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

Ruzzi, Alessandro; Shi, Xiangwei ; Wang, Xi; Li, Gengyan ; De Mello, Shalini; Chang, Hyung Jin ; Zhang, Xucong ; Hilliges, Otmar

DOI
10.1109/CVPR52729.2023.00933

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Citation (APA)
Ruzzi, A., Shi, X., Wang, X., Li, G., De Mello, S., Chang, H. J., Zhang, X., & Hilliges, O. (2023). GazeNeRF: 3D-Aware Gaze Redirection with Neural Radiance Fields. In L. O'Conner (Ed.), *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9676-9685). IEEE. https://doi.org/10.1109/CVPR52729.2023.00933

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

Alessandro Ruzzi[1*] Xiangwei Shi[2*] Xi Wang[1] Gengyan Li[1] Shalini De Mello[3]
Hyung Jin Chang[4] Xucong Zhang[2] Otmar Hilliges[1]
[1]Department of Computer Science, ETH Zürich [2]Computer Vision Lab, Delft University of Technology [3]NVIDIA
[4]School of Computer Science, University of Birmingham

2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-0129-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPR52729.2023.00933

Abstract

We propose GazeNeRF, a 3D-aware method for the task of gaze redirection. Existing gaze redirection methods operate on 2D images and struggle to generate 3D consistent results. Instead, we build on the intuition that the face region and eyeballs are separate 3D structures that move in a coordinated yet independent fashion. Our method leverages recent advancements in conditional image-based neural radiance fields and proposes a two-stream architecture that predicts volumetric features for the face and eye regions separately. Rigidly transforming the eye features via a 3D rotation matrix provides fine-grained control over the desired gaze angle. The final, redirected image is then attained via differentiable volume compositing. Our experiments show that this architecture outperforms naïvely conditioned NeRF baselines as well as previous state-of-the-art 2D gaze redirection methods in terms of redirection accuracy and identity preservation. Code and models will be released for research purposes.

1. Introduction

Gaze redirection is the task of manipulating an input image of a face such that the face in the output image appears to look at a given target direction, without changing the identity or other latent parameters of the subject. Gaze redirection finds applications in video conferencing [31], image and movie editing [3], human-computer interaction [23], and holds the potential to enhance life-likeness of avatars for the metaverse (e.g., [2,42]). It has furthermore been shown that gaze-redirected images can be used to synthesize training data for downstream tasks such as person-specific gaze estimation [7,41].

Existing gaze redirection methods formulate this task as a 2D image manipulation problem. Either by warping select pixels of the input image [3,35,36,38], or by synthesizing new images via deep generative models such as Generative Adversarial Networks (GANs) [7,10], encoder-decoder networks [22], or Variational Autoencoders (VAEs) [41]. Image warping methods can not model large changes due to the inability to generate new pixels. While 2D generative models can produce high-quality images and allow for large gaze direction changes, they do not take the 3D nature of the task into consideration. This can lead to spatio-temporal or identity inconsistencies where other latent variables are entangled with the gaze direction. Some 2D methods attempt to simulate the eyeball rotation by applying a 3D rotation matrix in latent space [22,41]. However, these injected implicit priors are weak and do not explicitly model the 3D nature of the task.

In this paper, we address these issues by reformulating gaze redirection as a 3D task and propose a novel 3D-aware gaze redirection method GazeNeRF. Our approach leverages recent advances in image-based conditional neural radiance fields [8] to inherit the ability to generate images of excellent quality. The physical face and eyes are not a monolithic 3D structure but

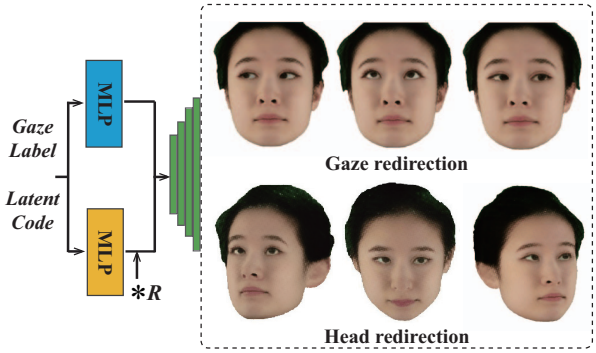

Figure 1. GazeNeRF consists of a NeRF-based two-stream-MLP structure conditioned on a target gaze label to generate the photo-realistic face images. A 3D rotation transformation **R** is applied on eyes stream of GazeNeRF.

*These two authors contributed equally to this work.

979-8-3503-0129-8/23/$31.00 ©2023 IEEE
DOI 10.1109/CVPR52729.2023.00933

are composed of two 3D structures – the face without eyes that deforms and the eyes only that rotates when we move our eyes. Hence, we model the two structures as separate feature volumes with neural radiance field (NeRF) models. To this end, our work shares similarities to EyeNeRF [16], but their focus is on high-fidelity rendering and relighting quality, whereas we are concerned with gaze redirection accuracy.

To endow NeRF architectures with 3D-aware gaze redirection capabilities, we propose a novel two-stream multilayer-perceptron (MLP) structure that predicts feature maps for the eye-balls (*eyes*) and the rest of the face region (*face only*) separately (see Fig. 1). The features of the eyes region are transformed via the desired 3D rotation matrix, before compositing the regions via differentiable volume rendering. With the explicit separation of the eyeballs, GazeNeRF rigidly rotates the 3D features which we show to be beneficial for gaze redirection accuracy. To be able to train the model, we propose the feature composition at end of the two-stream MLPs and additional training losses to enhance the functionality of gaze redirection.

