \boldsymbol{P}_0^f) < t_1 \\ \frac{t_2 - dist(x, \boldsymbol{P}_0^f)}{t_2 - t_1}, & dist(x, \boldsymbol{P}_0^f) \in [t_1, t_2] \\ 0, & dist(x, \boldsymbol{P}_0^f) > t_2 \end{cases}$$

with $\lambda_{pose}(x) = 1 - \lambda_{exp}(x)$, where $x \in \mathbf{X}_0^f$ denotes the position of a neutral Gaussian, $dist(x, \boldsymbol{P}_0^f)$ represents the minimum distance from point x to the 3D landmarks (without eyes) $\boldsymbol{P}_0^f$. Following the approach in Gaussian Head Avatar, the predefined hyperparameters are set as $t_1 = 0.15$ and $t_2 = 0.25$.

Color C^f of the Gaussians. Modeling dynamic details requires a color that varies with expressions. The color is directly predict by two color MLPs: ${}^{exp}_{col}\mathcal{F}^f \in \boldsymbol{\Phi}^f$ and ${}^{pose}_{col}\mathcal{F}^f \in \boldsymbol{\Phi}^f$:

$$\begin{aligned} C^f = &\lambda_{exp}(\boldsymbol{X}_0^f){}^{exp}_{col}\mathcal{F}^f(\boldsymbol{F}_0^f, \theta) \\ &+ \lambda_{pose}(\boldsymbol{X}_0^f){}^{pose}_{col}\mathcal{F}^f(\boldsymbol{F}_0^f, \beta), \end{aligned} \tag{3}$$

Rotation, **Scale** and **Opacity** $\{Q^f, S^f, A^f\}$ of the Gaussians. These three attributes are also dynamic, capturing detailed expression-related appearance changes. We just use another two attribute MLPs: ${}^{exp}_{att}\mathcal{F}^f \in \boldsymbol{\Phi}^f$ and ${}^{pose}_{att}\mathcal{F}^f \in \boldsymbol{\Phi}^f$ to predict their shift from the neutral value.

$$\begin{aligned} \{Q^f, S^f, A^f\} = \{&\mathbf{Q}_0^f, \mathbf{S}_0^f, \mathbf{A}_0^f\} \\ &+ \lambda_{exp}(\boldsymbol{X}_0^f){}^{exp}_{att}\mathcal{F}^f(\boldsymbol{F}_0^f, \theta) \\ &+ \lambda_{pose}(\boldsymbol{X}_0^f){}^{pose}_{att}\mathcal{F}^f(\boldsymbol{F}_0^f, \beta), \end{aligned} \tag{4}$$

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. The above face-only branch can be formulated as:

$$\begin{aligned} \mathcal{M}_f &= \mathcal{R}(\{X^f, C^f, Q^f, S^f, A^f\}) \\ &= \mathcal{R}(\boldsymbol{\Phi}^f(\boldsymbol{X}_0^f, \boldsymbol{F}_0^f, \boldsymbol{Q}_0^f, \boldsymbol{S}_0^f, \boldsymbol{A}_0^f; \theta, \beta)), \end{aligned} \tag{5}$$

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

4.2.2. Eye Rotation

For the eye branch, we also construct canonical neutral eye Gaussians with attributes $\{\boldsymbol{X}_0^e, \boldsymbol{F}_0^e, \boldsymbol{Q}_0^e, \boldsymbol{S}_0^e, \boldsymbol{A}_0^e\}$. These attributes share the same dimensionality as those in the face-only branch, except that $\boldsymbol{S}_0^e \in \mathbb{R}^{N\times 1}$ is constrained to be spherical, aligning with the rotational properties of the eyeball. Next, we describe the process of applying offsets to each Gaussian attribute.

