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

Anonymous CVPR submission

Paper ID 9700

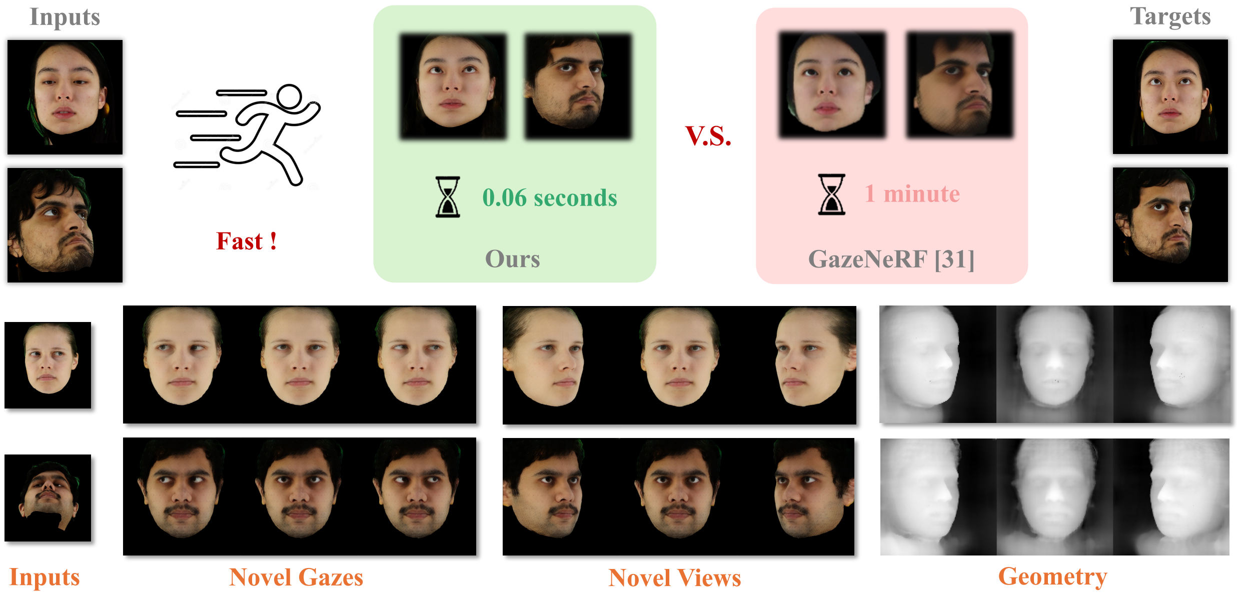

Figure 1. **3D-aware gaze redirection results** from our proposed LiveGaze, which generates photo-realistic face images under novel gazes and views with good 3D consistency in real time. Compared to the state-of-the-art 3D-aware gaze redirection method GazeNeRF [33], which requires approximately one minute during inference, our approach achieves real-time performance at **60ms** while maintaining superior image quality.

Abstract

Gaze redirection methods aim to generate realistic human face images with controllable eye movement. Recent methods usually struggle with good 3D consistency or have limited efficiency and quality, thereby limiting their applications. In this work, we present a real-time and high-quality gaze direction method. Our method leverages recent advancements in real-time radiance fields and fuses gaze with high-frequency features for redirection through cross-attention mechanism. Consequently, we distill a lightweight module from a 3D portrait generator, which provides prior knowledge of face geometry. The final redirected image is attained via differentiable volume rendering. We evaluate LiveGaze qualitatively and quantitatively on ETH-XGaze dataset and it outperforms the state of the art in both efficiency and image quality. Our system achieves real-time 3D-aware gaze redirection with a feedforward network (i.e., ≈ 0.06 sec/image), 1000× faster than the SOTA 3D-aware method. [1]

1. Introduction

Gaze is one of the most important facial features 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 [14, 27, 30, 47, 48], digital human [10, 16, 21, 36] and CG film-making [5, 37, 41]. Besides, it is also used to generate training data for downstream tasks such as gaze estimation as gaze data

[1] Upon publication of this paper, we will release the code and dataset for research use

collection is complex and time-consuming [39].

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 [12] or by generating new gaze images through deep generative models such as Generative Adversarial Networks (GANs) [13, 18], encoder-decoder networks [31], and Variational Autoencoders (VAEs) [49]. 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.

3D-based methods, on the other hand, construct a 3D representation of each input face image using techniques like the neural radiance field (NeRF) [29]. 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 [33] 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 fine-tuning and updating learnable latent codes before rendering [15, 33], 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 tackle the challenge of real-time 3D-aware gaze redirection by distilling a 3D portrait generator into a lightweight module [7, 35] that requires only a single image as input. We introduce a novel method, *LiveGaze*, for real-time gaze redirection with 3D awareness.

As illustrated in Fig. 2, our streamlined module takes an image and gaze label as inputs. Drawing inspiration from Stable Diffusion [32], we incorporate the gaze label by merging it with image features through cross-attention mechanism. The fused features are then encoded into a triplane representation [7] for volumetric rendering, enabling the model to be trained with a reconstruction loss. Direct optimization of appearance and shape in a compact model is challenging, particularly when deriving 3D geometry from a single image [24, 25]. To address this, we distill the prior knowledge of face geometry from a 3D portrait generator into our module, optimizing appearance and adjusting shape as needed.

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 [15, 33], LiveGaze offers superior image quality with a significant boost in inference speed.

In summary, our contributions are as follows:

1. We present a lightweight, real-time 3D-aware gaze redirection module that utilizes a cross-attention-based gaze fusion mechanism, maintaining strong 3D consistency.
2. We introduce the distillation of 3D face priors from 3D GANs into a lightweight module, enhancing 3D face generation quality.
3. Our method, tested on the ETH-XGaze dataset, surpasses state-of-the-art methods in both inference speed and image quality.

2. Related Work

2.1. Gaze Rediretion

Gaze redirection methods can generally be divided into two categories: 2D-based methods and 3D-based methods.