We find that GazeNeRF outperforms previous state-of-the-art methods [8, 41] for gaze redirection on multiple datasets in terms of gaze and head pose redirection accuracy and identity preservation, evidencing the advantage of formulating the task as a 3D-aware problem. In summary, our contributions are as follows:

- We re-formulate the task of gaze redirection as 3D-aware neural volume rendering.
- GazeNeRF learns to disentangle the features of the face and eye regions, which allows for the rigid transformation of the eyeballs to the desired gaze direction.
- State-of-the-art performance in gaze redirection accuracy under identity preservation across different datasets.

2. Related Work

2.1. Gaze redirection

Gaze redirection can be done with the graphic model that synthesizes the eye images with different gaze directions and head poses [32]. However, these methods are expensive due to the complex appearance modelling such as albedo, diffuse, shading and illuminations, which usually require accurate facial and eye landmarks detection. Recent gaze redirection methods mainly utilize image warping method [3, 35, 36] or generative models [7, 22, 41] to redirect the gaze and/or rotate the head. Image warping estimates warping matrices between source and target images and copies the pixels from the source image to the target image [3, 35]. However, the image warping method cannot generate new pixels out of the input image, which limits its ability for the target gaze label that is far away from the source image gaze direction.

To overcome this limitation, generative models have been used to synthesize the face or eye images with the target gaze label. He *et al.* [7] introduce a GAN-based approach to generate photo-realistic images with cycle consistency training. To regularize the generated images, they train a gaze estimator to produce the gaze estimation loss between the generated eye image and the ground truth eye image. Xia *et al.* propose controllable gaze redirection methods that explicitly control the gaze with an encoder-decoder structure [33]. STED [41] proposes a VAE architecture following FAZE [22] with the extension to generate the full-face image instead of the eye patch. Nonetheless, they require a pair of labeled samples during training.

However, none of the existing methods of gaze redirection is explicitly 3D aware, even though rigid eyeball rotation is inherently a 3D problem. STED [41] and FAZE [22] introduce an explicit rotation on the learned latent representations while it is a very weak prior nonetheless. The rotation matrix is applied to the 2D feature out of an encoder architecture which is mixed with the eyes and the rest of the face. Essentially, such rotation operation ignores the 3D nature of rigid eyeball rotation and the deformation of the rest of the face. To introduce the actual 3D eyeball rotation, EyeNeRF [16] presents a graphics-based method that fully models the eyeball and the periocular region, yet the focus of EyeNeRF is more on perceptual image quality and photo-realism applications and no result of redirection fidelity is reported in [16]. To train this complex model, EyeNeRF requires a large amount of high-quality data as input, including multiple videos from different camera views accounting for up to 40 minutes. In contrast, GazeNeRF incorporates the 3D awareness with gaze redirection by applying the rotation matrix to the disentangled eyeball feature maps.

2.2. Neural Radiance Fields

Mildenhall *et al.* [17] propose Neural Radiance Fields to represent a static scene with multi-layer perceptrons. NeRF implicitly learns a 3D-aware continuous function and maps the 3D positions and viewing directions to a density and radiance, which is used for generating novel views with volume rendering. Many following works [8, 9, 15, 20, 21, 43] focus on controlling the NeRF-based models to represent the dynamic scenes. Hong *et al.* [8] propose HeadNeRF, a NeRF-based model to generate high-fidelity head images by controlling the shape, expression, and albedo of the

faces with different illumination conditions. They bring the facial parameters from the 3D morphable model (3DMM) into the NeRF-based model and train the HeadNeRF to generate the dynamic head images conditioned on those learnable latent codes. HeadNeRF can synthesize head images with excellent perceptual quality and add the controllability of facial identity and motion. Some similar works to [8] also generate dynamic faces by controlling the shape, expression, and appearance latent codes of the faces in [43]. However, the existing NeRF-based methods of face generation lack gaze control. Different from the previous works, our work focuses on gaze redirection with a NeRF-based model. To control the gaze direction, we train the NeRF-based model conditioned on the gaze label and rigidly rotate the 3D features of the eyes.

3. Method

3.1. Recap: NeRF and HeadNeRF

Neural Radiance Fields, proposed by Mildenhall *et al.* [17], learns an implicit 3D representation that maps a 3D spatial point $\mathbf{x}$ and a view direction $\mathbf{d}$ to an RGB color $\mathbf{c}$ and a volume density σ. It parameterizes this continuous implicit function using an MLP as:

$$h_\theta : (\gamma(\mathbf{x}), \gamma(\mathbf{d})) \rightarrow (\mathbf{c}, \sigma), \tag{1}$$

where θ indicates the network's parameters, and γ denotes a positional encoding function [17, 29] transforming $\mathbf{x}$ and $\mathbf{d}$ into a high-dimensional space.

