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

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

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

Abstract

Gaze redirection methods aim to generate realistic human face images with controllable eye movement. However, recent methods often struggle with 3D consistency, efficiency, or quality, limiting their practical applications. In this work, we propose RTGaze, a real-time and high-quality gaze redirection method. Our approach learns a gaze-controllable facial representation from face images and gaze prompts, then decodes this representation via neural rendering for gaze redirection. Additionally, we distill face geometric priors from a pretrained 3D portrait generator to enhance generation quality. We evaluate RTGaze both qualitatively and quantitatively, demonstrating state-of-the-art performance in efficiency, redirection accuracy, and image quality across multiple datasets. Our system achieves real-time, 3D-aware gaze redirection with a feedforward network (~0.06 sec/image), making it 800× faster than the previous state-of-the-art 3D-aware methods.

Introduction

Gaze is one of the most important facial features (Cheng et al. 2021; Wang et al. 2023a; Cheng and Lu 2023; Wang et al. 2024) and it conveys human attention and intention in interaction. Gaze redirection involves redirecting the gaze of a face image to a given target direction without changing the identity. It has various applications including virtual reality (Pai et al. 2016; Mania, McNamara, and Polychronakis 2021; Tse et al. 2022; Zheng et al. 2023), digital human (Jack and Schyns 2015; Choi et al. 2025; Wang et al. 2023b) and CG film-making (Yang et al. 2022; Blanz et al. 2004).

Existing gaze redirection methods can be broadly divided into two categories: 2D-based and 3D-based, depending on whether they incorporate 3D representations. 2D-based methods achieve gaze redirection either by warping pixels in the input image (Ganin et al. 2016) or by generating new gaze images through deep generative models such as Generative Adversarial Networks (GANs) (He et al. 2019; Jindal and Wang 2023), encoder-decoder networks (Park et al. 2019), and Variational Autoencoders (VAEs) (Zheng et al. 2020). While effective to some extent, these methods do not capture the inherently 3D nature of gaze redirection, resulting in suboptimal performance under larger head poses.

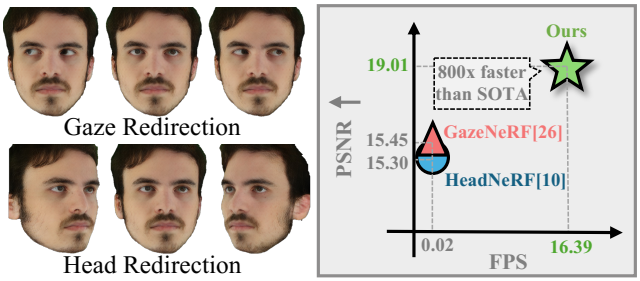

Figure 1: 3D-aware gaze redirection results from our proposed RTGaze, which generates photo-realistic face images under novel gazes and views with good 3D consistency in real time. Compared to the state-of-the-art 3D-aware gaze redirection method GazeNeRF (Ruzzi et al. 2023), which requires approximately one minute during inference, our approach achieves real-time performance at $61ms$ while maintaining superior image quality.

3D-based methods, on the other hand, construct a 3D representation of each input face image using techniques like the neural radiance field (NeRF) (Mildenhall et al. 2020). Once trained, these models can generate a full 3D face and, by adjusting camera poses, produce images with varied head orientations, ensuring strong 3D consistency across a wide range of poses. Among these, GazeNeRF (Ruzzi et al. 2023) is the state-of-the-art, employing two separate multilayer perceptrons (MLPs) to model the radiance fields for the face and eyes independently. GazeNeRF generates novel views using latent codes and gaze labels, but during inference, it requires GAN inversion and updating learnable latent codes before rendering (Hong et al. 2022; Ruzzi et al. 2023), a process that is time-consuming and delays gaze redirection. Balancing 3D consistency with real-time performance, therefore, remains an open challenge in gaze redirection.

In this paper, we introduce RTGaze, a novel method for real-time gaze redirection with 3D awareness, achieving both high-efficiency and high-quality generation. As illustrated in Fig. 2, our method takes images and gaze prompts as inputs, learns a gaze-controllable facial representation. We employ two distinct encoders to separately extract high-frequency and low-frequency features from images. The gaze prompt is injected via a cross-attention mechanism, merging it di-

†Corresponding author.

rectly with the high-frequency features, which are later fused with the low-frequency features. On the other hand, directly optimizing appearance and shape in a lightweight model is challenging, particularly when inferring 3D geometry from a single image (Liu et al. 2023, 2024). To address this, we distill prior knowledge of facial geometry from a pre-trained 3D portrait generator into our module. Our method utilizes NeRF's 3D structure learning and applies a distillation loss to the learned geometric depth images.

Our system achieves *real-time* 3D-aware gaze redirection through a feedforward network, processing each frame in just $\mathbf{61ms}$ on a standard consumer GPU. Extensive quantitative and qualitative evaluations validate our approach, and a series of ablation studies confirm the effectiveness of our design choices. Compared with existing methods (Ruzzi et al. 2023; Hong et al. 2022), RTGaze offers superior image quality and redirection accuracy with a significant boost in inference speed. In summary, our contributions are as follows:

1. We present a real-time 3D-aware gaze redirection model that achieves both high-efficiency and high-quality gaze-controllable image generation. Our method surpasses state-of-the-art methods in inference speed, redirection accuracy, and image quality across different datasets.
2. We propose a novel module for learning gaze-controllable facial representation from face images and a gaze prompt. It consists of two distinctive encoders for extracting facial features and a gaze injection module to effectively incorporate the gaze prompt into the facial representation.
3. We introduce the distillation of 3D face priors from a 3D portrait generation network into our gaze redirection model. By applying a distillation loss on the learned geometric depth images, our approach enhances the overall quality of gaze-redirection synthesis.