Positions X^e of the Gaussians. Directly applying the same deformation strategy as for the face branch would fail to fully leverage the unique characteristics of eyeball rotational motion, resulting in insufficient gaze redirection accuracy. Therefore, we first rotate the eye Gaussians in the canonical space and then incorporate the eye geometry information from the latent codes of different subjects to generate biases. Since the gaze labels may 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: ${}^{gaze}_{rot}\mathcal{F}^e \in \boldsymbol{\Phi}^e$ and ${}^{exp}_{def}\mathcal{F}^e \in \boldsymbol{\Phi}^e$ to predict the biases for Gaussian rotation and displacement.

$$X^e = {}^{exp}_{def}\mathcal{F}^f(\boldsymbol{X}_0^e, \theta) + {}^{gaze}_{rot}\mathcal{F}^e(\boldsymbol{X}_0^e, \varphi)\boldsymbol{X}_0^e, \tag{6}$$

Since eyes are relatively small and mainly influenced by the gaze direction, λ used in the face is omitted here.

Color C^e of the Gaussians. The color of the eye region is influenced by the gaze direction and latent codes. We use two MLPs: ${}^{exp}_{col}\mathcal{F}^e \in \boldsymbol{\Phi}^e$ and ${}^{gaze}_{col}\mathcal{F}^e \in \boldsymbol{\Phi}^e$ to predict it:

$$C^e = {}^{exp}_{att}\mathcal{F}^e(\boldsymbol{F}_0^e, \theta) + {}^{gaze}_{col}\mathcal{F}^e(\boldsymbol{X}_0^e, \varphi), \tag{7}$$

Rotation, **Scale** and **Opacity** $\{Q^e, S^e, A^e\}$ of the Gaussians. We just use another two attribute MLPs ${}^{exp}_{att}\mathcal{F}^e \in \boldsymbol{\Phi}^e$ and ${}^{gaze}_{att}\mathcal{F}^e \in \boldsymbol{\Phi}^e$ to predict their shift.

$$\begin{aligned} \{Q^e, S^e, A^e\} = \{\mathbf{Q}_0^e, \mathbf{S}_0^e, \mathbf{A}_0^e\} &+ {}^{exp}_{att}\mathcal{F}^e(\boldsymbol{F}_0^e, \theta) \\ &+ {}^{gaze}_{att}\mathcal{F}^e(\boldsymbol{F}_0^e, \varphi), \end{aligned} \tag{8}$$

Finally, we transform Gaussians in the canonical space to the world space. Then these eye Gaussians are rasterized into the feature maps. The eye branch is formulated as:

$$\begin{aligned} \mathcal{M}_e &= \mathcal{R}(\{X^e, C^e, Q^e, S^e, A^e\}) \\ &= \mathcal{R}(\boldsymbol{\Phi}^e(\boldsymbol{X}_0^e, \boldsymbol{F}_0^e, \boldsymbol{Q}_0^e, \boldsymbol{S}_0^e, \boldsymbol{A}_0^e; \theta, \varphi)), \end{aligned} \tag{9}$$

To obtain the full head rendering, we simply concat the two-stream Gaussians and rasterized them into feature maps:

$$\begin{aligned} \mathcal{M}_h = \mathcal{R}(\{&X^f, C^f, Q^f, S^f, A^f\} \\ &\{X^e, C^e, Q^e, S^e, A^e\}), \end{aligned} \tag{10}$$

4.3. Expression-Guided Neural Renderer

After obtaining the rasterized feature maps from Gaussians, a UNet-like neural renderer $\mathbf{R}$ opts to synthesize the final face-only, eyes, and head images $\{\mathcal{I}_f, \mathcal{I}_e, \mathcal{I}_h\}$:

$$\{\mathcal{I}_f, \mathcal{I}_e, \mathcal{I}_h\} = \mathbf{R}(\{\mathcal{M}_f, \mathcal{M}_e, \mathcal{M}_h\}), \tag{11}$$

To enhance the generalization ability across different subjects, we inject the latent codes θ into the neural renderer through a slice cross-attention module. Let $\boldsymbol{F}_b$ represent the bottleneck feature obtained from the encoder of $\mathbf{R}$. 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:

$$\boldsymbol{F}'_b = \boldsymbol{F}_b + \boldsymbol{F}_b \cdot \text{Attn}(q=\theta, k=\boldsymbol{F}_b, v=\boldsymbol{F}_b)), \tag{12}$$

where $\text{Attn}(\cdot)$ denotes the cross-attention operation that fuses the latent codes with the bottleneck feature. Then the refined feature $\boldsymbol{F}'_b$ 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, 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 the RGB colors. For each rendered image and its corresponding feature map, we apply the same loss functions. Taking the rendered eye image 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:

$$\begin{aligned}\mathcal{L}^e_{\mathcal{I}} = ||\mathcal{I}_{gt} - \mathcal{I}_e||_1 + \lambda_{SSIM}(1 - SSIM(\mathcal{I}_{gt}, \mathcal{I}_e)) \\ + \lambda_{VGG} VGG(\mathcal{I}_{gt}, \mathcal{I}_e),\end{aligned} \tag{13}$$