HeadNeRF [8] is a variant of NeRF for controllable multi-view synthesis and 3D modeling of human faces/heads. Formally, they adjust the MLP as follows:

$$h_\theta : (\gamma(\mathbf{x}), \mathbf{z}_{id}, \mathbf{z}_{exp}, \mathbf{z}_{alb}, \mathbf{z}_{ill}) \rightarrow (\sigma, \mathbf{f}). \tag{2}$$

Similar to [4, 20], HeadNeRF replaced the output RGB value with a high-dimensional feature vector $\mathbf{f}$. $\mathbf{z}_{id}$, $\mathbf{z}_{exp}$, $\mathbf{z}_{alb}$ and $\mathbf{z}_{ill}$ represent the latent codes of the shape, expression, albedo of the face and illumination condition, respectively. The initialization of these latent codes is obtained by fitting the 3D morphable model in [5] to the face.

3.2. GazeNeRF

We aim to bring 3D awareness to the gaze redirection task by leveraging the high-fidelity image generation and implicit 3D consistency powered by NeRF model. To this end, we propose *GazeNeRF*, a NeRF-based model with the two-stream MLPs and explicit 3D rotation on the eye region. Motivated by the fact that the face and eyes are two separate physiological entities that can move independently of each other, we propose to use two MLPs instead of one MLP [8, 17, 43] to separately model the eyes and face only explicitly, supervised by the segmented image patches. To introduce a strong 3D prior to the gaze redirection problem, we directly apply the rotation matrix defined by the target gaze direction on the eye stream due to the rigid movement of the eyeballs. An overview figure of GazeNeRF is presented in Fig. 2.

Two-stream MLPs. In contrast to previous gaze redirection methods that mix the eyes and the face only regions [7, 34, 41], we propose to explicitly disentangle the eyes from the rest of the face with a two-stream MLPs to model two separate radiance fields, h_{θ_e} and $h_{\theta_\text{fw/o}}$ with learnable parameters θ_e and $\theta_\text{fw/o}$ for the eyes and the face only regions respectively. It allows rigid rotation of the two eyeballs along with non-rigid deformation of the periocular areas, such as eyelids and eyebrows. More importantly, it allows for independent control of the transformation or deformation of the two regions.

The two MLPs in GazeNeRF are conditioned on a two-dimensional gaze label consisting of the pitch and yaw angles of the gaze vector in radians, denoted as g. In addition, inspired by [8, 43], GazeNeRF takes 3DMM parameters as input learnable latent codes to control different factors of the image appearance, such as the shape $\mathbf{z}_{sh}$, expression $\mathbf{z}_{ex}$, and texture of the face $\mathbf{z}_{te}$, and the illumination of the image $\mathbf{z}_{il}$. Both MLPs learn a mapping from 3D locations $\gamma(\mathbf{x}) \in \mathbb{R}^{L_\mathbf{x}}$ to a generic feature vector $\mathbf{f} \in \mathbb{R}^{L_\mathbf{f}}$ as:

$$h_{\theta_\text{fw/o}} / h_{\theta_\text{e}} : (\gamma(\mathbf{x}), \mathbf{z}_{sh}, \mathbf{z}_{ex}, \mathbf{z}_{te}, \mathbf{z}_{il}, g) \rightarrow (\sigma, \mathbf{f}). \tag{3}$$

With the output from $h_{\theta_\text{fw/o}}$ and h_{θ_e}, we use volume rendering to obtain two low-resolution volume feature maps $F_\text{fw/o}$ and $F_\text{e} \in \mathbb{R}^{64\times64\times258}$, which are then used to render 2D images. To ensure each stream generates feature maps corresponding to the correct regions, $F_\text{fw/o}$ and F_e are later mapped to the segmented face only and eyes regions, respectively.

3D awareness. Considering that NeRF-based models implicitly learn the 3D volumes of target objects, the feature maps F_e already incorporate the 3D volume information of the eyes. Moreover, previous works incorporate 3D awareness by directly applying the rotation matrix on the 3D volumes [18, 19]. Such rotation operations also have been shown to work even when rotating in 2D feature space for the gaze redirection task [22, 41]. These works, however, apply the rotation to the full face including the eyes ignoring the 3D nature of eyeball rotation and face deformation. Given the feature maps of the eyes F_e, we can apply the rotation matrix calculated by the target gaze direction on it to perform the rigid rotation of the eyeball thanks

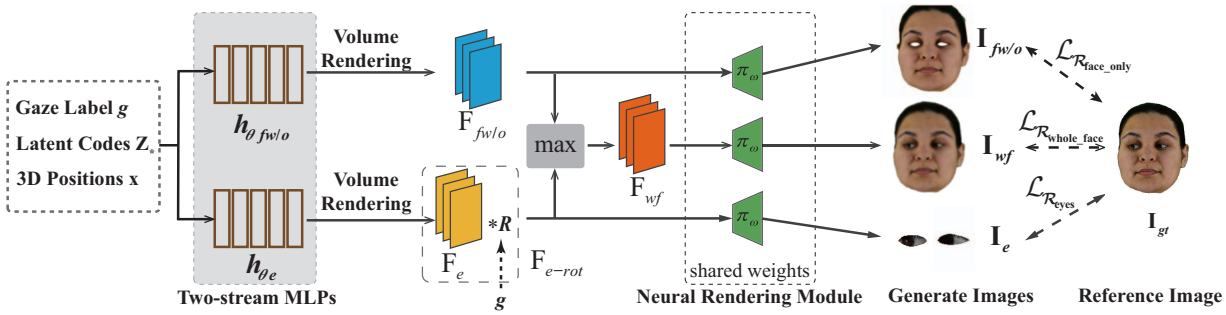

Figure 2. Overview of GazeNeRF pipeline. GazeNeRF trains a two-stream-MLP structure to learn the 3D-aware of the face without eyes feature $F_{\text{fw/o}}$ and the two eyes feature F_e separately via a NeRF-based model. To model the rigid rotation of two eyeballs, we explicitly multiply F_e with a gaze rotation matrix $\mathbf{R}$ to be $F_{\text{e-rot}}$. The $F_{\text{fw/o}}$ and $F_{\text{e-rot}}$ are merged via the max operation to be F_{wf}. All three features are used to render the face without eyes $\mathbf{I}_{\text{fw/o}}$, the eyes $\mathbf{I}_{\text{e}}$, and the completed face images $\mathbf{I}_{\text{fw}}$.