Related Work

Gaze redirection methods can generally be categorized into 2D-based methods and 3D-based methods, depending on whether they incorporate 3D representations.

2D-based Gaze Redirection

Deepwarp (Ganin et al. 2016) employs warping maps learned from pairs of eye images with different gaze directions, which requires extensive annotated data. To reduce this reliance on annotated real data, Yu, Liu, and Odobez (2019) incorporate a pretrained gaze estimator with synthetic eye images, further refined by Yu and Odobez (2020) with an unsupervised gaze representation learning network. GAN-based approaches (He et al. 2019) enable gaze redirection by leveraging generative models. FAZE introduces an encoder-decoder framework that encodes eye images into latent vectors, which are then manipulated with rotation matrices to produce synthetic images featuring redirected gaze. ST-ED (Zheng et al. 2020) builds on this by disentangling latent representations to perform both head and gaze redirection for full-face images, achieving highly accurate results. ReDirTrans (Jin et al. 2023) projects edited embeddings back into the original latent space, allowing for attribute replacement with minimal impact on other features and preserving the latent distribution. 2D-based methods often struggle with 3D consistency, as they lack an explicit 3D facial representation.

3D-based Gaze Redirection

3D-based gaze redirection methods offer improved 3D consistency via learned 3D representation. EyeNeRF (Li et al. 2022) combines explicit surface modeling for the eyeball with implicit volumetric representations of surrounding eye structures, enabling high-fidelity gaze redirection with photo-realistic effects using a minimal setup of lights and cameras. GazeNeRF (Ruzzi et al. 2023) employs a two-stream MLP architecture to separately model the face and eye regions via neural radiance fields, allowing for independent manipulation of the eyeball orientation. HeadNeRF (Hong et al. 2022) can be adapted for gaze redirection by integrating gaze labels as conditional inputs. Despite their robust 3D consistency, these methods often require complex, resource-intensive models, limiting their real-time applicability.

Method

Preliminary

This work introduces RTGaze, a novel method for real-time gaze redirection from a single image. Given an input image $\mathbf{I}$ and a target gaze direction $\mathbf{g}$, RTGaze generates a new face image $\hat{\mathbf{I}}$ with the specified gaze. Our pipeline consists of three main components. First, we build a feature extractor to obtain a gaze-controllable facial representation f from $\mathbf{I}$ and $\mathbf{g}$. Then, we feed f into a decoder $\mathcal{G}$ to generate the triplane representation $\mathbf{T}$ (Chan et al. 2022). Finally, we perform neural rendering $\mathcal{N}$ based on $\mathbf{T}$ to synthesize images under the target camera pose. The process could be formulated as:

$$f = \mathcal{F}(\mathbf{I}, \mathbf{g}), \quad \mathbf{T} = \mathcal{G}(f), \quad \hat{I} = \mathcal{N}(\mathbf{T}, \mathbf{c}) \qquad (1)$$

RTGaze employs a hybrid facial representation, leveraging two distinct networks to separately learn high-frequency and low-frequency facial features. To effectively incorporate gaze control, we design a gaze injection module that integrates the gaze input into the hybrid facial representation. Additionally, we distill 3D geometry priors from a pre-trained 3D portrait generation model to enhance the quality of gaze redirection.

Gaze-Controllable Facial Representation

Given an input image $\mathbf{I}$, we construct a hybrid image encoder comprising a high-frequency feature encoder $\mathcal{F}_h$ and a low-frequency feature encoder $\mathcal{F}_l$ (Trevithick et al. 2023). The high-frequency encoder captures fine-grained appearance details, while the low-frequency encoder extracts global geometric information. We feed $\mathbf{I}$ into $\mathcal{F}_h$ and $\mathcal{F}_l$, producing the corresponding high-frequency feature z_h and low-frequency feature z_l, respectively. In detail, We employ the DeepLabV3 network (Chen et al. 2019), pre-trained on ImageNet (Deng et al. 2009), to capture global contextual and semantic information. This information is then processed by a vision transformer encoder (Trevithick et al. 2023), which leverages self-attention mechanisms to refine the extracted global features, producing the final low-frequency feature representation. Simultaneously, a convolutional neural network