Among 2D-based methods, Deepwarp [12] 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 et al. [44] incorporate a pretrained gaze estimator with synthetic eye images, further refined by Yu and Odobez [43] with an unsupervised gaze representation learning network. GAN-based approaches, such as the one proposed by He et al. [13], 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 [49] builds on this by disentangling latent representations to perform both head and gaze redirection for full-face images, achieving highly accurate results. Expanding on ST-ED, ReDirTrans [17] projects edited embeddings back into the original latent space, allowing for attribute replacement with minimal impact on other features and preserving the latent distribution. While effective, 2D-based methods often struggle with 3D consistency, as they lack an explicit 3D facial representation.

In contrast, 3D-based methods offer improved 3D consistency. EyeNeRF [22] combines explicit surface modeling for the eyeball with implicit volumetric representations of surrounding eye structures, enabling high-fidelity gaze redirection with photorealistic effects using a minimal setup of lights and cameras. GazeNeRF [33] 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. Additionally, HeadNeRF [15] integrates gaze labels as conditional inputs to support gaze redirection. Despite their robust 3D consistency, these methods often require complex, resource-intensive models, limiting their real-time applicability. To address these limitations, our proposed method, LiveGaze, distills 3D facial geometry from 3D GANs into a streamlined, 3D-aware gaze redirection module, achieving real-time performance with minimal computational overhead.

2.2. 3D-Aware One-Shot Portrait Generation

Keeping good 3D consistency is challenging when generating a 3D-aware portrait from one single image. Efficient 3D representations, like neural radiance fields [29] or 3D meshes [23], are commonly used to impose geometric constraints on neural rendering to enhance view consistency. ROME [28] is a mesh-based method that estimates the head mesh as well as the neural texture. Although it achieves good view consistency, the limited resolution of polygonal meshes restricts the neural renderer's ability to produce high-fidelity geometric and appearance details.

HeadNeRF [15] and MofaNeRF [50] support direct control of the head pose of the generated images with NeRF-based parametric model. However, these methods require intensive test-time optimization, limiting their usability in real-time applications, and often struggle to preserve the source identity due to the use of compact latent vectors. EG3D [7] utilizes an advanced triplane-based neural field representation that efficiently encodes the 3D structure and appearance of an avatar's head, enabling detailed and structured 3D modeling. However, it still needs to update the latent codes for one specific head via GAN inversion [1–3].

Live3D [35] distills a lightweight neural network from EG3D to realize 3D portrait generation from one single image in real time. Our work relies on the structure of Live3D and inherits its real-time performance. While working efficiently, it lacks the disentanglement between the appearance and gaze direction and is unable to impose various driving gazes onto the input. To address the limitation, we propose a gaze fusion module to inject gaze as a condition for gaze controlling and train the model with pairs of images with different gazes. In addition, we further propose a distillation model of 3D face priors from the pre-trained Live3D model [35] to supplement shape information.

3. Method

3.1. Preliminary

3D-aware GANs [7, 35] have proven effective in generating highly realistic 3D images by leveraging collections of single-view images. To improve computational efficiency and image quality of 3D GANs, EG3D [7] proposes an efficient triplane-based 3D representation for portrait reconstruction. However, EG3D must do GAN inversion before 3D reconstruction given an image. The GAN inversion makes the whole inference process slow. To address this limitation, we rely on Live3D [35], a state-of-the-art single-image 3D portrait reconstruction model that distills a lightweight module from the pre-trained EG3D model. Live3D takes an image $\mathbf{I}$ as input instead of latent codes. It encodes the input image with two encoders, E_h and E_l, to extract the features with high frequency and low frequency. Then, a ViT-based decoder E_{tri} is used to transfer the concatenated features into the triplane $\mathbf{T}$:

$$\mathbf{T} = \mathrm{E}_{tri}\left(\mathrm{E}_h(\mathbf{I}), \mathrm{E}_l(\mathbf{I})\right). \tag{1}$$

The triplane $\mathbf{T}$ is followed with a volume renderer to render low-resolution images under the given camera pose c. High-resolution images are obtained via a super-resolution module which is the same as EG3D [7].

3.2. Network Architecture

The two encoders, E_h and E_l, are designed to extract high-frequency features and low-frequency features from the input image in Live3D. 3DPE [4] analyzes the two kinds of features by separately disabling the features and visualizing them. They find the features from E_h keep appearance information while losing the shape and the features from E_l retain the geometry while fail to capture the appearance. The human eyeball is close to a sphere and the geometric changes in eye region are tiny. Compared with geometric changes, redirecting gaze usually causes bigger changes in appearance. To this end, we choose to fuse the gaze condition with the high-frequency features from E_h to modify the face appearance and leave face geometry unchanged.

Injecting Gaze as Condition. Given a source image $\mathbf{I}_s$ and a gaze label $\mathbf{g}$, we first use the gaze label encoder E_g to create gaze embeddings whose dimensions are aligned with the high-frequency features $\mathbf{F}_h$. Inspired by Stable Diffusion [32], we fuse the gaze embedding and $\mathbf{F}_h$ via cross-attention mechanism. Specifically, we add a cross-attention layer after E_h to obtain the modified appearance features $\mathbf{F}_{app}$:

$$\mathbf{F}_{app} = \mathrm{CrossAttention}\left(\mathrm{E}_h(\mathbf{I}_s), \mathbf{F}_h\right), \tag{2}$$

In the cross attention, $\mathbf{F}_h$ functions as the query, while $\mathrm{E}_g(\mathbf{g})$ is used as both the key and value. The modified appearance features $\mathbf{F}_{app}$ and geometry feature $\mathbf{F}_{geo}$ are subsequently fed into the decoder E_{tri} to infer the triplane $\mathbf{T}_g$:

$$\mathbf{T}_g = \mathrm{E}_{tri}\left(\mathbf{F}_{app}, \mathbf{F}_{geo}\right), \tag{3}$$

The gaze-redirected image $\mathbf{I}_g$ and the depth map $\mathbf{D}_g$ under the target camera pose $\mathbf{c}$ are generated by a triplane renderer G same as EG3D [7]:

$$\mathbf{I}_g, \mathbf{D}_g = \mathrm{G}\left(\mathbf{T}_g, \mathbf{c}\right). \tag{4}$$