to two-stream MLPs disentanglement. Specifically, we reshape the $F_{\text{e}} \in \mathbb{R}^{64\times64\times258}$ to $F_{\text{e}} \in \mathbb{R}^{64\times64\times86\times3}$ and explicitly apply the following transformation to it as $F_{\text{e-rot}} = \mathbf{R}F_{\text{e}}$, where $\mathbf{R}$ is a 3D rotation matrix computed from the gaze label g [22, 41]. Specifically, we explicitly rotate the feature maps of the eyes F_{e} from the canonical space to $F_{\text{e-rot}}$ in the target space via a rigid rotation.

Merging features. To render the whole face image, we need to combine the feature maps from the two streams, $F_{\text{fw/o}}$ and $F_{\text{e-rot}}$. Similar to how the BlockGAN model combines object features into scene features [19], we apply the element-wise maximum between $F_{\text{fw/o}}$ and $F_{\text{e-rot}}$ to get the merged feature map F_{wf}. This feature map represents the whole face including both the face and eyes.

Rendering images. Finally, to render the final 2D images from the feature maps, a neural rendering module [20] is adopted. It gradually increases the resolution with learnable upsampling layers. The same strategy is used in [8, 26]. We render the images of the face only region $\mathbf{I}_{\text{fw/o}}$ with feature $F_{\text{fw/o}}$, the eyes region $\mathbf{I}_{\text{e}}$ with feature $F_{\text{e-rot}}$, and the whole face $\mathbf{I}_{wf}$ with feature F_{wf}. The weights for the rendering module π_ω are shared for all three images.

Given a reference image, we train GazeNeRF and update the learnable parameters including θ_{e} and $\theta_{\text{fw/o}}$ of two-stream MLPs, four latent codes $\mathbf{z}_*$ and the parameters of the neural rendering module π_ω through the minimization of the following objective function:

$$\mathcal{L}_{Overall} = \lambda_{\mathcal{R}}\mathcal{L}_{\mathcal{R}} + \lambda_{\mathcal{P}}\mathcal{L}_{\mathcal{P}} + \lambda_{\mathcal{F}}\mathcal{L}_{\mathcal{F}} + \lambda_{\mathcal{D}}\mathcal{L}_{\mathcal{D}}, \quad (4)$$

where $\mathcal{L}_{\mathcal{R}}$, $\mathcal{L}_{\mathcal{P}}$, $\mathcal{L}_{\mathcal{F}}$, $\mathcal{L}_{\mathcal{D}}$ represent the reconstruction loss, perceptual loss, functional loss, and disentanglement loss, respectively.

Reconstruction Loss. To generate realistic gaze-redirected images, we apply a reconstruction loss to minimize the pixel-wise distance between a generated image $\mathbf{I}_{wf}$ of the whole face and a target image $\mathbf{I}_{gt}$, which is formulated as:

$$\mathcal{L}_{\mathcal{R}_{\text{whole\_face}}}(\mathbf{I}_{wf}, \mathbf{I}_{gt}) = \frac{1}{|M_{wf} \odot \mathbf{I}_{gt}|} ||M_{wf} \odot (\mathbf{I}_{wf} - \mathbf{I}_{gt})||_1, \quad (5)$$

where M_{wf} is the whole face mask and $\odot$ stands for a pixel-wise Hadamard product operator.

To guarantee that the two streams produce $\mathbf{I}_{\text{e}}$ and $\mathbf{I}_{\text{fw/o}}$ respectively, we apply the similar losses (Eq. (5)) $\mathcal{L}_{\mathcal{R}_{\text{eyes}}}$ and $\mathcal{L}_{\mathcal{R}_{\text{face\_only}}}$ to the individual images generated by the two streams replacing the whole face mask M_{wf} with the eyes mask M_e and the face only mask M_f respectively. These pixel-wise losses associating masks and images ensure the full disentanglement of the eye and the rest of the face. It further enables us to apply the 3D-aware rotation matrix only to the learned features of the eyes. It is also helpful to prevent the generation of blurry eyes by applying the reconstruction loss on the eyes region, since eyes region is smaller than the face only region. Hence the final pixel-level reconstruction loss that we use can be written as:

$$\mathcal{L}_{\mathcal{R}} = \mathcal{L}_{\mathcal{R}_{\text{whole\_face}}} + \mathcal{L}_{\mathcal{R}_{\text{face\_only}}} + \mathcal{L}_{\mathcal{R}_{\text{eyes}}}. \quad (6)$$