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

$$\mathcal{L}_{\mathcal{I}} = \mathcal{L}^f_{\mathcal{I}} + \mathcal{L}^e_{\mathcal{I}} + \mathcal{L}^h_{\mathcal{I}} + \mathcal{L}^f_{\mathcal{M}} + \mathcal{L}^e_{\mathcal{M}} + \mathcal{L}^h_{\mathcal{M}}, \tag{14}$$

where $\mathcal{L}^f_{\mathcal{I}}, \mathcal{L}^e_{\mathcal{I}}, \mathcal{L}^h_{\mathcal{I}}$ represent the losses for the rendered face-only, eye, and head images, respectively. $\mathcal{L}^f_{\mathcal{M}}, \mathcal{L}^e_{\mathcal{M}}, \mathcal{L}^h_{\mathcal{M}}$ 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 [58] and GazeNeRF [34]. The gaze redirection loss can be formulated as:

$$\begin{aligned}\mathcal{L}_{\mathcal{G}}(\mathcal{I}_h, \mathcal{I}_{gt}) = \mathcal{E}_{\text{ang}}(\psi^g(\mathcal{I}_{wf}), \psi^g(\mathcal{I}_{gt})) \\ \mathcal{E}_{\text{ang}}(\mathbf{v},\ \hat{\mathbf{v}}) = \arccos \frac{\mathbf{v} \cdot \hat{\mathbf{v}}}{\|\mathbf{v}\|\ \|\hat{\mathbf{v}}\|},\end{aligned} \tag{15}$$

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:

$$\mathcal{L} = \lambda_{\mathcal{I}} \mathcal{L}_{\mathcal{I}} + \lambda_{\mathcal{G}} \mathcal{L}_{\mathcal{G}}, \tag{16}$$

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 [57], testing GazeGaussian alongside state-of-the-art gaze redirection and head generation methods. Next, we perform a cross-dataset comparison on ColumbiaGaze [37], MPIIFaceGaze [54, 55], and GazeCapture [20] to assess generalization. We also conduct an ablation study to analyze the contributions of each component in GazeGaussian. Additionally, we validate the impact of synthesized data on gaze estimator performance in the supplementary materials. Due to space limitations, please refer to the supplementary for more details on the experiment and visualization results.

5.1. Experimental Settings

Dataset Pre-processing. Following GazeNeRF's preprocessing, we normalize raw images [38, 56] 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 [62]. We also use the 3D face tracking method from [50] to produce identity and expression codes and camera poses for the input of our method. For consistency, gaze labels are converted to pitch-yaw angles in the head coordinate system across all datasets. Details are provided in the supplementary materials.

Baselines. We compare our method with the self-supervised gaze redirection approach STED [58], along with