3.3. Distillation of 3D Face Prior

Directly optimizing appearance and shape from one single image within a lightweight model is challenging since it lacks information to constrain a 3D scene. Existing methods generally generate multi-view images with 2D generation models [24, 25] or utilize 3D meshes [8] to provide shape priors. However, the generated multi-view images do not match

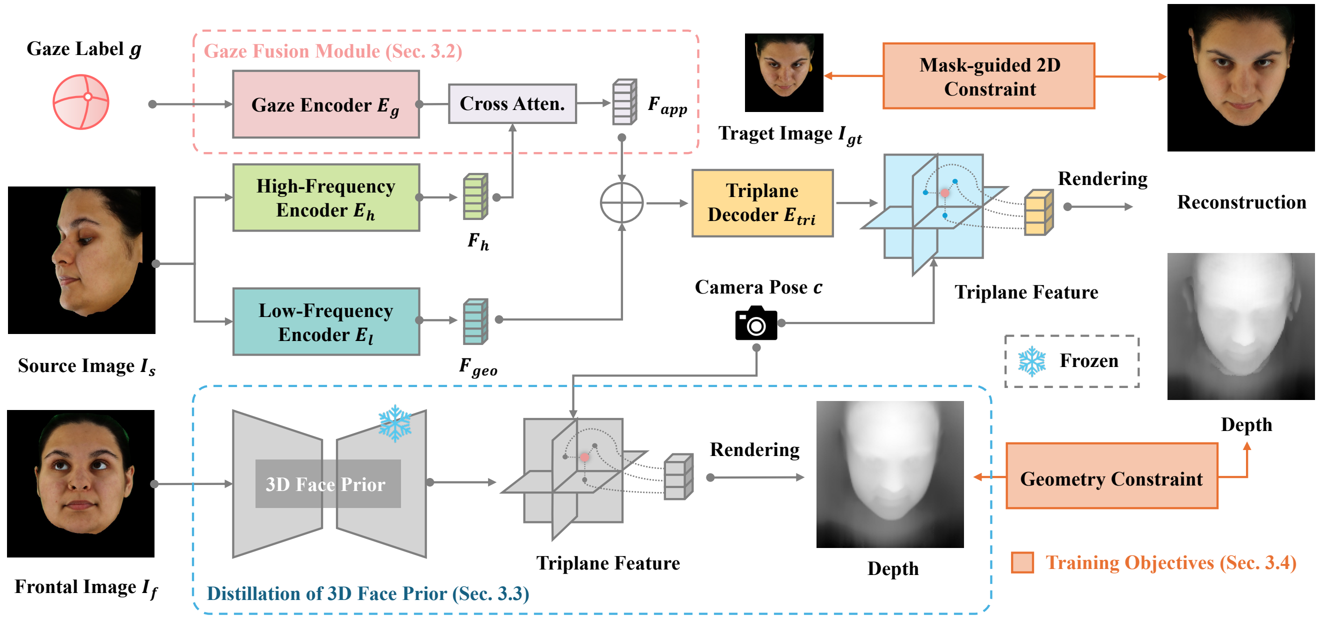

Figure 2. Overview of the LiveGaze pipeline. During training, our model takes three inputs: a gaze label $\mathbf{g}$, a source image $\mathbf{I}_s$, and a frontal image $\mathbf{I}_f$. First, features with different frequencies are extracted from $\mathbf{I}_s$ using two encoders, E_h and E_l. $\mathbf{g}$ is processed by a gaze encoder E_g, and the resulting gaze embedding is fused with the high-frequency features $\mathbf{F}_h$ via cross-attention mechanism to condition the model on gaze direction. The low-frequency features $\mathbf{F}_{geo}$ are then concatenated with the fused features $\mathbf{F}_{app}$, forming the final input to the triplane decoder E_{tri}, which generates a 3D face representation in the form of a triplane. This triplane representation is used to render the final gaze-redirected image and the corresponding depth map under the target camera pose $\mathbf{c}$. The target image $\mathbf{I}_{gt}$ provides a mask-guided 2D constraint along the eye mask. Additionally, geometry constraints are incorporated by distilling prior knowledge of 3D face geometry from a pre-trained 3D GANs, which takes $\mathbf{I}_f$ as input and outputs the reference depth map. During inference, our model is capable of performing real-time 3D-aware gaze redirection using just a single source image and gaze labels.

each other in most cases and result in poor reconstruction performance. While mesh priors are generally reasonable, face tracking is time-consuming. The mesh priors also have the low resolution issue which limits the performance of reconstructing details and most meshs obtained via face tracking only cover face region regardless of the hair region. Instead, we utilize the pre-trained Live3D [35], a state-of-the-art 3D portrait generation model, to generate 3D face prior. We find Live3D has the best performance when the face in the input image is oriented frontally. Specifically, we input a frontal image $\mathbf{I}_f$ whose identity and gaze are same as the target image $\mathbf{I}_{gt}$ to a pre-trained Live3D model and obtain the depth map $\mathbf{D}_{gt}$ under the target camera pose $\mathbf{c}$:

$$\mathbf{D}_{gt} = \mathrm{Live3D}\left(\mathbf{I}_f, \mathbf{c}\right), \tag{5}$$

Thanks to its good efficiency and the ability to generate high-resolution depth maps, we are able to obtain the detailed 3D face prior in real time. We formulate a geometry constraint to distill prior knowledge of 3D face shapes into our lightweight module. The 3D face prior guarantees the geometry quality of learned 3D faces and alleviates the complexity in training the model.

3.4. Training Objectives

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

$$\mathcal{L} = \lambda_{\mathcal{R}}\mathcal{L}_{\mathcal{R}} + \lambda_{\mathcal{D}}\mathcal{L}_{\mathcal{D}} + \lambda_{\mathcal{P}}\mathcal{L}_{\mathcal{P}}, \tag{6}$$

where $\mathcal{L}_{\mathcal{R}}$, $\mathcal{L}_{\mathcal{D}}$, $\mathcal{L}_{\mathcal{P}}$ represent the reconstruction loss, depth loss, and perceptual loss, respectively. As illustrated in Fig. 2, the reconstruction loss defines our mask-guided 2D constraint and the depth loss builds our geometry constraint.

