

# VOODOO 3D: Volumetric Portrait Disentanglement for One-Shot 3D Head Reenactment

Phong Tran<sup>1</sup> Egor Zakharov<sup>2</sup> Long-Nhat Ho<sup>1</sup> Anh Tuan Tran<sup>3</sup> Liwen Hu<sup>4</sup> Hao Li<sup>1,4</sup>

<sup>1</sup>MBZUAI <sup>2</sup>ETH Zurich <sup>3</sup>VinAI Research <sup>4</sup>Pinscreen

{the.tran, long.ho}@mbzui.ac.ae anhtt152@vinai.io ezakharov@ethz.ch  
liwen@pinscreen.com hao@hao-li.com

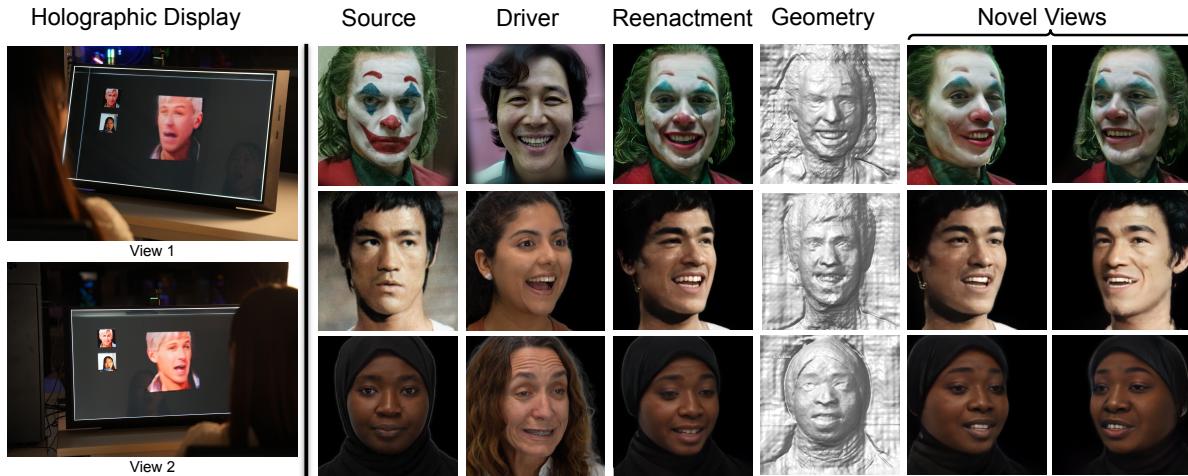


Figure 1. We introduce **VOODOO 3D**: a high-fidelity 3D-aware one-shot head reenactment technique. Our method transfers the expression of a driver to a source and produces view consistent renderings for holographic displays.

## Abstract

We present a 3D-aware one-shot head reenactment method based on a fully volumetric neural disentanglement framework for source appearance and driver expressions. Our method is real-time and produces high-fidelity and view-consistent output, suitable for 3D teleconferencing systems based on holographic displays. Existing cutting-edge 3D-aware reenactment methods often use neural radiance fields or 3D meshes to produce view-consistent appearance encoding, but, at the same time, they rely on linear face models, such as 3DMM, to achieve its disentanglement with facial expressions. As a result, their reenactment results often exhibit identity leakage from the driver or have unnatural expressions. To address these problems, we propose a neural self-supervised disentanglement approach that lifts both the source image and driver video frame into a shared 3D volumetric representation based on tri-planes. This representation can then be freely manipulated with expression tri-planes extracted from the driving images and rendered from an arbitrary view using neural radiance fields. We achieve this disentanglement via self-

supervised learning on a large in-the-wild video dataset. We further introduce a highly effective fine-tuning approach to improve the generalizability of the 3D lifting using the same real-world data. We demonstrate state-of-the-art performance on a wide range of datasets, and also showcase high-quality 3D-aware head reenactment on highly challenging and diverse subjects, including non-frontal head poses and complex expressions for both source and driver.

## 1. Introduction

Creating 3D head avatars from a single photo is a core capability in making a wide range of consumer AR/VR and telepresence applications more accessible, and user experiences more engaging. Graphics engine-based 3D avatar digitization methods [9, 14, 34, 38, 47, 48, 50, 59] are suitable for today’s video games and virtual worlds, and many commercial solutions exist (AvatarNeo [5], AvatarSDK [1], ReadyPlayerMe [6], in3D [2], etc.). However, the photorealism achieved by modern neural head reenactment tech-

niques is becoming increasingly appealing for advanced effects in video sharing apps and visual effects. For immersive telepresence systems that use AR/VR headsets, facial expression capture is typically achieved using tiny video cameras built into HMDs [30, 51, 57, 66, 72], while the identity of the source subject recorded using a separate process. However, the teleconferencing solutions based on holographic 3D displays (LookingGlass [4], LEIA [3], etc.) use regular webcams [84] or depth sensors [49]. As opposed to a video-based setting, head reenactment for immersive applications needs to be 3D-aware, meaning that in addition to generating the correct poses and expressions from a photo, multi-view consistency is critical.

While impressive facial reenactments results have been demonstrated using 2D approaches [27, 28, 85, 98, 103, 104], they typically struggle with preserving the likeness of the source and exhibit significant identity changes when varying the camera pose. More recently, 3D-aware one-shot head reenactment methods [37, 44, 54, 55, 61, 100] have used either 3D meshes or tri-plane neural radiance fields as a fast and memory efficient volumetric data representations for neural rendering. However, the expression and identity disentanglement in these methods is based on variants of linear face and expression models [15, 53] which lack expressiveness and high-frequency details. While these methods can achieve view consistency, facial expressions are often uncanny, and preserving the likeness of the input source portrait is challenging, especially for views different than the source image. Hence, input sources with extreme expressions and non-frontal poses are often avoided.

In this paper, we introduce the first 3D aware one-shot head reenactment technique that disentangles source identities and the target expressions fully volumetrically, and without the use of explicit linear face models. Our method is real-time and designed with holographic displays in mind, where a large number of views (up to 45) can be rendered in parallel based on their viewing angle. We leverage the fact that real-time 3D lifting for human heads has recently been made possible [84] with the help of Vision Transformers (ViT) [26], which avoids the need for inefficient optimization-based GAN-inversion process [70]. In particular, 3D lifting allows us to map 2D face images into a canonical tri-plane representation for both source and target subjects and treat identity and expression disentanglement independently from the head pose.