Perceptual Loss. The Perceptual loss [11] is designed to measure perceptual and semantic differences between two images with an image classification network ϕ, which has been proved effective in previous works [8, 10]. We employ a perceptual loss to supervise the generated image $\mathbf{I}_{wf}$ to perceptually match with the ground truth image $\mathbf{I}_{gt}$, which is formulated as:

$$\mathcal{L}_{\mathcal{P}_{\text{whole\_face}}} = \sum_i \frac{1}{|\phi_i(\mathbf{I}_{gt})|} ||\phi_i(\mathbf{I}_{wf}) - \phi_i(\mathbf{I}_{gt})||_1, \quad (7)$$

	Gaze↓	Head Pose↓	SSIM↑	PSNR↑	LPIPS↓	FID↓	Identity Similarity↑
STED	16.217	13.153	0.726	**17.530**	0.300	115.020	24.347
HeadNeRF	12.117	4.275	0.720	15.298	0.294	**69.487**	**46.126**
GazeNeRF	**6.944**	**3.470**	**0.733**	15.453	**0.291**	81.816	45.207

Table 1. Comparison of the GazeNeRF to other state-of-the-art methods on the ETH-XGaze dataset in terms of gaze and head redirection errors in degree, redirection image quality (SSIM, PSNR, LPIPS, FID), and identity similarity.

where i denotes the i-th layer of VGG16 network [24] pre-trained on ImageNet [14]. Following the same structure of the reconstruction loss, we compute the perceptual losses, $\mathcal{L}_{\mathcal{P}_{\text{face only}}}$ and $\mathcal{L}_{\mathcal{P}_{\text{eyes}}}$, for the face only and the eyes images from the two streams, $\mathbf{I}_{fw/o}$ and $\mathbf{I}_e$. The total perceptual loss is defined as:

$$\mathcal{L}_{\mathcal{P}} = \mathcal{L}_{\mathcal{P}_{\text{whole\_face}}} + \mathcal{L}_{\mathcal{P}_{\text{face only}}} + \mathcal{L}_{\mathcal{P}_{\text{eyes}}}. \tag{8}$$

Functional Loss. To improve task-specific performance and remove task-relevant inconsistencies between the target image $\mathbf{I}_{gt}$ and the reconstructed image $\mathbf{I}_{wf}$, we adopt the functional loss from STED [41]. We only include the content-consistency loss formulated as:

$$\mathcal{L}_{\mathcal{F}_{\text{content}}}(\mathbf{I}_{wf}, \mathbf{I}_{gt}) = \mathcal{E}_{\text{ang}}(\psi^g(\mathbf{I}_{wf}), \psi^g(\mathbf{I}_{gt})), \tag{9}$$

$$\mathcal{E}_{\text{ang}}(\mathbf{v},\ \hat{\mathbf{v}}) = \arccos \frac{\mathbf{v} \cdot \hat{\mathbf{v}}}{\|\mathbf{v}\|\,\|\hat{\mathbf{v}}\|}, \tag{10}$$

where $\psi^g(*)$ represents the gaze direction estimated by a gaze estimator network, and $\mathcal{E}_{\text{ang}}(*,*)$ represents the angular error function. Our final functional loss is formulated as follows:

$$\mathcal{L}_{\mathcal{F}} = \lambda_{\mathcal{F}_{\text{content}}} \mathcal{L}_{\mathcal{F}_{\text{content}}}. \tag{11}$$

Disentanglement Loss. Inherited from HeadNeRF [8], to disentangle the effect of the latent codes, we minimize the distance between learned latent codes and the initialization to avoid obvious variations as:

$$\mathcal{L}_{\mathcal{D}} = \sum \frac{w_*}{|\mathbf{z}_*^0|} ||\mathbf{z}_* - \mathbf{z}_*^0||^2, \tag{12}$$

where $\mathbf{z}_*^0$ denotes the initial values of the four latent codes obtained from 3DMM parameters, and w_* represents the loss weight.

4. Experiments

To demonstrate the effectiveness of GazeNeRF, we first train GazeNeRF on the ETH-XGaze dataset [37] and compare it to the current state-of-the-art gaze redirection and face generation methods with multiple evaluation metrics. We then conduct cross-dataset evaluations with key evaluation metrics to show the generalization of GazeNeRF. We further analyze the contribution of the various individual components of GazeNeRF to the performance with an ablation study.

4.1. Datasets

ETH-XGaze [37] is a large-scale gaze dataset of high-resolution images with extreme head pose and gaze variation, which was acquired under a multi-view camera system with different illumination conditions. There are 756K frames of 80 subjects in the training set. Each frame is composed of 18 different camera view images. The person-specific test set contains 15 subjects with two hundred images from each subject provided with ground truth gaze labels.
MPIIFaceGaze [40] is an additional dataset for appearance-based gaze estimation based on MPIIGaze dataset [39]. MPIIFaceGaze provides 3000 face images with two-dimensional gaze labels for every 15 subjects.
ColumbiaGaze [27] consists of 5880 high-resolution images taken from 56 subjects. For each subject, the images were acquired with the same five fixed head poses and 21 fixed gaze directions per head pose.
GazeCapture [13] is a large-scale dataset were taken with different poses and under different illumination conditions via crowd-sourcing. During the cross-dataset evaluation, we use its test set only, which contains 150 subjects.