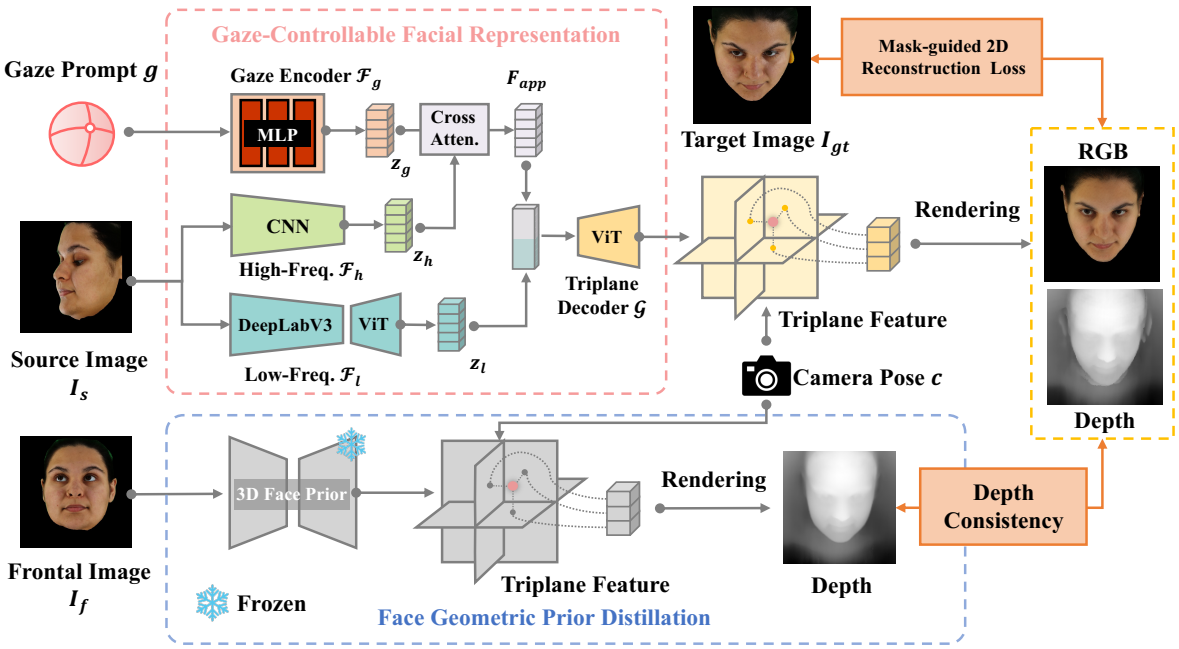

Figure 2: Our model takes three inputs: gaze prompts, source images, and frontal images during training. It consists of a gaze-controllable facial representation learning module and face geometric prior distillation module. First, the model extracts high-frequency and low-frequency features from the source images and injects the gaze prompt into the high-frequency features. The final representation is a fusion of the injected gaze features and the low-frequency features. This combined representation is then fed into a triplane decoder, which generates a 3D face representation in the form of a triplane. This triplane representation is used to render the final gaze-redirected image. The target image provides a mask-guided 2D constraint along the eye region. Additionally, we aim to distill 3D face geometry prior from a pre-trained 3D portrait generation model. We compute depth images from both the pre-trained model and our model, and apply a distillation loss.

(CNN) is used to extract high-frequency features, focusing on fine details within the input image.

Gaze Prompt Injection. Our aim is to inject the gaze prompt into the hybrid facial representation to obtain a gaze-controllable facial representation. The gaze serves as a prompt for generating the target image, ensuring accurate eye appearance. The gaze prompt $\mathbf{g} \in \mathbb{R}^2$ consists of pitch and yaw angles of eyeball rotation. Since gaze redirection primarily affects eye appearance rather than geometry, our strategy first injects the gaze prompt into the high-frequency feature. The injected feature is then fused with the low-frequency feature to produce the final gaze-controllable facial representation. More concretely, given the gaze prompt $\mathbf{g}$, we first embed it using an MLP layer, ensuring that the embedding length matches that of the high-frequency feature. Next, we employ a cross-attention layer (Rombach et al. 2022) to inject the gaze embedding into the high-frequency feature. In this process, the high-frequency feature serves as the query, while the gaze embedding acts as both the key and value. Finally, we integrate the injected high-frequency feature with the low-frequency feature to obtain the final facial representation.

Face Geometric Prior Distillation

Directly optimizing appearance and shape from a single image is challenging, as it lacks sufficient constraints to accurately represent a 3D scene. In this work, we distill face geometric priors from a pre-trained model to enhance the quality of the generated images. Notably, the pre-trained model is typically not designed for gaze redirection.

In detail, we distill face geometric prior from a pre-trained 3D portrait generation model (Trevithick et al. 2023). We find the model has the best performance when the face in the input image is oriented frontally. Therefore, we input a frontal image whose identity and gaze are same as the target image to the pre-trained model and obtain a NeRF representation, *i.e.*, the triplane feature $\mathbf{T}$. The triplane feature enables the rendering of multi-view images and dense depth maps by predicting the color and density of 3D points along camera rays. In this work, we focus solely on depth map generation and apply a distillation loss on the predicted depth, as the synthesized images do not always maintain appearance consistency with the target images. Specifically, we sample 3D points along the camera rays $\mathbf{r}$ under the target camera pose $\mathbf{c}$. For each ray $\mathbf{r}$, the depth value $D(\mathbf{r})$ is computed as:

$$D(\mathbf{r}) = \sum_{i=1}^{N} W_i \left(1 - \exp\left(-\sigma_i \delta_i\right)\right) \mathbf{d}_i, \tag{2}$$

and

$$W_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right). \tag{3}$$

N is the number of samples along the ray, $\mathbf{d}_i$ is the distance between the i-th sample and the camera, σ_i is the density of the i-th sample, and δ_i is the distance between the i-th and $(i+1)$-th samples. The depth map $\mathbf{D}$ under the camera pose $\mathbf{c}$ is obtained by integrating all depth values. We compute the teacher depth map $\mathbf{D^t}$ from the pre-trained model and the student depth map $\mathbf{D^s}$ from our model, applying an L1 loss to enforce depth consistency between them.

$$\mathcal{L}_{\mathcal{D}} = \left\| \mathbf{D}^t - \mathbf{D}^s \right\|_1. \tag{4}$$