Once the source image and driver frame are lifted into a pose-normalized tri-plane representation, we extract appearance features from the source subject and expressions from the driver. The pose of the driver is estimated separately using a 3D face tracker and used as input to a neural renderer. Tri-plane-based feature extraction ensures view-consistent rendering, while facial appearance and driver expression feature use frontalized views from the 3D lifting

to enable robust and high-fidelity facial disentanglement. To handle highly diverse portraits (variations in facial appearance, hairstyle, head covering, eyewear, etc.), we propose a new method for fine-tuning Lp3D on real datasets by introducing a mixed loss function based on real and synthetic datasets. Our volumetric disentanglement and rendering framework is trained only using in-the-wild videos from the CelebV-HQ dataset [113] in a self-supervised fashion.

We not only demonstrate that our volumetric face disentanglement approach produces qualitative superior head reenactments than existing ones, but also show on a wide and diverse set of source images how non-frontal poses and extreme expressions can be handled. We have quantitatively assessed our method on multiple benchmarks and outperform existing 2D and 3D state-of-the-art techniques in terms of fidelity, expression, and likeness accuracy metrics. Our 3D aware head reenactment technique is therefore suitable for AR/VR-based immersive applications, and we also showcase a teleconferencing system using a holographic display from LookingGlass [4]. We summarize the main contributions as follows:

- First fully volumetric disentanglement approach for real-time 3D aware head reenactment from a single photo. This method combines 3D lifting into a canonical tri-plane representation and formalized facial appearance and expression feature extraction.
- A 3D lifting network that is fine-tuned on unconstrained real-world data instead of only generating synthetic ones.
- We demonstrate superior fidelity, identity preservation, and robustness w.r.t. current state-of-the-art methods for facial reenactment on a wide range of public datasets. We plan to release our code to the public.

## 2. Related Work

**2D Neural Head Reenactment.** The problem of generating animations of photorealistic human heads given images or video inputs has been thoroughly explored using various neural rendering techniques in the past few years, outperforming traditional 3DMM-based methods [8, 27, 32, 45, 65, 68, 81, 82, 97] which often appear uncanny due to their compressed linear space. These approaches can be categorized into one-shot and multi-shot ones. While multi-shot methods generally achieve high-fidelity results, they are not suitable for many consumer applications as they typically require an extensive amount of training data, such as a monocular video capture [10, 11, 18, 21, 31, 35, 94, 109–111, 114], and sometimes even a calibrated multi-view stereo setup [13, 30, 57, 60, 72]. More recently, few-shot techniques [105] have also been introduced.

To maximize accessibility, a considerable number of methods [17, 27, 28, 33, 36, 40, 68, 75–77, 79, 80, 85, 88, 89, 98, 102–104, 108] use a single portrait as input by leveraging advanced generative modeling techniques based

on in-the-wild video training data. While most methods rely on linear face models to extract facial expressions, the head reenactment technique from Drobyshev et al. [28] directly extract expression features from cropped 2D face regions, allowing them to obtain better face disentanglements, which results in higher fidelity face synthesis. While similar to our proposed approach in avoiding the use of low dimensional linear face models, their method is purely 2D and struggle with ensuring identity and expression consistency when novel views are synthesized.

**3D-Aware One-Shot Head Reenactment.** Due to potential inconsistencies when rendering from different views or poses, a number of 3D-aware single shot head reenactment techniques [7, 19, 20, 25, 64, 67, 73, 78, 90, 93, 95, 96] have been introduced. These methods generally use an efficient 3D representation, such as neural radiance fields or 3D mesh, to geometrically constraint the neural rendering and improve view consistency. ROME [44] for instance is a mesh-based method using FLAME blendshapes [52] and neural textures. While view-consistent results can be produced for both face and hair regions, the use of low resolution polygonal meshes hinders the neural renderer to generate high-fidelity geometric and appearance details.

Implicit representations such as HeadNeRF [37] and MofaNeRF [39] use a NeRF-based parametric model which supports direct control of the head pose of the generated images. While real-time rendering is possible, these methods require intensive test-time optimization and often fail to preserve the identity of the source due to the use of compact latent vectors. Most recent methods [54, 55, 100] adopt the highly efficient tri-plane-based neural fields representation [20] to encode the 3D structure and appearance of the avatars head. Compared to the previous works on view-consistent neural avatars [37, 44, 54, 55, 61, 100], we refrain from depending on parametric head models for motion synthesis and, instead, learn the volumetric motion model from the training data. This methodology enables us to narrow the identity gap between the source and generated images and yield a superior fidelity of the generated motion compared to competing approaches, and hence a higher quality disentanglement for reenactment.

**3D GAN Inversion.** When training a whole reconstruction and disentangled reenactment model end-to-end on facial performance videos, one can introduce substantial overfitting and reduce the quality of the results. To address this problems, we focus our training approach to an inversion of pre-trained 3D-aware generative models for human heads.

We use tri-plane-based generative network EG3D [20] as the foundational generator, due to its proficiency in producing high-fidelity and view-consistent synthesis of human heads. For a given image, an effective 3D GAN inver-

sion method should leverage these properties for estimating latent representations, which can be decoded into outputs that maintain view consistency and faithfully replicate the contents of the input. One naive approach is to adapt GAN inversion methods that were initially designed for 2D GANs to the EG3D pre-trained network. These methods either do a time consuming but more precise optimization [43, 70] or train a fast but less accurate encoder network [69, 83] to obtain the corresponding latent vectors. They often produce incorrect depth prediction, leading to clear artifacts in novel view synthesis. Hence, some methods are specifically designed for inverting 3D GANs, which either do multi-view optimization [46, 92] or predict residual features/tri-plane maps for refining the initial inversion results [12, 84, 99, 101].

In this work, we rely on the state-of-the-art EG3D inversion method Lp3D [84]. While achieving excellent novel-view synthesis results, it lacks disentanglement between the appearance and expression of the provided image and is unable to impose various driving expressions onto the input. To address this limitation, we propose a new method that introduces appearance-expression disentanglement in the latent space of tri-planes using our new self and cross-reenactment training pipeline while relying on a pre-trained but fine-tuned Lp3D network for regularization which enables highly consistent view synthesis.