4.2. Implementation details

We employ Adam [12] as our optimizer with $1e^{-4}$ as the learning rate. We use a VGG-based [25] network pre-trained on ImageNet and fine-tune it on the ETH-XGaze training set for the functional loss $L_{\mathcal{F}}$ as the pre-trained gaze estimator. We train another ResNet50 backbone as in [6] on the ETH-XGaze training set that outputs gaze and head pose for evaluation purposes. Finally, we empirically set the total loss coefficients in equation (4) to $\lambda_{\mathcal{R}} = \lambda_{\mathcal{P}} = \lambda_{\mathcal{F}} = \lambda_{\mathcal{D}} = 1$, and the disentanglement weights in equation (12) to $w_{\text{sh}} = w_{\text{te}} = w_{\text{il}} = 1 \times 10^{-3}$ and $w_{\text{ex}} = 1.0$. While $\lambda_{\mathcal{F}_{\text{content}}}$ in equation (11) is set to 1×10^{-3} and is increased by 1×10^{-3} after each epoch.

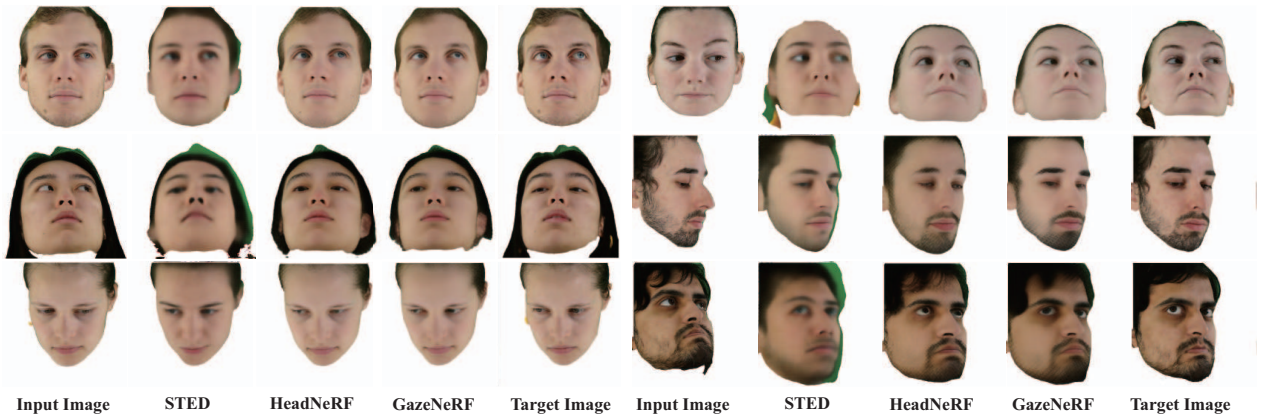

Figure 3. Visualization of generated images from ETH-XGaze with our GazeNeRF, STED and HeadNeRF. All faces are applied with face masks to remove the background. Our GazeNeRF can generate photo-realistic face images with different gaze directions and head poses. STED suffers from losing identity information, and HeadNeRF cannot generate fine-grained eyes (the second row).

	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

Table 2. Comparison of GazeNeRF to other state-of-the-art methods on ColumbiaGaze, MPIIFaceGaze, and GazeCapture datasets in terms of gaze and head redirection errors in degree, LPIPS, and Identity similarity (ID).

4.3. Experimental setup

Pre-processing procedure. We apply the data normalization method [28, 38] and resize the face images into 512×512 pixels. To guarantee that our two-stream MLPs architecture learns to render the face only and eyes regions separately, we utilize face parsing models [44] to generate masks for them. We also adopt the 3D face parametric model from [5] to generate the four latent codes as input into our model. We convert the provided gaze labels from all datasets into pitch-yaw angle labels in the head coordinate system for consistency across subjects and datasets. See the details in the supplementary.

Evaluation metrics. We evaluate all models with various metrics, which can be divided into three different categories: redirection error, redirection image quality, and identity similarity. Similar to STED, redirection error is composed of gaze and head pose angular errors estimated by the ResNet50 [6]-based estimator. These errors are measured with the estimated gaze directions between the redirected images and the corresponding ground truth images. To measure the quality of reconstructed images, we adopt four different metrics, including Structure Similarity Index(SSIM), Peak Signal-to-Noise Ratio(PSNR), Learned Perceptual Image Patch Similarity(LPIPS), and Fréchet Inception Distance(FID). Identity similarity is measured based on the face recognition model from FaceX-Zoo [30]. It measures the differences in identity between the redirected images and ground truth images.

4.4. Comparison to state of the art

To show the superiority of GazeNeRF, we compare GazeNeRF with several previous works in two different experiments: within-dataset evaluation in Tab. 1 and cross-dataset evaluation in Tab. 2. In both experiments, all models are trained with 14.4K images from 10 frames per subject, 18 camera view images per frame, and 80 subjects on the ETH-XGaze training set.

Methods. We compare against the existing state-of-the-art gaze redirection model STED [41], and other variants of the NeRF-based HeadNeRF [8] models that could also be modified to redirect gaze. STED is the current state-of-the-art gaze redirection method applied to full-face images. It performs better than

previous works from He [7], DeepWarp [3] and StarGAN [1]. HeadNeRF is a NeRF-based method that generates high-fidelity face images with 3DMM latent codes controlling different factors of faces. We adapted STED to our setting by increasing the input and target images dimension from 128×128 to 512×512 pixels. To adapt the NeRF-based model for gaze redirection, we concatenated the two-dimensional gaze labels to the original inputs of HeadNeRF directly and conditioned the MLP to learn gaze-related information.