Training Objectives

We train the model using a pair of images $\mathbf{I}$ and $\mathbf{I}_t$ with the same identity but different gazes. The 3D face prior is obtained from an additional frontal image with the same identity and gaze as $\mathbf{I}_t$. We optimize our model using the following objective function:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\mathcal{R}} + \beta \cdot \mathcal{L}_{\mathcal{D}} + \gamma \cdot \mathcal{L}_{\mathcal{P}}, \tag{5}$$

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

Mask-Guided 2D Reconstruction Loss. We apply a reconstruction loss that minimizes the differences between the generated image $\hat{\mathbf{I}}$ and the target image $\mathbf{I}_t$ in pixel level. To improve the eye generation quality, we apply an eye region mask to the reconstruction loss and enhance eye region reconstruction by applying a greater loss coefficient specifically to the eye area.

$$\mathcal{L}_{\mathcal{R}} = \alpha_1 \cdot \mathcal{L}_{\mathcal{R}}^{face} + \alpha_2 \cdot \mathcal{L}_{\mathcal{R}}^{eye}, \tag{6}$$

where $\mathcal{L}_{\mathcal{R}}^{face}$ and $\mathcal{L}_{\mathcal{R}}^{eye}$ stand for face region reconstruction loss (excluding the eyes) and eye region reconstruction loss respectively. The $\mathcal{L}_{\mathcal{R}}^{face}$ is formulated as:

$$\mathcal{L}_{\mathcal{R}}^{face} = \frac{1}{|\mathbf{M}_f \odot \mathbf{I}_t|} \left\| \mathbf{M}_f \odot \left(\hat{\mathbf{I}} - \mathbf{I}_t \right) \right\|_1, \tag{7}$$

where $\mathbf{M}_f$ is the face region mask and $\odot$ denotes the pixel-wise Hadamard product operator. $\mathcal{L}_{\mathcal{R}_e}$ is formulated as:

$$\mathcal{L}_{\mathcal{R}}^{eye} = \frac{1}{|\mathbf{M}_e \odot \mathbf{I}_t|} \left\| \mathbf{M}_e \odot \left(\hat{\mathbf{I}} - \mathbf{I}_t \right) \right\|_1, \tag{8}$$

where $\mathbf{M}_e$ is the eye region mask, and $\mathbf{M}_f = \mathbf{1} - \mathbf{M}_e$.

Perceptual Loss. We utilize a perceptual loss (Johnson, Alahi, and Fei-Fei 2016) function to ensure perceptual alignment between the gaze-redirected image $\hat{\mathbf{I}}$ and the target image $\mathbf{I}_t$:

$$\mathcal{L}_{\mathcal{P}} = \sum_i \frac{1}{|\phi_i(\mathbf{I}_t)|} \left\| \phi_i(\hat{\mathbf{I}}) - \phi_i(\mathbf{I}_t) \right\|_1, \tag{9}$$

where ϕ_i denotes the i-th layer of a VGG16 (Simonyan and Zisserman 2014) network pre-trained on ImageNet (Krizhevsky, Sutskever, and Hinton 2012).

During inference, our model only takes a single 2D portrait image and a specified gaze direction as input and produces a gaze-redirected, triplane-based 3D face NeRF. Our model enables photorealistic view synthesis, allowing for highly realistic visualizations from multiple perspectives.

Experiments

Datasets

ETH-XGaze (Zhang et al. 2020) is a large-scale gaze dataset with high-resolution images covering diverse head poses and gaze directions. Collected via a multi-camera setup under various lighting conditions, it includes 756K frames from 80 subjects, each frame captured from 18 angles. A personalized test set comprises 15 subjects, each contributing 200 images with accurate gaze labels. Following GazeNeRF (Ruzzi et al. 2023), we train RTGaze on 14.4K images from 10 frames per subject, with 18 views per frame, using the ETH-XGaze training set. Models are tested on the personalized test set.

MPIIFaceGaze (Zhang et al. 2017) is an extension of the MPIIGaze dataset (Zhang et al. 2015), designed for appearance-based gaze estimation. This dataset comprises 3000 facial images, each annotated with two-dimensional gaze labels for a total of 15 subjects.

ColumbiaGaze (Smith et al. 2013) comprises 5880 high-resolution images collected from 56 subjects. For each subject, the images were captured using five consistent head poses, each linked to 21 fixed gaze directions.

Experimental Setup

Data Preparation. We first normalize the data and resize the face images into a resolution of 512x512 following the method provided in ETH-XGaze (Zhang et al. 2020). Then we process the normalized data following EG3D (Chan et al. 2022) to get the camera pose for each image. To realize the mask-guided 2D constraint, we generate face region masks and eye region masks with face parsing models (Yu et al. 2018). We convert the provided gaze labels into pitch-yaw labels in the head coordinate system for convenience of gaze controlling in 3D space.

Implementation Details. Our model is trained in an end-to-end manner. We employ Adamw (Loshchilov and Hutter 2017) as our optimizer for whole model The learning rates are set to $1e^{-5}$ and $1e^{-5}$ for the encoding part and the rendering part respectively. We train our model with a batch size of 4 for 50 epochs. We empirically set the loss coefficients ($\mathcal{L}_{\mathcal{R}}$, $\mathcal{L}_{\mathcal{D}}$, $\mathcal{L}_{\mathcal{P}}$) in equation (5) to 1, 1, 0.8 respectively. The coefficients ($\mathcal{L}_{\mathcal{R}}^{face}$, $\mathcal{L}_{\mathcal{R}}^{eye}$) of in equation (6) are assigned with 1 and 2 separately. It takes around 18 hours to train the whole model on two NVIDIA A100 GPUs with 40GB memory.