Mask-Guided 2D Reconstruction Loss. To achieve realistic gaze redirection, we introduce a 2D reconstruction loss that minimizes the differences between the generated image $\mathbf{I}_g$ and the target image $\mathbf{I}_{gt}$ in pixel level. To improve the generation quality of eyes, we apply a face region mask and an eye region mask to the reconstruction loss and enhance eye region reconstruction by setting the loss in eye region with a larger weight:

$$\mathcal{L}_{\mathcal{R}} = \lambda_{\mathcal{R}_f}\mathcal{L}_{\mathcal{R}_f} + \lambda_{\mathcal{R}_e}\mathcal{L}_{\mathcal{R}_e}, \tag{7}$$

where $\mathcal{L}_{\mathcal{R}_f}$ and $\mathcal{L}_{\mathcal{R}_e}$ stand for face region reconstruction loss and eye region reconstruction loss respectively. We set $\lambda_{\mathcal{R}_e} > \lambda_{\mathcal{R}_f}$ in our case.

The face region reconstruction loss $\mathcal{L}_{\mathcal{R}_f}$ is formulated as:

$$\mathcal{L}_{\mathcal{R}_f} = \frac{1}{|M_f \odot \mathbf{I}_{gt}|} \left\| M_f \odot (\mathbf{I}_g - \mathbf{I}_{gt}) \right\|_1 , \tag{8}$$

where M_f is the face region mask and $\odot$ denotes the pixel-wise Hadamard product operator. And the eye region reconstruction loss $\mathcal{L}_{\mathcal{R}_e}$ is formulated as:

$$\mathcal{L}_{\mathcal{R}_e} = \frac{1}{|M_e \odot \mathbf{I}_{gt}|} \left\| M_e \odot (\mathbf{I}_g - \mathbf{I}_{gt}) \right\|_1 , \tag{9}$$

where M_e is the eye region mask.

Depth Loss. To distill the 3D face prior into our lightweight module, we utilize the depth map $\mathbf{D}_{gt}$ from pre-trained Live3D to supervise the generated depth map $\mathbf{D}_g$ of the redirected face:

$$\mathcal{L}_{\mathcal{D}} = \left\| \mathbf{D}_g - \mathbf{D}_{gt} \right\|_1 . \tag{10}$$

Perceptual Loss. Perceptual loss [19] is to assess the quality of generated images by comparing high-level feature representations rather than pixel-level differences. It leverages a pre-trained network (such as VGG [34]) to capture the semantic and structural similarity between generated and target images. The models trained with perceptual loss tend to produce more visually realistic and detailed outputs by focusing on perceptual similarity, which better aligns with human visual perception. It has been widely used in image synthesis [15, 18] and approved as effective. We utilize a perceptual loss function to ensure perceptual alignment between the gaze-redirected image $\mathbf{I}_g$ and the target image $\mathbf{I}_{gt}$:

$$\mathcal{L}_{\mathcal{P}} = \sum_i \frac{1}{|\phi_i(\mathbf{I}_{gt})|} \left\| \phi_i(\mathbf{I}_g) - \phi_i(\mathbf{I}_{gt}) \right\|_1 , \tag{11}$$

where ϕ_i denotes the i-th layer of a VGG16 [34] network pre-trained on ImageNet [20].

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.

4. Experiments

To demonstrate the effectiveness of LiveGaze, we train our model on the ETH-XGaze [46] dataset and compare it to the state-of-the-art methods regarding efficiency and image quality both qualitatively and quantitatively. Then we show the results of novel view synthesis and gaze redirection. Finally, we validate our design choices via ablation studies.

4.1. Experimental Setup

Dataset. We train our model on the ETH-XGaze dataset which is widely used in gaze redirection task [33, 39] and gaze estimation task [9, 38]. ETH-XGaze is a large-scale gaze dataset comprising high-resolution images that capture a wide range of head poses and gaze directions. It was collected using a multi-camera setup under various lighting conditions to enhance diversity. The training set includes 756K frames across 80 subjects, with each frame featuring images from 18 distinct camera angles. Additionally, a personalized test set contains 15 subjects, with each providing 200 images along with accurate ground truth gaze labels for evaluation. Following GazeNeRF [33], we train LiveGaze with 14.4K images from 10 frames per subject, 18 images with different views per frame, and 80 subjects on the ETH-XGaze training set. We test all the models on the personalized test set.

Data Preparation. We first normalize the data and resize the face images into a resolution of 512x512 following the method provided in ETH-XGaze [46]. Then we process the normalized data following EG3D [7] 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 [42]. 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 [26] 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 (6) to 1, 1, 0.8 respectively. The coefficients ($\mathcal{L}_{\mathcal{R}_f}$ and $\mathcal{L}_{\mathcal{R}_e}$) of in equation (7) 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 and generated image quality. To evaluate the efficiency of models, we report the 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) [40], Peak Signal-to-Noise Ratio(PSNR), Learned Perceptual Image Patch Similarity(LPIPS) [45], and Fréchet Inception Distance(FID).