### 3. 3D-Aware Head Reenactment

As illustrated in Fig. 2, our head reenactment pipeline consists of three stages: 1) 3D Lifting, 2) Volumetric Disentanglement, and 3) Tri-plane Rendering. Given a pair of source and driver images, we first frontalize them using a pre-trained but fine-tuned tri-plane-based 3D lifting module [84]. This driver alignment step is crucial and allows our model to disentangle the expressions from the head pose, which prevents overfitting. Then, the frontalized faces are fed into two separate convolutional encoders to extract the face features  $F_s$  and  $F_d$ . These extracted features are concatenated with the ones extracted from the tri-planes of the source, and all are fed together into several transformer blocks [91] to produce the expression tri-plane residual, which is added to the tri-planes of the source image. The final target image can be rendered from the new tri-planes using a pre-trained tri-plane renderer using the driver’s pose.

#### 3.1. Fine-Tuned 3D Lifting

We adopt Lp3d [84] as a 3D face-lifting module, which predicts the radiance field of any given face image in real-time. Instead of using an implicit multi-layer perceptron [63] or sparse voxels [29, 74] for the radiance field, Lp3D [84] uses tri-planes [20], which can be computed using a single forward of a deep learning network. Specifically, for a given source image  $x_s$ , we first extract the tri-planes  $T$  using a

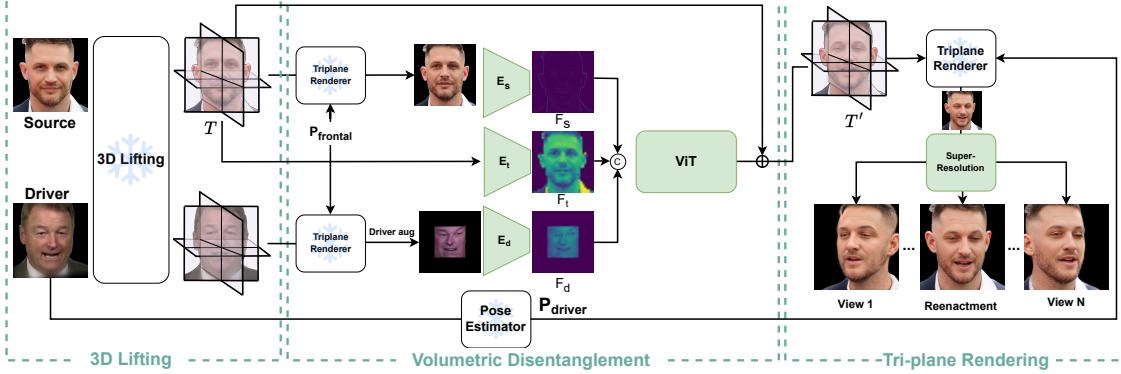


Figure 2. Given a pair of source and driver images, our method processes them in three steps: **3D Lifting** into tri-plane representations, **Volumetric Disentanglement**, which consists of source and driver frontalization and tri-plane residual generation, and **Tri-plane Rendering** via volumetric ray marching with subsequent super-resolution.

transformer-based appearance encoder  $\mathbf{E}_{\text{app}}$ :

$$\mathbf{E}_{\text{app}}(x_s) = T \in \mathbb{R}^{3 \times H \times W \times C} = \{T_{xy}, T_{yz}, T_{zx}\}. \quad (1)$$

The color  $c$  and density  $\sigma$  of each point  $p = (x, y, z)$  in the radiance field can be obtained by projecting  $p$  onto the three planes and by summing up the features at the projected positions:

$$c, \sigma = \mathbf{D}(F_{xy} + F_{yz} + F_{zx}), \quad (2)$$

where  $\mathbf{D}$  is a shallow MLP decoder for the tri-plane rendering,  $F_{xy}$ ,  $F_{yz}$ , and  $F_{zx}$  are the feature vectors at the projected positions on  $xy$ ,  $yz$ , and  $zx$  planes, respectively, calculated using bilinear interpolation. The rendered  $128 \times 128$  image is then upsampled using a super-resolution module to produce a high-resolution output. To train the encoder  $\mathbf{E}_{\text{app}}$ , Lp3D [84] uses synthetic data generated from a 3D-aware face generative model [20]. While these synthetic data have ground truth camera poses, they are limited to the face distribution of the generative model. As a result, Lp3D can fail to generalize to in-the-wild images as shown in Fig. 5. To prevent this, we fine-tune the pre-trained Lp3D on a large-scale real-world dataset. We also replace the original super-resolution module in Lp3D with a pre-trained GF-PGAN [87], which is then fine-tuned together with Lp3D (see Sec. 3.4).

### 3.2. Disentangling Appearance and Expression

Separating facial expression from the identity appearance in a 3D radiance field is very challenging especially when source and driver subjects have misaligned expressions. In order to simplify the problem, we use our 3D lifting approach to bring both source and driver heads into a pose-oriented space where faces are frontalized. Here, we denote frontalized source and driver images as  $x_s^f$  and  $x_d^f$ , respectively. These images are then fed into two separate convolutional source and driver encoders  $\mathbf{E}_s$  and  $\mathbf{E}_d$  to produce

coarse feature maps:

$$\begin{aligned} F_s &= \mathbf{E}_s(x_s^f) \\ F_d &= \mathbf{E}_d(x_d^f) \end{aligned}$$

Since we already have the source’s tri-plane, which encodes the 3D shape of the source, we use another encoder to encode this tri-plane and concatenate it together with the coarse frontalized feature maps of the images to produce expression feature  $F$ :

$$\begin{aligned} F_t &= \mathbf{E}_t(T) \\ F &= F_s \oplus F_d \oplus F_t \end{aligned}$$

Even though face frontalization aligns the source and the driver, there is still some misalignment between the two faces, e.g., the positions of the eyes may be different, or one mouth is open while the other is closed. Therefore, we feed the concatenation of the feature maps into several transformer blocks to produce the final residual tri-plane  $\mathbf{E}_v(F)$ . This residual is then added back to the source’s tri-planes to change the source’s expression to the driver’s expression  $T' = T + \mathbf{E}_v(F)$ . Unlike LPR [55], we do not use a 3D face model to compute the expression but instead use the RGB images of the source and the driver directly, allowing the model to learn high-fidelity and realistic expressions.

### 3.3. Tri-Plane Rendering

The resulting tri-planes are then volumetrically rendered into one or multiple output images using pose parameters and viewing angles in the case of a holographic display. Following EG3D [20], we use a neural radiance fields (NeRFs)-based volumetric ray marching approach [62]. However, instead of encoding each point in space via positional encodings [62], the features of the points along rays are calculated using their projections onto tri-planes. Since tri-planes are aligned with the frontal face, we can compute

these rays directly using camera extrinsics  $P_{\text{driver}}$  predicted by an off-the-shelf 3D head pose estimator [24].