Evaluation Metrics. We evaluate our model with various metrics regarding model efficiency, generated image quality, redirection accuracy, and identity preservation. To evaluate the efficiency of models, we report inference time (including encoding time and rendering time) measured on a single NVIDIA 3090 GPU in the inference stage with an average of 100 samples. To evaluate the quality of generated image, we report four widely used metrics including Structure Similarity Index (SSIM) (Wang et al. 2004), Peak Signal-to-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018), and Fréchet Inception Distance (FID). To evaluate the accuracy of gaze redirection, we report gaze error and head error in degrees. Identity similarity (ID) is assessed using the face recognition model from FaceX-Zoo (Wang et al. 2021). This model evaluates the discrepancies in

Method	3D-based	FID ↓	PSNR ↑	LPIPS ↓	SSIM ↑	Enc Time ↓	Rend Time ↓	Total Time ↓
ST-ED	✗	115.020	17.530	0.300	0.726	-	-	-
HeadNeRF	✓	69.487	15.298	0.294	0.720	$60s$	$0.058s$	$60.058s$
GazeNeRF	✓	81.816	15.453	0.291	**0.733**	$60s$	$0.060s$	$60.060s$
RTGaze (ours)	✓	**38.346**	**19.007**	**0.262**	0.715	**0.026s**	**0.035s**	**0.061s**

Table 1: Quantitative comparisons on ETH-XGaze dataset. We compare our model with state-of-the-art methods regarding image quality (SSIM, PSNR, LPIPS, FID) and inference speed (Encode Time, Render Time, Total Time). For fairness, we only report inference speed metrics for 3D methods. Image quality is evaluated on the personalized test set from ETH-XGaze, while inference speed is averaged over 100 samples on a single NVIDIA 3090 GPU. RTGaze achieves real-time performance, processing each image in 61ms, outperforming other methods in most quality metrics and maintaining competitive SSIM scores.

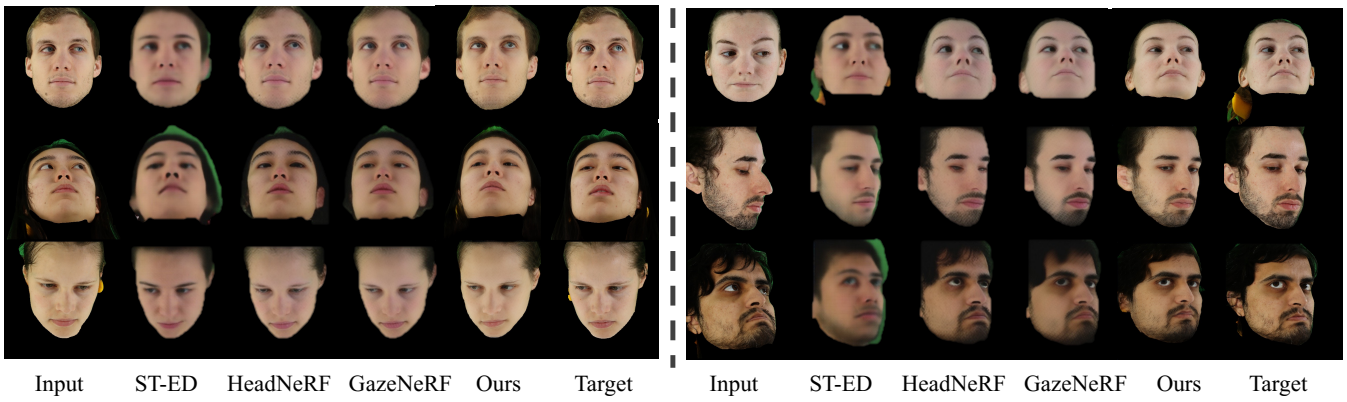

Figure 3: Qualitative comparisons on ETH-XGaze dataset. The background is removed by applying face masks. The images generated from RTGaze are photo-realistic and have extensive details. ST-ED (Zheng et al. 2020) struggles to preserve identity information while retaining the unmasked green background which is not found in 3D-based methods. HeadNeRF (Hong et al. 2022) and GazeNeRF (Ruzzi et al. 2023) suffer from losing facial details.

identity between the redirected images and the corresponding ground truth images.

Baseline Methods. We compare our model against the state-of-the-art gaze redirection methods including 2D-based method ST-ED (Zheng et al. 2020) and 3D-based method GazeNeRF (Ruzzi et al. 2023). ST-ED realizes gaze redirection on full-face images by disentangling latent vectors with a novel self-transforming encoder-decoder architecture. GazeNeRF disentangles eye and face with two-stream MLPs and achieves 3D-aware gaze redirection based on NeRF representation. We also compare our model with HeadNeRF (Hong et al. 2022), a state-of-the-art NeRF-based 3D portrait generation model. It is adapted to gaze redirection task by simply adding two-dimension gaze labels as additional input.

Quantitative Comparison in Image Generation

We show the quantitative results of the comparison with SOTA methods in Table 1. We evaluate our model against other state-of-the-art methods in terms of generated image quality, using widely used metrics including SSIM, PSNR, LPIPS, and FID. To ensure fairness, we only compare our model with other 3D-based methods on inference speed, examining encoding time, rendering time, and total inference time. Image quality is assessed on the personalized test set from the ETH-XGaze dataset, and inference speed is measured by averaging results from 100 samples on a single NVIDIA 3090 GPU.