4.4.1 Within-dataset evaluation

Since GazeNeRF and other methods require the gaze label as input, we evaluate their performance on the person-specific test set of ETH-XGaze. There are 15 subjects in the person-specific test set, where 200 images per subject have been annotated with head pose and gaze labels. We randomly pair these 200 labeled images per subject as input and target images. The same pairs of images are evaluated for all models.

Tab. 1 shows the evaluation results of GazeNeRF and other methods. From the table, we can see that GazeNeRF achieves better results than STED and HeadNeRF for most of the error metrics. Especially, GazeNeRF achieves the best performance on the gaze and head redirection as the core criteria of a gaze redirection method. Compared to the HeadNeRF conditioned on the gaze label with single MLP, GazeNeRF applies an explicit rotation to the learned feature maps of the two eyes, which provides better control of gaze direction with smaller gaze error. Although also explicitly applies rotation to the feature maps, STED performs worse than GazeNeRF in terms of gaze and head pose errors. It is because STED utilizes 2D generative model that lacks 3D awareness in its feature maps, and it does not separate the eyes from the face. GazeNeRF achieves similar results as HeadNeRF in terms of the image quality error metrics SSIM, PSNR and LPIPS. This shows that the increase in gaze redirection accuracy does not come at the cost of image fidelity.

We show a qualitative comparison of the various methods in Fig. 3. It clearly shows that GazeNeRF generates photo-realistic face images for variant gaze directions and head poses. STED suffers from the loss of personal identity information in the generated face images, which is quantitatively verified as the 'identity similarity' in Tab. 1. Moreover, STED has difficulty in dealing with extreme head poses (the second row left and the first row right), where the generated faces shift from the target poses. As for HeadNeRF, the feature maps from the single MLP conditioned on gaze labels as inputs alone are not strong enough to control the appearance of eyes with various gaze directions (the last row). Even though most of the results of HeadNeRF can preserve the face identity, the rest of them fail to generate fine-grained eyes (the second row). Compared with these two state-of-the-art methods, GazeNeRF generates better face images with fine-grained eyes, even with extreme head poses (the middle two rows from right).

4.4.2 Cross-dataset evaluation

To evaluate the generalization of GazeNeRF, we conduct a cross-dataset evaluation. For the cross-dataset evaluation, we train the same methods as for the within-dataset evaluation and test on three other datasets, namely ColumbiaGaze, MPIIFaceGaze, and the test set of GazeCapture. Similar to the within-dataset evaluation, we randomly pair the images per subject and fix the pairs for all models. We adopt gaze and head pose angular errors, LPIPS and identity similarity as the evaluation metrics.

The results from Tab. 2 show that GazeNeRF achieves the best performance on the three datasets for most evaluation metrics. As the same as the within-dataset evaluation, GazeNeRF significantly outperforms the other two methods in terms of gaze angular and head pose errors with big margins only except the gaze error in the MPIIFaceGaze dataset. Compared to the HeadNeRF, GazeNeRF achieves better performance in terms of the head pose redirection, although the head rotation operation is the same for both methods. It could be because the two-stream MLPs architecture adds additional ability control for the face only region by separating the face and eye. All three models have similar performance in image quality as in Tab. 1. STED still suffers from the loss of personal identity.

4.5. Ablation study

We analyze GazeNeRF through a number of ablation experiments in Tab. 3. We show the strength of GazeNeRF by comparing it with alternative design choices as listed in the following.

Vanilla-GazeNeRF. We utilize a single MLP and concatenate the two-dimensional gaze label with its other inputs instead of adopting our proposed two-stream MLPs and adding a 3D-aware rotation matrix. Different from the HeadNeRF from Tab. 1, we adopt $\mathcal{L}1$ reconstruction loss used in GazeNeRF instead of $\mathcal{L}2$ photometric loss used in the original HeadNeRF [8] for a fair comparison. During training, all training objectives except functional loss are used.

3D awareness. To verify the individual contributions

	Gaze↓	Head Pose↓	SSIM↑	PSNR↑	LPIPS↓	FID↓	Identity Similarity↑
vanilla-GazeNeRF	11.427	4.581	0.722	15.254	0.291	71.971	47.751
vanilla-GazeNeRF+rotation	9.279	4.458	0.724	15.273	0.296	75.112	47.642
Two-stream	8.609	3.527	0.731	15.431	**0.286**	**69.339**	**48.649**
Two-stream+rotation	8.437	3.563	0.730	15.368	0.296	79.289	48.284
vanilla-GazeNeRF+$L_{\mathcal{F}}$	7.777	4.127	0.729	15.404	0.306	92.201	38.395
GazeNeRF (Two-stream + rotation + $L_{\mathcal{F}}$)	**6.944**	**3.470**	**0.733**	**15.453**	0.291	81.816	45.207

Table 3. Comparison of GazeNeRF to its other variations 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.

of the various components of GazeNeRF, namely its two-stream MLPs architecture and its 3D-aware rotation, we train three more models, *vanilla GazeNeRF+rotation*, *Two-stream* and *Two-stream+rotation* for comparison. Similar to the previous work, STED and FAZE, both of which apply the rotation matrix to the intermediate feature maps of the whole face, *vanilla-GazeNeRF + rotation* applies the gaze rotation matrix to the feature maps of the single MLP. *Two-stream* has a two-stream MLP structure for generating the face without two eyes and two-eye regions separately. Both streams only take the gaze label as input without applying a rotation matrix to the feature maps of two eyes. *Two-stream + rotation* multiplies the 3D-aware feature maps from the two eyes stream with the rotation matrix. Similar to *vanilla-GazeNeRF*, the functional loss is not used for optimizing *Two-stream* and *Two-stream + rotation*.