While the renderings are highly view-consistent, the large number of points evaluated for each ray still limits the output resolution for real-time performance. We therefore follow [55] and employ a 2D upsampling network [86] based on StyleGAN2 [42], which in our experiments produced higher quality results than the upsampling approach in EG3D [20]. Finally, for holographic displays, we generate a number of renderings based on their viewing angles and simply using the head pose parameter. Real-time performance is achieved using efficient inference libraries such as TensorRT, half-precision, and batched inference over multiple GPUs.

### 3.4. Training Strategy

**Fine-Tuning Lp3D.** To make Lp3D work with in-the-wild images, we fine-tune it on a large-scale real-world video dataset [112]. Unlike the use of synthetic data, real-world data do not have ground-truth camera parameters and facial expressions in monocular videos are typically inconsistent over time. While the camera parameters can be estimated using standard 3D pose estimators, the expression differences are difficult to determine. However, we found that we can ignore this expression difference and fine-tune Lp3D using real data together with continuous training on synthetic data. In particular, our experiments indicate that the fine-tuned model can still faithfully reconstruct 3D faces from the input without changing expressions and still generalize successfully on in-the-wild images. Specifically, on real video data, we sample two frames  $x_s^r$  and  $x_d^r$  and estimate their camera parameters  $P_s^r$  and  $P_d^r$ . Similar to [20], we assume a fixed intrinsics for standard portraits for all images. Then we use  $E_{\text{app}}$  from Lp3D to calculate the tri-planes of  $x_s^r$ , render it using the two poses, and calculate reconstruction losses on the two rendered images:

$$\mathcal{L}_{\text{real}} = \|\text{Lp3D}(x_s^r, P_d^r) - x_d^r\| + \|\text{Lp3D}(x_s^r, x_s^r) - x_s^r\|,$$

where  $\text{Lp3D}(x, P)$  is the face in  $x$  re-rendered using camera pose  $P$  and  $\mathcal{L}_{\text{real}}$  is the loss for real images. Simultaneously, we render two synthetic images employing an identical latent code but through varying camera views and calculate the synthetic loss  $\mathcal{L}_{\text{syn}}$ :

$$\mathcal{L}_{\text{syn}} = \|\text{Lp3D}(x_s^f, P_d^s) - x_d^s\| + \|\text{Lp3D}(x_s^s, P_s^s) - x_s^s\|$$

$$\mathcal{L}_{\text{tri}} = \|E_{\text{app}}(x_s^f) - T\|,$$

where  $T$  is the ground-truth tri-planes returned by EG3D [20] and  $\mathcal{L}_{\text{tri}}$  is the tri-plane loss adopted directly from Lp3D. The final loss  $\mathcal{L}_{\text{app}}$  for fine-tuning Lp3D can be formulated as:

$$\mathcal{L}_{\text{app}} = \mathcal{L}_{\text{real}} + \lambda_{\text{syn}} \mathcal{L}_{\text{syn}} + \lambda_{\text{tri}} \mathcal{L}_{\text{tri}}$$

where  $\lambda_{\text{syn}}$  and  $\lambda_{\text{tri}}$  are tunable hyperparameters.

**Disentangling Appearance and Expressions.** In this stage, we also use real-world videos as training data. For a pair of source and driver images  $x_s$  and  $x_d$  sampled from the same video, we apply the reconstruction loss  $\mathcal{L}_{\text{recon}}$  which is a combination of  $L1$ , perceptual [106], and identity losses, between the reenacted image  $x_{s \rightarrow d}$  and the corresponding ground-truth  $x_d$ :

$$\begin{aligned} \mathcal{L}_{\text{recon}} = & \|x_{s \rightarrow d} - x_d\|_1 + \phi(x_{s \rightarrow d}, x_d) \\ & + \|\text{ID}(x_{s \rightarrow d}) - \text{ID}(x_d)\|_1, \end{aligned}$$

where  $\phi$  is the perceptual loss and  $\text{ID}(\cdot)$  is a pretrained face recognition model. Similar to other works that use RGB images directly to calculate expressions [28], our proposed encoder also suffers from an “identity leaking” issue. Since there is no cross-reenactment dataset, the expression module is trained with self-reenactment video data. Therefore, without proper augmentation and regularization, the expression module can leak identity information from the driver to the output, making the model fail to generalize to cross-reenactment tasks. Hence, we introduce a *Cross Identity Regularization*. Specifically, we further sample an additional driver frame  $x_{d'}$  from another video. We incorporate a GAN loss where real samples are  $\text{Lp3D}(x_s, P^d)$  and fake samples are  $x_{s \rightarrow d'}$ . This GAN loss is also conditioned on the identity vector of the source  $\text{ID}(x_s)$ . Following [28], we also apply strong augmentation (random warping and color jittering) and additionally mask the border of the driver randomly to further reduce potential identity leaks. The loss for expression training can be summarized as:

$$\mathcal{L}_{\text{exp}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{CIR}} \mathcal{L}_{\text{CIR}},$$

where  $\mathcal{L}_{\text{CIR}}$  and  $\lambda_{\text{CIR}}$  are cross identity regularization and its hyperparameter, respectively.

**Global Fine-Tuning.** After training both Lp3D and the expression module, we iteratively fine-tune the two modules using the same losses as the previous sections. Specifically, for every 10000 iterations, we freeze one module and fine-tune the other and vice versa. In addition, we add a GAN loss on the super-resolution output of the Lp3D module.

## 4. Experiments

**Implementation Details.** We train our model on CelebV-HQ dataset [113] using 7 NVIDIA RTX A6000 ADA (50Gb memory each). We use AdamW [58] to optimize the parameters with a learning rate of  $10^{-4}$  and batch size of 28. The Lp3D finetuning takes 5 days for 500K iterations to converge. Training the expression module takes 2 days, and the iterative fine-tuning takes another 5 days. More training details, such as hyperparameter fine-tuning or architecture of the networks, can be found in the supplementary materials.