For inference speed, RTGaze achieves real-time performance in both encoding and rendering stages, with a total processing time of $61ms$ per image. This is attributed to the efficient triplane-based lightweight module distilled from a pre-trained 3D GAN (Trevithick et al. 2023), as well as the avoidance of the inversion process by requiring only a single image as input. In contrast, both HeadNeRF and GazeNeRF are based on the same parametric head model with NeRF representation. Their inputs are parameters of a specific head instead of images. Therefore, they have to conduct an inversion process to update the parameters with the input image, which takes a great amount of time like one minute. They suffer from slower encoding times due to the involved inversion process. Regarding image quality, RTGaze beats the other SOTA methods on most metrics (PSNR, LPIPS, FID) and achieves a comparable result on SSIM. Notably, our model outperforms other methods on FID by a large margin.

Qualitative Comparison in Image Generation

We show the qualitative results of the comparison with SOTA methods in Fig. 3. Following GazeNeRF (Ruzzi et al. 2023), we pair the images with the different gazes from the personalized test set of ETH-XGaze to get the input and target images.

	ETH-XGaze				ColumbiaGaze				MPIIFaceGaze			
	LPIPS↓	**ID↑**	**Gaze↓**	**Head↓**	**LPIPS↓**	**ID↑**	**Gaze↓**	**Head↓**	**LPIPS↓**	**ID↑**	**Gaze↓**	**Head↓**
ST-ED	0.300	24.347	16.217	13.153	0.413	6.384	17.887	14.693	0.288	10.677	14.796	11.893
HeadNeRF	0.294	46.126	12.117	4.275	0.349	23.579	15.250	6.255	0.288	31.877	14.320	9.372
GazeNeRF	0.291	45.207	**6.944**	**3.470**	0.352	23.157	9.464	3.811	0.272	30.981	14.933	7.118
RTGaze	**0.262**	**60.708**	9.047	3.631	**0.249**	**61.765**	**7.625**	**3.326**	**0.251**	**46.098**	**9.409**	**6.444**

Table 2: Comparison of gaze and head redirection on ETH-XGaze, ColumbiaGaze, and MPIIFaceGaze datasets. Lower values are better for Gaze, Head, and LPIPS, while higher values are better for ID. Our model demonstrates a consistent superiority over other SOTA methods across all key metrics on the ColumbiaGaze and MPIIFaceGaze datasets. These results highlight the robustness and adaptability of our model in effectively executing gaze redirection tasks, even when applied to diverse datasets.

Figure 4: Visualization of generated results under novel views and gazes. Our model is able to generate 3D faces with controllable gazes using one single image as input. It can generate photorealistic face images in a large range of head pose and gaze directions. The results under novel views show that our model keeps good 3D consistency in the generation process. Its ability to generate consistent gaze images is also demonstrated by the results under novel gazes. Please zoom in for better visualization.

Injecting into	**FID ↓**	**ID ↑**	**Gaze ↓**	**Head ↓**
Low-Frequency Feature	67.298	38.517	18.973	5.409
High-Frequency Feature	**38.346**	**60.708**	**9.047**	**3.631**

Table 3: Ablation study on gaze prompt injection. The results indicate that injecting gaze prompt into low frequency feature cannot achieve competitive performance.

	FID ↓	**ID ↑**	**Gaze ↓**	**Head ↓**
$\mathcal{L}_{\mathcal{R}}$	101.053	47.251	9.332	4.208
$\mathcal{L}_{\mathcal{R}} + \mathcal{L}_{\mathcal{P}}$	54.682	52.518	10.911	3.700
$\mathcal{L}_{\mathcal{R}} + \mathcal{L}_{\mathcal{P}} + \mathcal{L}_{\mathcal{D}}$	**38.346**	**60.708**	**9.047**	**3.631**

Table 4: Ablation study on loss functions. Note that, $\mathcal{L}_{\mathcal{D}}$ denotes the inclusion of 3D face prior distillation.

Our model takes an image and a target gaze label as inputs, generating a photorealistic gaze-adjusted image.

As shown in Fig. 3, ST-ED (Zheng et al. 2020) suffers from preserving identity information tending to generate similar faces with different inputs. Besides, the results from ST-ED preserve the unmasked green background by mistake which is barely found in 3D-based methods. 2D-based methods only learn a mapping from the input image and gaze label to the target image, while 3D-based methods are trained to build 3D face representations by integrating extensive multi-view information. It demonstrates the robustness of 3D-based methods in handling defective inputs. GazeNeRF (Ruzzi et al. 2023) generates gaze-redirected images whose gazes are aligned with target images, while it struggles to preserve more facial details including facial texture and fine-grained hair. In contrast, our model can generate photorealistic face images with extensive details while maintaining the ability to redirect gaze accurately. Notably, our model works in real time which outperforms all the existing 3D-aware methods.