Functional loss. *Vanilla-GazeNeRF+$L_{\mathcal{F}}$* is trained to verify the power of the functional loss for the gaze redirection task. Compared to *vanilla-GazeNeRF*, the functional loss is used along with the other losses.

From the results shown in Tab. 3, we can see that the baseline model *vanilla-GazeNeRF* performs the worst in terms of gaze error. Comparing *vanilla-GazeNeRF+rotation* with the *vanilla-GazeNeRF*, both the gaze and the head pose angular errors drop. The performance of two angular errors profits from the rotation matrix is applied to the feature maps incorporating the information of the whole face. Moreover, the smaller gaze and head pose angular errors of *Two-stream* are due to the two-stream-MLP structure that separates the whole face into the face only and eyes parts. We can also see that applying rotation matrix to the eyes stream on the basis of *Two-stream* benefits both angular errors of *Two-stream+rotation*. In addition, adding the functional loss $L_{\mathcal{F}}$ as shown with *Vanilla-GazeNeRF+$L_{\mathcal{F}}$* improves the gaze error greatly since it uses an additional gaze estimator to minimize the gaze-relevant inconsistency between the generated and ground-truth images.

Among all ablations, GazeNeRF achieves the best performance in terms of gaze and head pose angular errors by taking advantage of the combination of two-stream-MLP structure, applying a rotation matrix to the eyes stream, and using the function loss $L_{\mathcal{F}}$. As for the image quality, GazeNeRF achieves the best performance regarding SSIM and PSNR score and is comparable to best performances with slight differences for image quality and identify similarities metrics. Again, we emphasize that our goal is not to improve the overall image quality but rather to improve gaze redirection accuracy.

5. Conclusion and Discussion

We propose GazeNeRF, the first method that introduces 3D awareness to the gaze redirection task. By considering the 3D nature of the gaze redirection task itself, GazeNeRF consists of a two-stream MLPs and explicit rotation on the disentangled eye volumes feature. The 3D-aware design endows the advantage of GazeNeRF for the gaze redirection task, which has been proven by the leading performance on multiple datasets and ablation studies. We believe GazeNeRF has great potential for downstream applications with the benefits of 3D awareness. Notwithstanding the above advantages, GazeNeRF shares the same limitation of the group of NeRF models that it takes a long time to train. We leave reducing the burden of training time as our future work.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00608, Artificial intelligence research about multi-modal interactions for empathetic conversations with humans)

References

[1] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 7

[2] 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. 1

[3] Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *European conference on computer vision*, pages 311–326. Springer, 2016. 1, 2, 7

[4] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *International Conference on Learning Representations*, 2022. 3

[5] Yudong Guo, Lin Cai, and Juyong Zhang. 3d face from x: Learning face shape from diverse sources. *IEEE Transactions on Image Processing*, 30:3815–3827, 2021. 3, 6

[6] 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, 6

[7] 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. 1, 2, 3, 7

[8] 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, 3, 4, 5, 6, 7

[9] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12949–12958, 2021. 2

[10] Swati Jindal and Xin Eric Wang. Cuda-ghr: Controllable unsupervised domain adaptation for gaze and head redirection. *arXiv preprint arXiv:2106.10852*, 2021. 1, 4

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 4

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*, 2015. 5

[13] 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

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 5

[15] Verica Lazova, Vladimir Guzov, Kyle Olszewski, Sergey Tulyakov, and Gerard Pons-Moll. Control-nerf: Editable feature volumes for scene rendering and manipulation. *arXiv preprint arXiv:2204.10850*, 2022. 2

[16] 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

[17] 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. 2, 3

[18] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 3

[19] Thu H Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *Advances in Neural Information Processing Systems*, 33:6767–6778, 2020. 3, 4

[20] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 2, 3, 4

[21] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 2

[22] 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. 1, 2, 3, 4

[23] Wooyeong Park, Jeongyun Heo, and Jiyoon Lee. Talking through the eyes: User experience design for eye gaze redirection in live video conferencing. In *International Conference on Human-Computer Interaction*, pages 75–88. Springer, 2021. 1

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recogni-

tion. In *International Conference on Learning Representations*, 2015. 5

[26] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 4

[27] 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, Oct 2013. 5

[28] 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. 6

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[30] 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

[31] Lior Wolf, Ziv Freund, and Shai Avidan. An eye for an eye: A single camera gaze-replacement method. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 817–824. IEEE, 2010. 1

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

[33] 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

[34] Yu Yu, Gang Liu, and Jean-Marc Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11937–11946, 2019. 3

[35] 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, 2

[36] 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, 2

[37] 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

[38] 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, 6

[39] 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, June 2015. 5

[40] 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

[41] 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. 1, 2, 3, 4, 5, 6

[42] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021. 1

[43] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. In *European Conference on Computer Vision*, 2022. 2, 3

[44] zllrunning. Using modified bisenet for face parsing in pytorch. https://github.com/zllrunning/face-parsing.PyTorch. 6