4.2. Efficient 3D-Aware Gaze Redirection

Baseline Methods. We compare our model against the state-of-the-art gaze redirection methods including 2D-based method ST-ED [49] and 3D-based method GazeNeRF [33]. ST-ED realizes gaze redirection on full-face images by disentangling latent vectors with a novel self-transforming

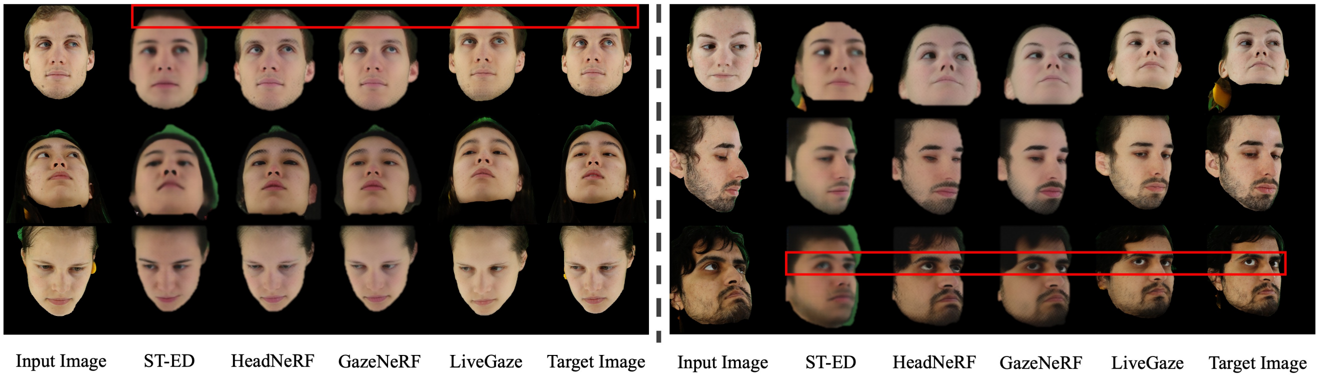

Figure 3. Qualitative comparisons. We conduct the comparison on ETH-XGaze dataset [46]. The background is removed by applying face masks. The images generated from our LiveGaze are photo-realistic and have extensive details. ST-ED [49] encounters challenges in preserving identity information while retaining the unmasked green background which is not found in 3D-based methods. HeadNeRF [15] and GazeNeRF [33] suffer from losing facial details.

Table 1. Quantitative comparisons. We compare our LiveGaze model with other state-of-the-art methods based on image quality (SSIM, PSNR, LPIPS, FID) and inference speed (Encode Time, Render Time, Total Time). For fairness, we report inference speed metrics only 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. LiveGaze achieves real-time performance, processing each image in just 61ms, outperforming other methods in most quality metrics and maintaining competitive SSIM scores. In contrast, HeadNeRF and GazeNeRF experience slower performance due to increased encoding times caused by inversion.

Method	**3D-based**	**SSIM** ↑	**PSNR** ↑	**LPIPS** ↓	**FID** ↓	**Enc Time** ↓	**Rend Time** ↓	**Total Time** ↓
ST-ED [49]	✗	0.726	17.530	0.300	115.020	-	-	-
HeadNeRF [15]	✓	0.720	15.298	0.294	69.487	60s	0.058s	60.058s
GazeNeRF [33]	✓	**0.733**	15.453	0.291	81.816	60s	0.060s	60.060s
Ours	✓	0.715	**19.007**	**0.272**	**38.346**	**0.026s**	**0.035s**	**0.061s**

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, 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.

Qualitative Results. We show the qualitative results of the comparison with SOTA methods in Fig. 3. Following GazeNeRF [33], we pair the images with the different gazes from the personalized test set of ETH-XGaze to get the input images and target images. Our model takes one single image as well as a target gaze label as input and generates a photorealistic gaze-redirected image. As shown in Fig. 3, ST-ED [49] 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 (HeadNeRF, GazeNeRF, and our LiveGaze). 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 explains the robustness of 3D-based methods in handling defective inputs, which verifies our choice of NeRF-based architecture. Even though HeadNeRF [15] generates face images with the correct identity, it fails to redirect the gaze accurately and loses details in eye regions. GazeNeRF [33] 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 as shown in the red box on the left. 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.

Quantitative Results. 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 compare inference speed with other 3D-based methods, examining encoding time, rendering time, and total 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, LiveGaze achieves real-time performance in both encoding and rendering stages, with a total