Gaze Redirection Accuracy Evaluation

We further assess the performance of our model on the ETH-XGaze, ColumbiaGaze and MPIIFaceGaze datasets regarding redirection accuracy and identity preservation. We compare our model with ST-ED (Zheng et al. 2020), HeadNeRF (Hong et al. 2022), and GazeNeRF (Ruzzi et al. 2023) in terms of gaze error, head error, LPIPS, and identity similarity. The detailed results of this evaluation are presented in Table 2. Our model outperforms the competing methods regarding image quality (LPIPS) and identity preservation (ID) and achieves comparable redirection accuracy on ETH-XGaze

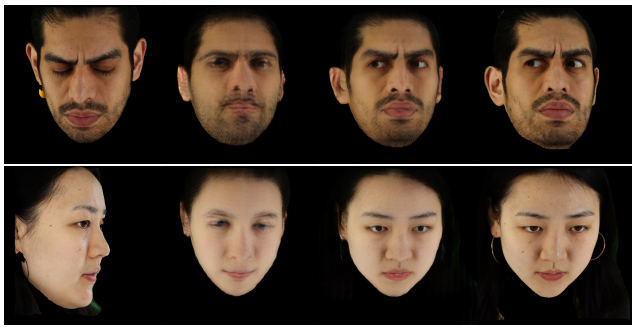

Figure 5: Visualization of ablation on fusion feature choices. The findings suggest that solely relying on low-frequency geometric features leads to blurriness and inaccurate gaze redirection. Conversely, combining high-frequency appearance features with gaze embedding maintains facial structure while enabling efficient gaze redirection.

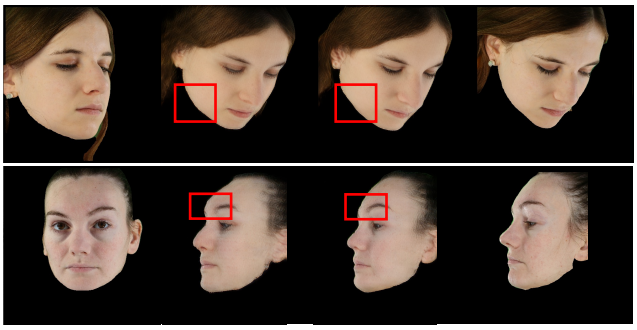

Figure 6: Visualization of ablation on 3D face prior distillation. The results without the distillation show shape distortions, whereas the model utilizing the 3D prior accurately reconstructs the 3D shape.

dataset. Notably, our model consistently outperforms competing methods across all key metrics on ColumbiaGaze and MPIIFaceGaze datasets. These findings underscore the robustness and adaptability of our model in performing gaze redirection tasks, even when applied to diverse datasets.

Face Rendering under Novel Views and Gazes

To showcase the effectiveness of our model in generating 3D-consistent results and achieving consistent gaze redirection, we show the visualization of face rendering under novel views and gazes in Fig. 4. We set the gaze as looking forward during the generation under novel views and interpolate the gaze from left to right under a frontal view in the generation under novel gazes. The results demonstrate that our model can generate face images with strong 3D consistency and enables smooth and coherent gaze interpolation. The good performance relies on our expressive gaze-controllable facial representation and the simple but effective face geometric prior distillation. It is important to note that our model allows the generation of photorealistic face images across a large range of head poses and gaze directions.

Ablation Study

Gaze-Controllable Facial Representation. RTGaze extracts high-frequency and low-frequency features, and the gaze prompt injection module inject gaze prompt embedding with high-frequency feature. We first perform ablation study on the choices of features for gaze prompt injection. The result is shown in Table 3, where injection with low-frequency feature cannot achieve competitive performance. We also try to inject gaze prompt into both low-frequency and high-frequency features, but the training fails to converge due to the difficulty of modifying appearance and geometry simultaneously. We present the generated images with the fusion of low-frequency features in Fig. 5. The results exhibit overall blurriness, and the gaze redirection fails to produce accurate outcomes. When the gaze embedding is fused with low-frequency features, the geometry of the 3D face entangled with the input gaze labels. Ideally, the model should modify only the geometry around the eye region; however, without specific geometric constraints, the model struggles to focus on the eye region alone. Instead, it tends to alter the entire face, leading to noticeable instability in the generated results. In our approach, we fuse the high-frequency features with the gaze embedding while keeping the geometric features unchanged. This enables the model to perform gaze redirection by adjusting only the appearance of the eye region, ensuring a stable face shape throughout the process.

3D Face Prior Distillation. We also conduct an ablation study on the 3D face prior distillation and the proposed loss functions. Results are shown in Table 4. $\mathcal{L}_{\mathcal{D}}$ denotes the inclusion of 3D face prior distillation in our method. The results demonstrate that incorporating 3D face prior distillation and proposed loss functions significantly improves the performance of the model. We also show the qualitative results with and without 3D face prior distillation in Fig. 6. The results without 3D prior exhibit distortions in shape, while the model with 3D prior successfully reconstructs the 3D shape of the input face. This demonstrates that the chosen 3D GAN prior provides effective information on 3D face shape, improving the final generation performance.

Conclusion

We propose RTGaze, a real-time 3D-aware gaze redirection method from a single image. Our model achieves real-time inference by employing an expressive gaze-controllable facial representation to directly fuse gaze prompt into image space and distilling the prior knowledge from 3D GANs into a lightweight module. Benefiting from the gaze-controllable facial representation and the face geometric prior distillation, our model realizes accurate gaze redirection while maintaining superior image quality. With its exceptional real-time performance and high-quality generation, our model holds great potential for numerous downstream applications, particularly in scenarios with high real-time requirements.

Acknowledgments

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). This research was also supported by National Natural Science Foundation of China (Grant No. 62302252). The research utilized the Baskerville Tier 2 HPC service (https://www.baskerville.ac.uk/) funded by the Engineering and Physical Sciences Research Council (EPSRC) and UKRI through the World Class Labs scheme (EP/T022221/1) and the Digital Research Infrastructure programme (EP/W032244/1) operated by Advanced Research Computing at the University of Birmingham. Hengfei Wang was supported by China Scholarship Council Grant No.202006210057.