CVPR #3250

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

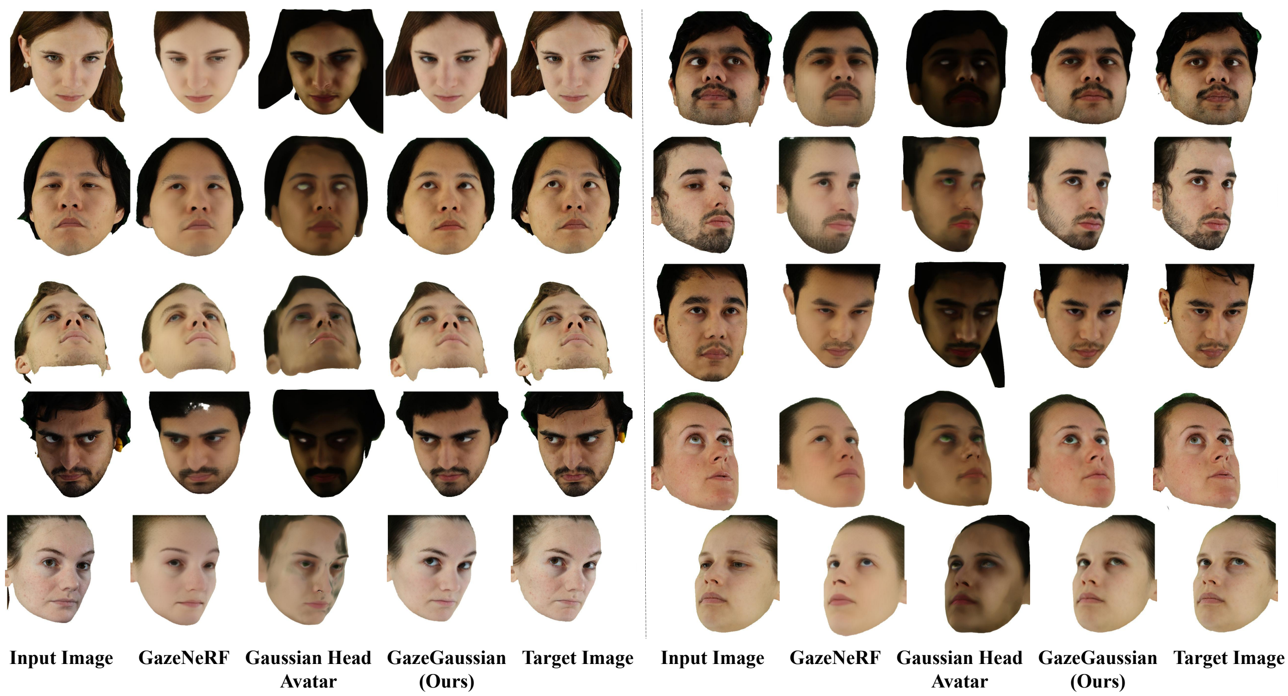

Figure 3. Within-dataset comparison: Visualization of generated images from the ETH-XGaze test set using our GazeGaussian, GazeNeRF, and Gaussian Head Avatar. All faces are masked to remove the background. GazeGaussian generates photo-realistic images with the target gaze direction, preserving identity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

NeRF-based models such as HeadNeRF [15] and the state-of-the-art method GazeNeRF [34], as well as the latest 3DGS-based head synthesis method, Gaussian Head Avatar [50]. As the NeRF-based methods, NeRF-Gaze [51] and Wang *et al*. [40] 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, and rendering speed. Redirection accuracy is measured by gaze and head poses angular errors, using a ResNet50 [13]-based estimator, as in GazeNeRF [34]. Image quality is assessed with SSIM, PSNR, LPIPS, and FID. Identity preservation is evaluated with FaceX-Zoo [41], 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. The evaluation is conducted on the person-specific test set of the ETH-XGaze dataset. This test 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-

CVPR
#3250

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**

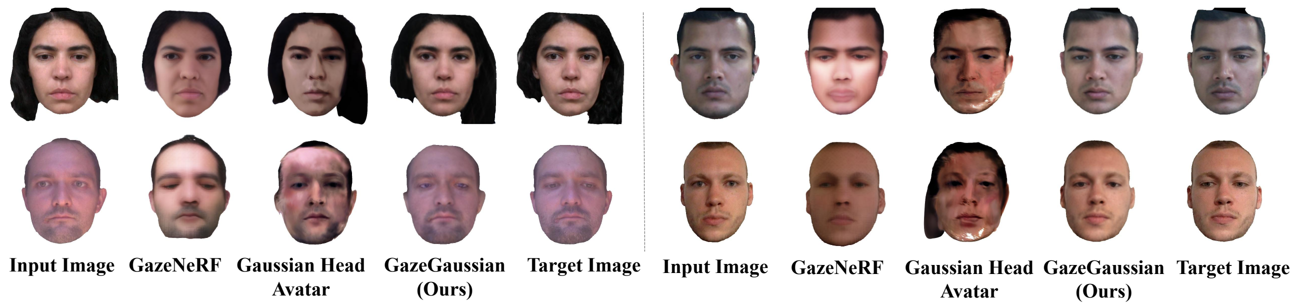

Figure 4. Cross-dataset comparison: Visualization of generated images from the MPIIFaceGaze test set using our GazeGaussian, GazeNeRF, and Gaussian Head Avatar. Please refer to the supplementary for more visualization.