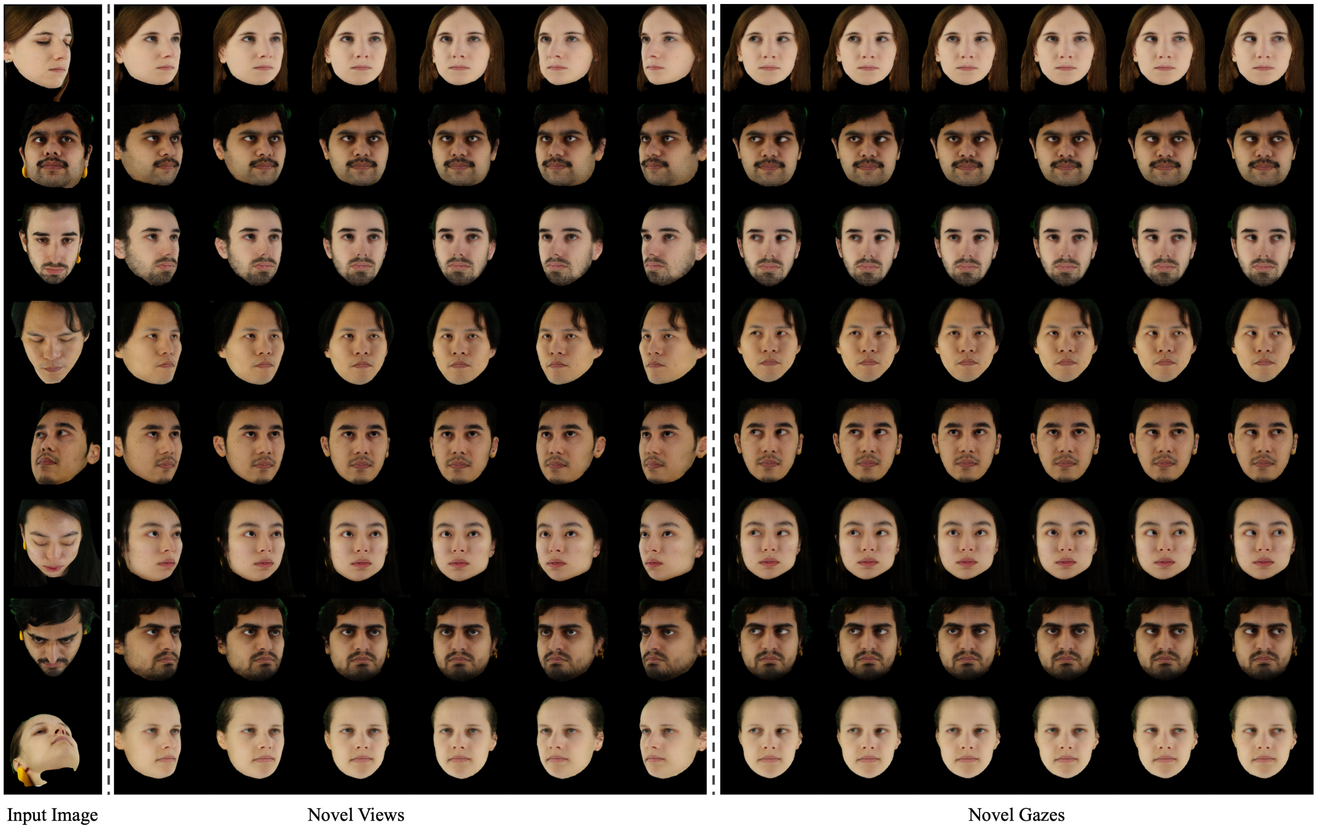

Figure 4. Visualization of generated results under novel views and gazes. Our model is able to generate 3D faces with controllable gazes using one single image as input. It can generate photorealistic face images in a large range of head pose and gaze directions. The results under novel views show that 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.

processing time of $61ms$ per image. This is attributed to the efficient triplane-based lightweight module distilled from a pre-trained 3D GAN [35], 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, LiveGaze 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.

4.3. 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 triplane-based 3D face representation and the simple but effective gaze fusion module. 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. Please see the supplementary materials for more details.

4.4. Ablation Study

Ablation on Feature Choices. LiveGaze separates high-frequency features (related to appearance) and low-frequency features (related to geometry) before mapping them to the triplane representation. We perform an ablation study on these features, and the results are shown in Fig. 5. Images generated using only low-frequency features appear blurry across the entire image, with gaze redirection failing to achieve accurate results. When the gaze embedding is fused with low-frequency features, the geometry of the 3D face becomes entangled with the input gaze labels. Ideally,

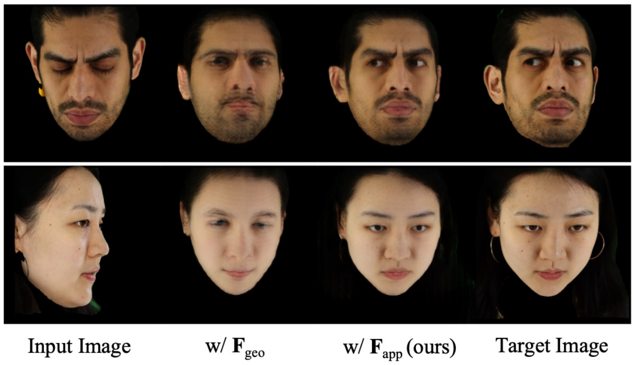

Figure 5. Ablation on feature choices. The results highlight that using only low-frequency geometric features results in blurry images and inaccurate gaze redirection. Fusing low-frequency features with gaze embedding also causes unintended changes across the entire face, making it challenging to isolate modifications to the eye region alone. In contrast, incorporating high-frequency appearance features with gaze embedding preserves stable face geometry, enabling effective gaze redirection.

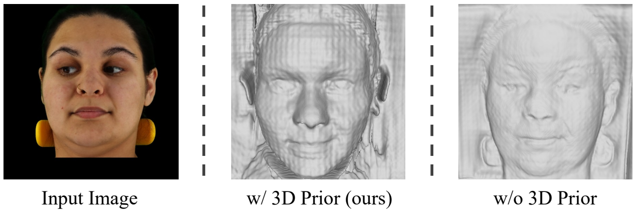

Figure 6. Ablation on 3D face prior distillation. We generate 3D meshes with and without 3D prior. The result shows that the model with 3D prior reconstructs the 3D shape of the input face successfully while the model without without 3D prior fails to capture the depth information of human head by generating a flat face mesh. This result demonstrates that the chosen 3D GAN prior can provide effective information on 3D face shape which improves the final generation performance.

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.

Ablation on 3D Face Prior Distillation. We conduct an ablation study on 3D face prior distillation and show the reconstructed 3D meshes in Fig. 6. Our model, leveraging 3D priors, clearly reconstructs 3D face shapes with high fidelity. In contrast, the mesh generated without 3D priors appears flat across the face, lacking significant depth information.

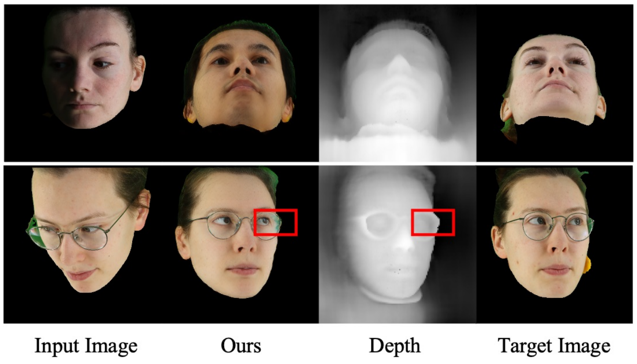

Figure 7. Failure cases under challenging conditions. LiveGaze demonstrates robust gaze redirection but struggles with low lighting and faces wearing glasses. When processing the images in a dark environment, the model may incorrectly interpret the subject's identity. Additionally, glasses are challenging to reconstruct accurately, often interfering with eye region generation. Nonetheless, our model consistently redirects gaze correctly, underscoring its effective gaze-redirecting capability.

4.5. Limitations

While LiveGaze is capable of generating photorealistic gaze-redirected face images in real time, it faces challenges under certain conditions, such as varying illumination and the presence of glasses. Failure cases are illustrated in Fig. 7. In low-light environments, our model may mistakenly perceive faces as belonging to different subjects with darker facial tones. Additionally, its performance declines when processing faces with glasses. As shown in the second row of Fig. 7, the model struggles to reconstruct glasses correctly, and their presence interferes with the generation of the eye region. Despite these limitations, it is worth noting that our model successfully redirects gaze correctly in both scenarios, underscoring its robust gaze redirection capability.