References

Blanz, V.; Scherbaum, K.; Vetter, T.; and Seidel, H.-P. 2004. Exchanging faces in images. In *Computer Graphics Forum*, volume 23, 669–676. Wiley Online Library.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16123–16133.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2019. Rethinking atrous convolution for semantic image segmentation. arXiv 2017. *arXiv preprint arXiv:1706.05587*, 2: 1.

Cheng, Y.; and Lu, F. 2023. DVGaze: Dual-View Gaze Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 20632–20641.

Cheng, Y.; Wang, H.; Bao, Y.; and Lu, F. 2021. Appearance-based Gaze Estimation With Deep Learning: A Review and Benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Choi, Y.; Wang, H.; Cheng, Y.; Kim, B.; Chang, H. J.; Choi, Y.; and Choi, S.-I. 2025. Roll Your Eyes: Gaze Redirection via Explicit 3D Eyeball Rotation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 10516–10524.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Ganin, Y.; Kononenko, D.; Sungatullina, D.; and Lempitsky, V. 2016. DeepWarp: Photorealistic Image Resynthesis for Gaze Manipulation. In *The European Conference on Computer Vision*.

He, Z.; Spurr, A.; Zhang, X.; and Hilliges, O. 2019. Photo-Realistic Monocular Gaze Redirection Using Generative Adversarial Networks. In *The IEEE International Conference on Computer Vision*.

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

Jack, R. E.; and Schyns, P. G. 2015. The human face as a dynamic tool for social communication. *Current Biology*, 25(14): R621–R634.

Jin, S.; Wang, Z.; Wang, L.; Bi, N.; and Nguyen, T. 2023. Redirtrans: Latent-to-latent translation for gaze and head redirection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5547–5556.

Jindal, S.; and Wang, X. E. 2023. Cuda-ghr: Controllable unsupervised domain adaptation for gaze and head redirection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 467–477.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 694–711. Springer.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Li, G.; Meka, A.; Mueller, F.; Buehler, M. C.; Hilliges, O.; and Beeler, T. 2022. EyeNeRF: a hybrid representation for photorealistic synthesis, animation and relighting of human eyes. *ACM Transactions on Graphics (TOG)*, 41(4): 1–16.

Liu, M.; Xu, C.; Jin, H.; Chen, L.; Varma T, M.; Xu, Z.; and Su, H. 2024. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36.

Liu, R.; Wu, R.; Van Hoorick, B.; Tokmakov, P.; Zakharov, S.; and Vondrick, C. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9298–9309.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mania, K.; McNamara, A.; and Polychronakis, A. 2021. Gaze-aware displays and interaction. In *ACM SIGGRAPH 2021 Courses*, 1–67.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

Pai, Y. S.; Tag, B.; Outram, B.; Vontin, N.; Sugiura, K.; and Kunze, K. 2016. GazeSim: simulating foveated rendering using depth in eye gaze for VR. In *ACM SIGGRAPH 2016 Posters*, 1–2.

Park, S.; Mello, S. D.; Molchanov, P.; Iqbal, U.; Hilliges, O.; and Kautz, J. 2019. Few-Shot Adaptive Gaze Estimation. In *The IEEE International Conference on Computer Vision*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

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

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smith, B. A.; Yin, Q.; Feiner, S. K.; and Nayar, S. K. 2013. Gaze Locking: Passive Eye Contact Detection for Human-Object Interaction. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology*, 271–280. ISBN 9781450322683.

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

Tse, T. H. E.; Zhang, Z.; Kim, K. I.; Leonardis, A.; Zheng, F.; and Chang, H. J. 2022. S 2 Contact: Graph-Based Network for 3D Hand-Object Contact Estimation with Semi-supervised Learning. In *European Conference on Computer Vision*, 568–584. Springer.

Wang, H.; Oh, J. O.; Chang, H. J.; Na, J. H.; Tae, M.; Zhang, Z.; and Choi, S.-I. 2023a. GazeCaps: Gaze Estimation With Self-Attention-Routed Capsules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2668–2676.

Wang, H.; Zhang, Z.; Cheng, Y.; and Chang, H. J. 2023b. High-fidelity eye animatable neural radiance fields for human face. *BMVC*.

Wang, H.; Zhang, Z.; Cheng, Y.; and Chang, H. J. 2024. Textgaze: Gaze-controllable face generation with natural language. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7143–7151.

Wang, J.; Liu, Y.; Hu, Y.; Shi, H.; and Mei, T. 2021. Facexzoo: A pytorch toolbox for face recognition. In *Proceedings of the 29th ACM international conference on multimedia*, 3779–3782.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Yang, G.-W.; Zhou, W.-Y.; Peng, H.-Y.; Liang, D.; Mu, T.-J.; and Hu, S.-M. 2022. Recursive-NeRF: An efficient and dynamically growing NeRF. *IEEE Transactions on Visualization and Computer Graphics*.

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

Yu, Y.; Liu, G.; and Odobez, J.-M. 2019. Improving Few-Shot User-Specific Gaze Adaptation via Gaze Redirection Synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Yu, Y.; and Odobez, J.-M. 2020. Unsupervised Representation Learning for Gaze Estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

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

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2015. Appearance-Based Gaze Estimation in the Wild. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2017. It's written all over your face: Full-face appearance-based gaze estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2299–2308.

Zheng, L.; Wang, C.; Sun, Y.; Dasgupta, E.; Chen, H.; Leonardis, A.; Zhang, W.; and Chang, H. J. 2023. HS-Pose: Hybrid Scope Feature Extraction for Category-level Object Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17163–17173.

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