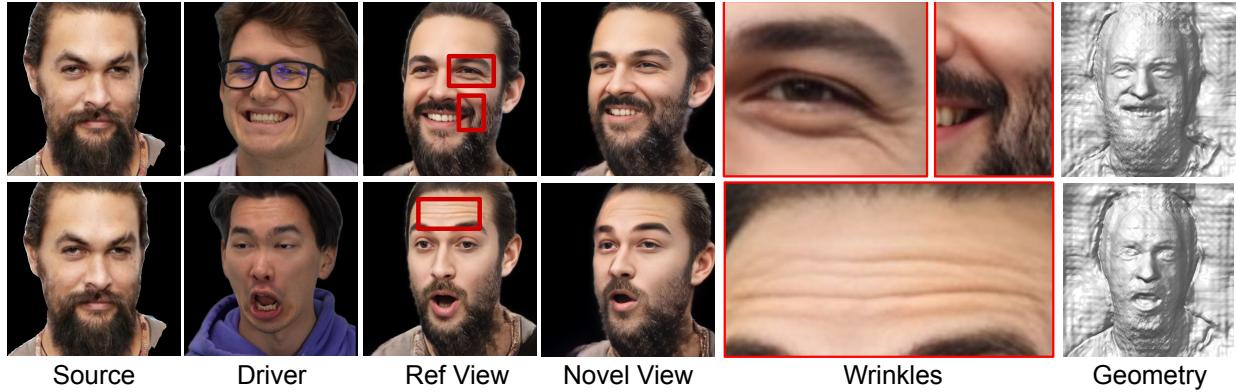


Figure 3. Expression dependent high-fidelity details, incl. eye and forehead wrinkles, as well as nasolabial folds (see zoom-ins)

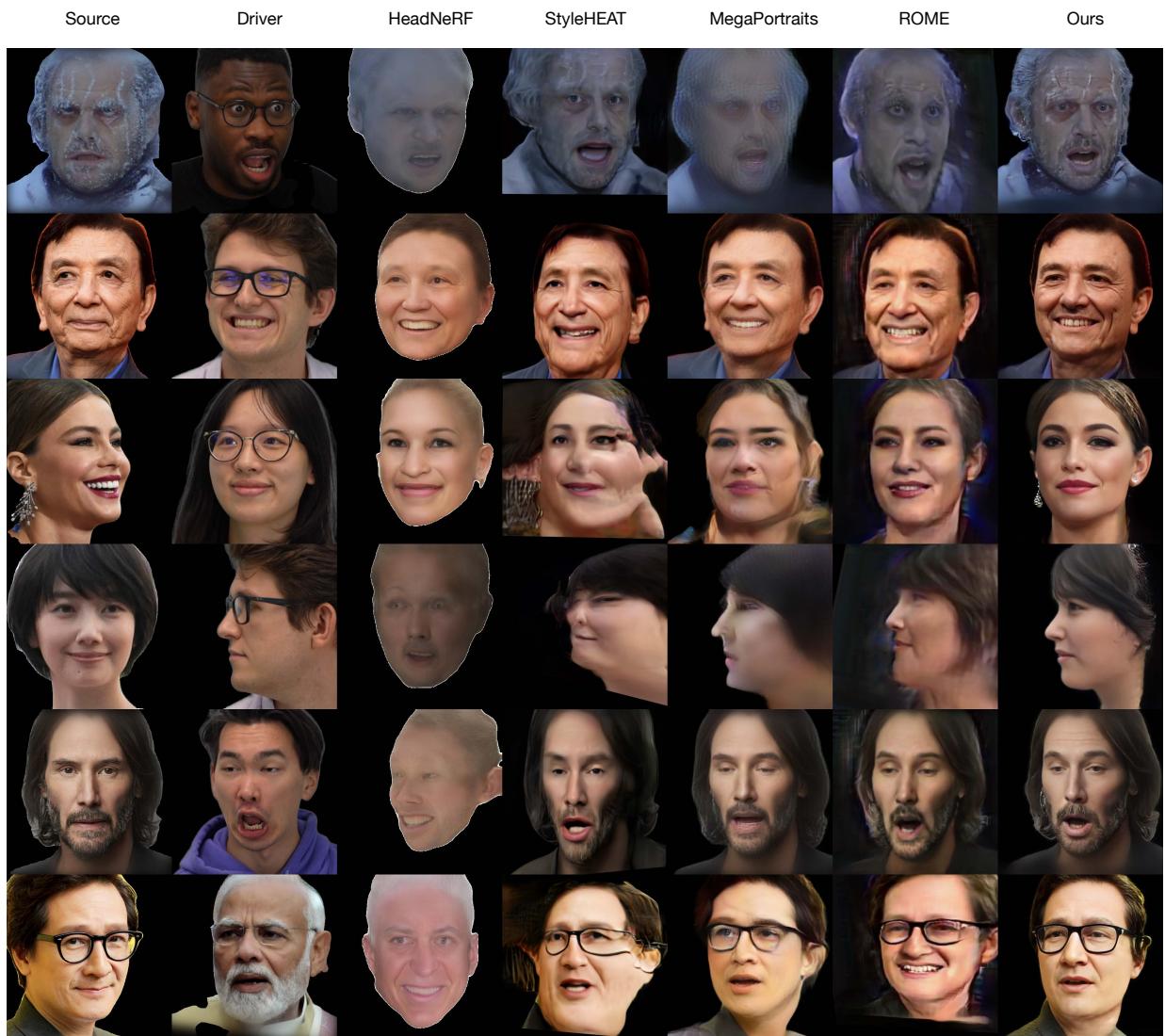


Figure 4. A qualitative comparison with the baselines on in-the-wild photos. Notice that our method is capable of producing a variety of facial expressions, and handle highly diverse subjects, with and without accessories, as well as extreme head poses, such as rows 3 and 4.

Method	Self-reenactment							Cross-reenactment			Method	Cross-reenactment		
	PSNR ↑	SSIM ↑	LPIPS ↓	NAKD ↓	ECMD ↓	FID ↓	CSIM ↑	ECMD ↓	FID ↓	CSIM ↑	ECMD ↓	FID ↓		
ROME [44]	18.46	0.488	0.351	0.030	0.594	138	0.507	<b>0.740</b>	172	0.519	0.91	52.6	ROME [44]	
StyleHeat [98]	19.73	0.689	0.278	0.035	0.748	89.8	0.398	0.744	95.5	0.346	0.88	113	HeadNeRF [37]	
OTAvatar [61]	19.28	0.749	0.289	0.035	0.651	67.0	0.462	0.901	72.4	0.467	0.85	50.2	StyleHeat [98]	
MegaPortraits [28]	21.10	0.731	0.291	0.022	0.755	52.0	0.729	0.771	61.7	<b>0.647</b>	<b>0.77</b>	29.2	MegaPortraits [28]	
Ours	<b>22.83</b>	<b>0.768</b>	<b>0.168</b>	<b>0.012</b>	<b>0.426</b>	<b>40.5</b>	<b>0.754</b>	0.754	<b>36.4</b>	0.608	0.79	<b>23.6</b>	Ours	

Table 1. Evaluation on HDTF [107] dataset. Our method outperforms the competitors across almost all of the metrics for both self- and cross-reenactment scenarios.