5. Conclusion

We propose LiveGaze, a real-time 3D-aware gaze redirection method from a single image. Our model achieves real-time inference by employing an efficient triplane-based NeRF architecture and distilling the prior knowledge from 3D GANs into a lightweight module. Benefiting from the simple but effective gaze fusion module and the dedicated choice of fused features, 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. While LiveGaze offers significant advantages, its performance declines under challenging conditions like varying lighting and the presence of glasses. Addressing these challenges is reserved for future work.

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

Supplementary Material

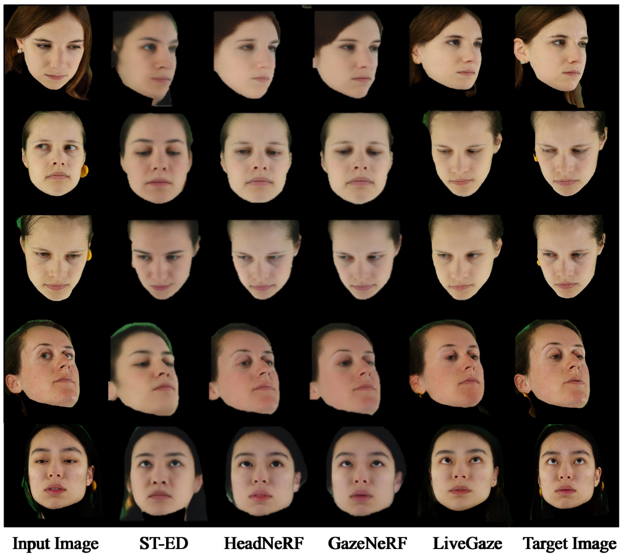

Figure 8. Additional visualization of generated images from ETH-XGaze with our LiveGaze, ST-ED, HeadNeRF, and GazeNeRF. The background is eliminated using face masks. Our model can generate photo-realistic images with extensive details. In contrast, ST-ED struggles to preserve identity information. HeadNeRF and GazeNeRF face challenges in maintaining facial details.

6. Details of Data Pre-processing and Training

The resolution of raw images in ETH-XGaze [46] is 6Kx4K. We first normalize the raw images using the method in [46] and get the normalized head poses and gaze directions. The normalized distance between the camera and the center of the face is fixed at 950mm and the focal length in the normalized camera projection matrice is set to 1600. To align the data format with Live3D [35] and EG3D [7], we resize the normalized images to 512x512 and estimate camera poses using the model in [11]. To apply our mask-guided 2D constraint, we use the face parsing model [42] to segment the whole and the eye region. We use the detected landmarks [6] to do the segmentation when the face parsing model does not work when processing some challenging images.

The personalized test set in ETH-XGaze includes 200 labeled images for each subject. We split the personalized test set into an input group and a target group following GazeNeRF [33]. The input group contains 100 images for each subject and the target group includes the other 100 images. We train our model with 80 subjects in the train set of ETH-XGaze first and then finetune the model with images from the input group for 10 epochs. We generate the images in the target group during evaluation.

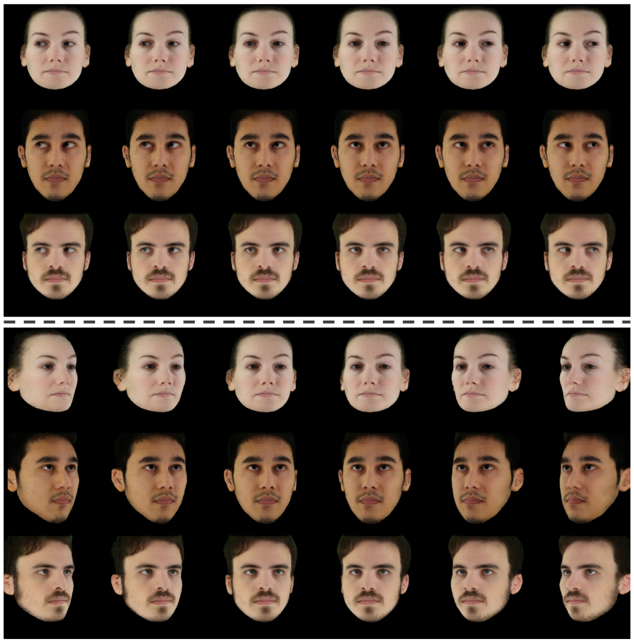

Figure 9. Additional visualization of generated images under novel gazes and novel views. The upper part is the generated images under novel gazes which are from left to right. The lower part is the generated results under novel views. Our model generates 3D faces with controllable gazes from a single image, producing photorealistic results across diverse head poses and gaze directions while maintaining 3D and gaze consistency.

7. Additional Qualitative Results

In Fig. 8, we show additional qualitative results comparing our model to the SOTA methods. All the models are evaluated on the personalized test set of the ETH-XGaze dataset. Our model generates images with superior quality while ST-ED [49] encounters difficulties in preserving identity information and both HeadNeRF [15] and GazeNeRF [33] struggle to retain facial details.

In Fig. 9, we show additional qualitative results of generation under novel gazes and novel views. Our model can generate 3D faces with controllable gazes from a single input image. It produces photorealistic face images across a wide range of head poses and gaze directions. The results under novel viewpoints demonstrate the model's strong 3D consistency throughout the generation process. Additionally, its capability to produce consistent gaze images is validated by the results under novel gaze directions.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4432–4441, 2019. 3

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8296–8305, 2020.