Gaussian consistently outperforms prior methods across all metrics. Specifically, our approach achieves the lowest errors in both gaze and head redirection (6.622° and 2.128°, respectively), demonstrating its superior precision in gaze and head control. Compared to the previous SOTA method GazeNeRF, which applies rotation to feature map for gaze redirection, GazeGaussian adopts a Gaussian eye rotation to explicitly control eye movement. Such a technique not only improves redirection accuracy but also significantly boosts rendering quality. 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 model built on Gaussian-based representations, struggles to deliver competitive performance in gaze and head redirection tasks. The lack of dedicated mechanisms for gaze disentanglement and explicit eye region modeling in GHA leads to poor 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.

We present a qualitative comparison of different methods in Fig. 3. GHA struggles to preserve personal identity in the generated face 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, it struggles with rendering fine-grained facial details and exhibits notable artifacts in hair rendering, particularly in the last two rows. Overall, the inability to accurately handle eye details in both GHA and GazeNeRF limits their effectiveness 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 perform a cross-dataset evaluation 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, GazeGaussian can retain facial details across various subjects. 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

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.

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	38.521
✓	✓		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**

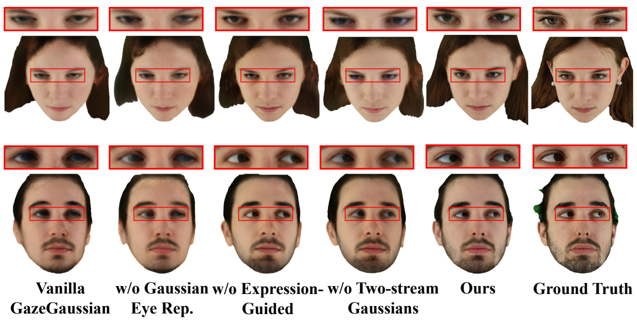

Figure 5. Qualitative ablation study on the ETH-XGaze dataset.

datasets and complex gaze redirection tasks. Please refer to supplementary material for more visualization results on the cross-dataset evaluation.

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.

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 proposed expression-guided neural renderer and rely solely on the 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 effectively preserve dynamic facial expressions, leading to less accurate gaze redirection. The synthesized images also exhibit lower fidelity in capturing facial details and subtle expression changes.

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 decouples the face and eye regions, enables more precise modeling of each region's unique characteristics, improving gaze accuracy and image quality. Furthermore, when comparing this version to the vanilla GazeGaussian (where no proposed components are used), we observe a substantial performance improvement, validating the effectiveness of the proposed techniques and their contribution to gaze redirection and head avatar synthesis.

Among all the ablation experiments, the full GazeGaussian achieves the best performance. This improvement results from the combination of the two-stream Gaussian structure, which decouples the face and eye regions for more precise modeling, 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.

6. Conclusion

We present GazeGaussian, a high-fidelity gaze redirection pipeline that uses a two-stream model to represent face and eye regions separately. We present a new Gaussian-based representation of the eye to accurately depict eye rotations, along with an expression-conditional neural renderer that enhances the fidelity of gaze redirection. Numerous experiments have shown 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. 6

[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. 6

[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] 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

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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

[25] 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

[26] 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

[27] 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

[28] 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

[29] 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

[30] 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

[31] 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

[32] 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

[33] 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

[34] 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, 3, 5, 6

[35] 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

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

[37] 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

[38] 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

[39] 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. 1

[40] 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, 6

[41] 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. 6

[42] 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

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

[44] 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

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

[46] 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

[47] 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

[48] 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

[49] 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. *arXiv preprint arXiv:2312.03029*, 2023. 2

[50] 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, 5, 6

[51] 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, 6

[52] 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

[53] 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

[54] 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

[55] 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

[56] 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

[57] 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

[58] 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

[59] 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

[60] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pages 250–269. Springer, 2022. 1

[61] 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

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