Figure 5. Our implementation of Lp3D [84] before and after CelebV-HQ [113] fine-tuning.

	CSIM ↑	ECMD ↓
Lp3D	0.548	0.82
Lp3D-FT	0.670	0.76
w/o frontal	0.668	1.01
w/o CIR	0.570	0.97
Ours	0.608	0.79

Table 3. Ablation studies conducted on CelebA-HQ [41] dataset. FT is a fine-tuned version of Lp3D, and “frontal” denotes frontalization of the source and driver.



Figure 6. Ablation study for source and driver frontalization and cross identity regularization (CIR).



Figure 7. Qualitative comparison with LPR [55] method on the samples from HDTF [107] dataset.

Unlike Lp3D, our method reenacts faces without relifting in 3D for every frame. For each driver, we perform only a single frontalization (0.0115 ms), one inference for expression encoding (0.0034 ms), and one tri-plane rendering at  $128 \times 128$  resolution (0.0071 ms), and one neural upsampling (0.0099 ms). Each view runs at 31.9 fps on an Nvidia RTX 4090 GPU including I/O. More details on performance can be found in the supplemental materials.

We compare our method with state-of-the-art 3D-

Table 2. Evaluation on CelebA-HQ [41] dataset.

based [37, 44, 61] and 2D-based [28, 98] models. For MegaPortraits [28], we use our own implementation that was trained on the CelebV-HQ dataset. Similar to previous works, we evaluate our method using public benchmarks, including CelebA-HQ [41] and HDTF [107]. For CelebA-HQ, we split the data into two equal sets. Each set contains around 15K images. Then, we use one set as the source and the rest as driver images. For the HDTF dataset, we perform cross-reenactment by using the first frame of each video as source and 200 first frames of other videos as drivers, which is more than 60K data pairs. Similarly, to evaluate self-reenactment, we also use the first frames of each video as sources and the rest of the same video as the driver. Furthermore, we also collected 100 face images on the internet and around 100 high-quality videos for qualitative comparison purposes. We provide the video results in the supplementary materials.

**Quantitative Comparisons.** Given a source image  $x_s$ , a driver image  $x_d$ , and reenacted output  $x_{s \rightarrow d}$  we use EMO-CAv2 [23] to extract the FLAME [53] expression coefficients of the prediction and the driver, as well as the shape coefficients of the source. We then compute 2 FLAME meshes using the predicted shape coefficients in world coordinates, one with the expression coefficients of the driver and one with the expression coefficients of the reenacted output. We measure the distance between the 2 meshes and denote this expression metric as ECMD. Moreover, we also use cosine similarity between the embeddings of a face recognition network (CSIM) [102], normalized average keypoint distance (NAKD) [16], perceptual image similarity (LPIPS) [106], peak signal-to-noise ratio (PSNR), and structure similarity index measure (SSIM).

We provide quantitative comparisons on HDTF and CelebA-HQ datasets in Tab. 1 and Tab. 2, respectively, and show that our method outperforms existing methods on both datasets. We also note that our FID and CSIM scores are significantly more reliable than the others, while expression-based metrics such as NAKD and ECMD are either better or very close to the best baseline, w.r.t output quality, expression accuracy, and identity consistency.

**Qualitative Results.** Fig. 4 and Fig. 3 showcase the qualitative results of cross-identity reenactment on in-the-wild images. Compared to the baselines [28, 37, 44, 98], our

reenactment faithfully reconstructs intricate and complex elements, such as hairstyle, facial hair, glasses, and facial makeups. Furthermore, our method effectively generates realistic and fine-scale dynamic details that match the driver’s expressions including substantial head pose rotations. We also conduct a comparative analysis of our results with the current state-of-the-art 3D-aware method LPR [55] in Fig. 7. Compared to LPR, our method achieves superior identity consistency. We further refer to the supplemental video for a live demonstration of our holographic telepresence system and animated head reenactment results and comparisons, with and without disentangled poses.

**Ablation Study.** We compare Lp3D with and without fine-tuning on the CelebA-HQ dataset in Tab. 3 and show several examples in Fig. 5. Without fine-tuning on real data, our implementation of Lp3D fails to preserve the identity of the input image, resulting in a considerably lower CSIM score. We also try without any facial frontalization in the expression module and instead use the source and driver images directly to calculate the expression tri-plane residual. We observe in Fig. 6 that without face frontalization, the model completely ignores the expression of the driver and keeps the expression of the input source instead. We show in Tab. 3, that facial frontalization leads to much better ECMD score. We then measure the effectiveness of the GAN-based cross-identity regularization on the CelebA-HQ dataset,  $\mathcal{L}_{CIR}$ . Without this loss, identity characteristics (hairstyle or color) can leak from the driver to the output. See column 4 in Fig. 6. Tab. 3 also shows that cross-identity regularization can reduce identity leaking and improve the CSIM score. Lastly, we have also attempted to train our model end-to-end using the same losses and optimization process instead of our proposed iterative fine-tuning. Even with a lower learning rate and the use of pre-trained Lp3D weights, we were unable to succeed.

**Limitations.** Limitations of our approach are illustrated in Figure 8. For source images that are extremely side ways (i.e., over 90°), our method can produce a plausible frontal face, but the likeness cannot be guaranteed due to insufficient visibility. For very highly stylized portraits, such as cartoons, our framework often produces photorealistic facial elements such as teeth which can be inconsistent in style. Due to the dependence on training data volume and diversity, accessories such as dental braces or glasses may disappear or look different during synthesis. We believe that providing more and better training data can further improve the performance of our algorithm.

## 5. Discussion

We have demonstrated that a fully volumetric disentanglement of facial appearance and expressions is possible

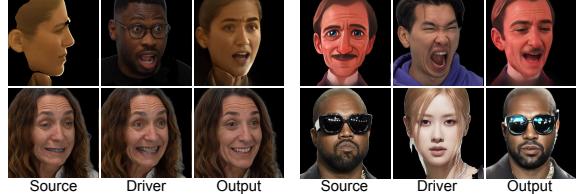


Figure 8. Failure cases of our method include side views in the source, extreme expressions, modeling of cartoonish characters and paintings, as well as modeling the reflections and semi-transparency of the eyewear.

through a shared canonical tri-plane representation. In particular, an improved disentanglement also leads to higher fidelity and more robust head reenactment, when compared to existing methods that use linear face models for expressions, especially for non-frontal poses. A critical insight of our approach is that head frontalization via 3D lifting is particularly effective for extracting features that can encode fine-scale details and expressions such as wrinkles and folds. The resulting reenactment is also highly view-consistent for large angles, making our solution suitable for holographic displays. We have also shown that the 3D lifting model can still be successfully trained with real data despite the fact that different frames with the same subject have varying facial expressions. Without a fine-tuned 3D lifting model, our 3D-aware reenactment framework would struggle with preserving the identity of the source, especially for side views. Our experiments indicate that our results achieve better visual quality and are more robust to extreme poses, which is validated via an extensive evaluation on multiple datasets.