[3] Qingyan Bai, Yinghao Xu, Jiapeng Zhu, Weihao Xia, Yujiu Yang, and Yujun Shen. High-fidelity gan inversion with padding space. In *European Conference on Computer Vision*, pages 36–53. Springer, 2022. 3

[4] Qingyan Bai, Zifan Shi, Yinghao Xu, Hao Ouyang, Qiuyu Wang, Ceyuan Yang, Xuan Wang, Gordon Wetzstein, Yujun Shen, and Qifeng Chen. Real-time 3d-aware portrait editing from a single image. In *European Conference on Computer Vision*, pages 344–362. Springer, 2025. 3

[5] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging faces in images. In *Computer Graphics Forum*, pages 669–676. Wiley Online Library, 2004. 1

[6] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, pages 1021–1030, 2017. 1

[7] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022. 2, 3, 5, 1

[8] Yujin Chen, Yinyu Nie, Benjamin Ummenhofer, Reiner Birkl, Michael Paulitsch, Matthias Müller, and Matthias Nießner. Mesh2nerf: Direct mesh supervision for neural radiance field representation and generation. In *European Conference on Computer Vision*, pages 173–191. Springer, 2025. 3

[9] Yihua Cheng and Feng Lu. Dvgaze: Dual-view gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20632–20641, 2023. 5

[10] H Onan Demirel and Vincent G Duffy. Applications of digital human modeling in industry. In *Digital Human Modeling: First International Conference on Digital Human Modeling, ICDHM 2007, Held as Part of HCI International 2007, Beijing, China, July 22-27, 2007. Proceedings 1*, pages 824–832. Springer, 2007. 1

[11] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 1

[12] Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *The European Conference on Computer Vision*, 2016. 2

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

[14] Anders Henrysson, Mark Billinghurst, and Mark Ollila. Face to face collaborative ar on mobile phones. In *Fourth ieee and acm international symposium on mixed and augmented reality (ismar'05)*, pages 80–89. IEEE, 2005. 1

[15] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 2, 3, 5, 6, 1

[16] Rachael E Jack and Philippe G Schyns. The human face as a dynamic tool for social communication. *Current Biology*, 25 (14):R621–R634, 2015. 1

[17] Shiwei Jin, Zhen Wang, Lei Wang, Ning Bi, and Truong Nguyen. Redirtrans: Latent-to-latent translation for gaze and head redirection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5547–5556, 2023. 2

[18] Swati Jindal and Xin Eric Wang. Cuda-ghr: Controllable unsupervised domain adaptation for gaze and head redirection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 467–477, 2023. 2, 5

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 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*, pages 694–711. Springer, 2016. 5

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 5

[21] Tommer Leyvand, Daniel Cohen-Or, Gideon Dror, and Dani Lischinski. Digital face beautification. In *ACM Siggraph 2006 Sketches*, pages 169–es. 2006. 1

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

[23] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 3

[24] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[25] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2, 3

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[27] Katerina Mania, Ann McNamara, and Andreas Polychronakis. Gaze-aware displays and interaction. In *ACM SIGGRAPH 2021 Courses*, pages 1–67, 2021. 1

[28] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022. 3

[29] 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. In *ECCV*, 2020. 2, 3

[30] Yun Suen Pai, Benjamin Tag, Benjamin Outram, Noriyasu Vontin, Kazunori Sugiura, and Kai Kunze. Gazesim: simulating foveated rendering using depth in eye gaze for vr. In *ACM SIGGRAPH 2016 Posters*, pages 1–2, 2016. 1

[31] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *The IEEE International Conference on Computer Vision*, 2019. 2

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 3

[33] Alessandro Ruzzi, Xiangwei Shi, Xi Wang, Gengyan Li, Shalini De Mello, Hyung Jin Chang, Xucong Zhang, and Otmar Hilliges. Gazenerf: 3d-aware gaze redirection with neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 2, 5, 6

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[35] Alex Trevithick, Matthew Chan, Michael Stengel, Eric Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Ravi Ramamoorthi, and Koki Nagano. Real-time radiance fields for single-image portrait view synthesis. 2023. 2, 3, 4, 7, 1

[36] Tze Ho Elden Tse, Zhongqun Zhang, Kwang In Kim, Ales Leonardis, Feng Zheng, and Hyung Jin Chang. S 2 contact: Graph-based network for 3d hand-object contact estimation with semi-supervised learning. In *European Conference on Computer Vision*, pages 568–584. Springer, 2022. 1

[37] P Vanezis, RW Blowes, AD Linney, AC Tan, R Richards, and R Neave. Application of 3-d computer graphics for facial reconstruction and comparison with sculpting techniques. *Forensic science international*, 42(1-2):69–84, 1989. 1

[38] Hengfei Wang, Jun O Oh, Hyung Jin Chang, Jin Hee Na, Minwoo Tae, Zhongqun Zhang, and Sang-Il Choi. Gazecaps: Gaze estimation with self-attention-routed capsules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2668–2676, 2023. 5

[39] Hengfei Wang, Zhongqun Zhang, Yihua Cheng, and Hyung Jin Chang. High-fidelity eye animatable neural radiance fields for human face. *BMVC*, 2023. 2, 5

[40] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[41] Guo-Wei Yang, Wen-Yang Zhou, Hao-Yang Peng, Dun Liang, Tai-Jiang Mu, and Shi-Min Hu. Recursive-nerf: An efficient and dynamically growing nerf. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 1

[42] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 5, 1

[43] Yu Yu and Jean-Marc Odobez. Unsupervised representation learning for gaze estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[44] Yu Yu, Gang Liu, and Jean-Marc Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[46] 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 *The European Conference on Computer Vision*, 2020. 5, 6, 1

[47] Zhongqun Zhang, Wei Chen, Linfang Zheng, Aleš Leonardis, and Hyung Jin Chang. Trans6d: Transformer-based 6d object pose estimation and refinement. In *European Conference on Computer Vision*, pages 112–128. Springer, 2022. 1

[48] Linfang Zheng, Chen Wang, Yinghan Sun, Esha Dasgupta, Hua Chen, Aleš Leonardis, Wei Zhang, and Hyung Jin Chang. 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*, pages 17163–17173, 2023. 1

[49] Yufeng Zheng, Seonwook Park, Xucong Zhang, Shalini De Mello, and Otmar Hilliges. Self-learning transformations for improving gaze and head redirection. *Advances in Neural Information Processing Systems*, 2020. 2, 5, 6, 1

[50] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. In *European conference on computer vision*, pages 268–285. Springer, 2022. 3