**Risks and Potential Misuse.** The proposed method is intended to promote avatar-based 3D communication. Nevertheless, our AI-based reenactment solution produces synthetic but highly realistic face videos from only a single photo, which could be hard to distinguish from a real person. Like deepfakes and other facial manipulation methods, potential misuse is possible and hence, we refer to the supplemental material for more discussions.

**Future Work.** We are also interested in expanding our work to upper and full body reenactment, where hand gestures can be used for more engaging communication. To this end, we plan to investigate the use of canonical representations for human bodies, such as T-poses. As our primary motivation, we have showcased a solution using holographic displays for immersive 3D teleconferencing. However, we believe that our approach can also be extended to AR/VR HMD-based settings where full 360° head views are possible. The recent work by An et al. [7] is a promising avenue for future exploration.

# VOODOO 3D: Volumetric Portrait Disentanglement for One-Shot 3D Head Reenactment

## Supplementary Material

### 6. Training Details

**Training Data.** We fine-tune Lp3D using CelebV-HQ dataset [113]. For the expression modules, we also use the CelebV-HQ dataset but adopt an expression re-sampling process to make the expressions of the sources and drivers during training more different. Specifically, for a given video, we use EMOCA [23] to reconstruct the mesh of every frame without the head pose. Let these obtained meshes be  $\{M_1, M_2, \dots, M_n\}$ , we first pick two frames  $x^*$  and  $y^*$  such that the distance between their meshes are maximized:

$$x^*, y^* = \arg \max_{x, y} \|M_x - M_y\|_2.$$

Then we pick the third frame  $z^*$  such that:

$$z^* = \arg \max_z \min (\|M_{x^*} - M_z\|, \|M_{y^*} - M_z\|).$$

We use this frame selection process for all the videos in the CelebV-HQ dataset [113] and use the re-sampled frames to train the expression modules. A few examples from this selection process are shown in Fig. 9.



Figure 9. Some examples of our training data extracted from the CelebV-HQ dataset [113]

**Driver Augmentation.** To prevent identity leaking from the driver to the output, we apply several augmentations to

```
Conv2d(96, 96, kernel_size=3, stride=2, padding=1)
ReLU()
Conv2d(96, 96, kernel_size=3, stride=1, padding=1)
ReLU()
Conv2d(96, 128, kernel_size=3, stride=2, padding=1)
ReLU()
Conv2d(128, 128, kernel_size=3, stride=1, padding=1)
ReLU()
Conv2d(128, 128, kernel_size=3, stride=1, padding=1)
```

Table 4. Architecture of  $E_T$

the frontalized driver images, including: (1) Kornia color jiggle<sup>1</sup> with parameters for brightness, contrast, saturation, hue set to 0.3, 0.4, 0.3, and 0.4, respectively; (2) random channel shuffle; (3) random warping<sup>2</sup>; and (4) random border masking with the mask ratio uniformly sampled from 0.1 to 0.3. During testing, we removed all the augmentations except the random masking and fixed the mask ratio to 0.25. This random masking greatly improves the consistency in the output, especially for border regions. In addition, since we mask the border with a fixed rate, we can modify the renderer to only generate the center of the frontalized driver and further improve the performance.

**Architecture Details.** Our architecture design is inspired by Lp3D [84]. Specifically, for  $E_s$  and  $E_d$ , we use two separate DeepLabV3 [22] with all normalization layers removed. Since the triplane already captures deep 3D features of the source, we adopt a simple convolutional network for  $E_t$ , which is given in Tab. 4. Recall that:

$$F = F_s \oplus F_d \oplus F_t$$

For the final transformer that is applied on the concatenations of the feature maps  $F$ , we use a slight modification of  $E_{low}$  (light-weight version) in Lp3D [84]. The architecture of this module is given in Tab. 5 where block used is the transformer block in SegFormer [91]. As mentioned in our paper, we use a pretrained GFPGAN as the super-resolution module. This module is loaded from a public pretrained weight GFPGAN v1.4 [87] and fine-tuned end-to-end with the network.

<sup>1</sup><https://kornia.readthedocs.io/en/latest/augmentation.module.html#kornia.augmentation.ColorJiggle>

<sup>2</sup><https://github.com/deepfakes/faceswap/blob/a62a85c0215c1d791dd5ca705ba5a3fef08f0ffd/lib/training/augmentation.py#L318>

```

PatchEmbed(64, patch=3, stride=2, in=640, embed=1024)
Block(dim=1024, num_heads=4, mlp_ratio=2, sr_ratio=1)
Block(dim=1024, num_heads=4, mlp_ratio=2, sr_ratio=1)
PixelShuffle(upscale_factor=2)
upsample(scale_factor=2, mode=bilinear)
Conv2d(256, 128, kernel_size=3, stride=1, padding=1)
ReLU()
upsample(scale_factor=2, mode=bilinear)
Conv2d(128, 128, kernel_size=3, stride=1, padding=1)
ReLU()
Conv2d(128, 96, kernel_size=3, stride=1, padding=1)

```

Table 5. Architecture of the transformer network used in the expression module.

**Training Losses.** To train the model used in our experiments, we set  $\lambda_{\text{syn}} = 0.1$ ,  $\lambda_{\text{tri}} = 0.01$ , and  $\lambda_{\text{CIR}} = 0.01$ . For GAN-based losses, we use hinge loss [56] with projected discriminator [71].

## 7. Implementation Details for Holographic Display System

We implement our model on a Looking Glass monitor 32”<sup>3</sup>. To visualize results on a holographic display, we must render multiple views for each frame using camera poses with a yaw angle that spans the range from  $-17.5^\circ$  to  $17.5^\circ$ . In our case, we find that using 24 views is sufficient for the user experience. While our model can run at 32FPS using a single NVIDIA RTX 4090 on a regular monitor, which only requires a single view at a time, it cannot run in real-time when rendering 24 views simultaneously. Thus, to achieve real-time performance for the Looking Glass display, we ran the holographic telepresence demo on seven NVIDIA RTX 6000 ADA GPUs.

We parallelize the rendering process to four GPUs, so each one needs to render six views in a batch. We dedicate one GPU for driving image pre-processing and another one for disentangled tri-plane estimation. We use the last GPU to run the looking-glass display itself. This setup results in 25 FPS for the whole application. We showcase the results rendered on the holographic display in the supplementary videos.

## 8. Additional Comparisons with LPR [55]

In this section, we compare our method with the current state-of-the-art in 3D aware one-shot head reenactment, LPR [55] using their test data from HDTF [107] and CelebA-HQ datasets [41]. In particular, for CelebA-HQ, they use even-index frames as sources and odd-index frames as drivers, while in contrast, in our experiment section, we use the first half as sources and the rest as

<sup>3</sup><https://lookingglassfactory.com/looking-glass-32>

drivers. For the HDTF dataset, they use a single driver (WRA\_EricCantor\_000) and the first frame of each video as source image. Compared to our split, this reduces the diversity in the driver images. We provide the comparison results in Tab. 6 and Tab. 7. The ECMD scores on both datasets show that our method is more accurate in transferring expression from the driver to the source images. On the HDTF dataset, our results have much higher CSIM. Our FID score is better than LPR [55] on CelebA-HQ but worse on the HDTF dataset. We found that the HDTF’s ground-truth images have poor quality while our outputs are higher in quality; this mismatch causes our FID to be unimpressive on this dataset. Hence, this FID arguably does not correctly reflect the performance of our model. According to the qualitative examples in Fig. 14, our method captures the driver’s expression more accurately than LPR. However, we note that our quality is even higher than the input, as can be observed in Fig. 14.

We also provide extensive qualitative comparisons in Fig. 16 and Fig. 14. The expression of our output images is more realistic and faithful to the driver, which is particularly more visible in the mouth/teeth/jaw region, as well as for driver or source side views. Notably, in Fig. 15, it can be observed that LPR fails to remove the smiling from the source, resulted in inaccurate expression in the reenacted output while our method can still successfully transfer the expression from the driver to the source image.

Method	Cross-reenactment		
	CSIM	ECMD	FID
LPR [55]	0.531	0.912	<b>25.26</b>
Ours	<b>0.774</b>	<b>0.860</b>	54.15

Table 6. Quantitative comparisons with LPR [55] on HDTF dataset using the test split proposed in [55].

Method	Cross-reenactment		
	CSIM	ECMD	FID
LPR [55]	<b>0.643</b>	0.483	47.39
Ours	0.628	<b>0.473</b>	<b>34.27</b>

Table 7. Quantitative comparisons with LPR [55] on CelebA-HQ dataset using the test split proposed in [55].

## 9. Additional Qualitative Comparisons

We provide additional qualitative comparisons with other methods in Fig. 18, Fig. 19, Fig. 20, Fig. 21, Fig. 22, Fig. 23, Fig. 24, Fig. 25, Fig. 26, Fig. 27, Fig. 28, Fig. 29, Fig. 30, and Fig. 31.

In Fig. 17, we evaluate the ability to synthesize novel views of our method. In addition, we also reconstruct the 3D mesh of the reenacted results.

In Fig. 10, we evaluate our model on self-reenactment task using HDTF and our collected datasets.

In Fig. 11, we compares our method with the others on source images that have jewelries. As can be seen, other methods struggle to reconstruct the jewelries while our results still have the jewelries from the source input.

## 10. Additional Experiments with PTI [70]

Our method can achieve high-quality results without noticeable identity change without additional fine-tuning, which is known to be computationally expensive. In this section, we try to fine-tune [70] the super-resolution module using PTI [70] for 100 iterations, which takes around 1 minute per subject. Without PTI, our pipeline runs instantly similarly to [55]. For most cases, the difference between results with and without fine-tuning is negligible. However, for out-of-domain images such as Mona Lisa, PTI fine-tuning helps retain the oil-painting style and fine-scale details from the input source. For the fine-tuning results, please refer to the supplementary video.

## 11. Additional Limitations

Besides the limitations that we discussed in the paper, we also notice that the model cannot transfer tongue-related expressions or certain asymmetric expressions due to limited training data for our 3D lifting and expressions module. Since our method is not designed to handle the shoulder pose, the model uses the head pose as a single rigid transformation for the whole portrait. This issue would be an interesting research direction for future work. Also, our model sometimes fails to produce correct accessories when the input has out-of-distribution sunglasses. These failure cases are illustrated in Fig. 12.

## References

- [1] ItSeez3D AvatarSDK, <https://avatarsdk.com>. 1
- [2] in3D, <https://in3d.io>. 1
- [3] Leia, <https://www.leiainc.com>. 2
- [4] Looking Glass Factory, <https://lookingglassfactory.com>. 2
- [5] Pinscreen Avatar Neo, <https://www.avatarneo.com>. 1
- [6] ReadyPlayerMe, <https://readyplayer.me>. 1
- [7] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 8
- [8] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20364–20373, 2022. 2
- [9] Haoran Bai, Di Kang, Haoxian Zhang, Jinshan Pan, and Linchao Bao. Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 362–371, 2023. 1
- [10] Ziqian Bai, Feitong Tan, Zeng Huang, Kripasindhu Sarkar, Danhang Tang, Di Qiu, Abhimitra Meka, Ruofei Du, Ming-song Dou, Sergio Orts-Escalano, Rohit Pandey, Ping Tan, Thabo Beeler, Sean Fanello, and Yinda Zhang. Learning personalized high quality volumetric head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [11] Shrisha Bharadwaj, Yufeng Zheng, Otmar Hilliges, Michael J. Black, and Victoria Fernandez Abrevaya. FLARE: Fast learning of animatable and relightable mesh avatars. *ACM Transactions on Graphics*, 42:15, 2023. 2
- [12] Ananta R Bhattachari, Matthias Nießner, and Artem Sevastopolsky. Triplanenet: An encoder for eg3d inversion. *arXiv preprint arXiv:2303.13497*, 2023. 3
- [13] Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn McPhail, Ravi Ramamoorthi, Yaser Sheikh, and Jason M. Saragih. Deep relightable appearance models for animatable faces. *ACM Transactions on Graphics (TOG)*, 40, 2021. 2
- [14] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 1
- [15] Volker Blanz and Thomas Vetter. *A Morphable Model For The Synthesis Of 3D Faces*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. 2
- [16] 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. 7
- [17] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13786–13795, 2020. 2
- [18] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shouo-I Yu, Yaser Sheikh, and Jason Saragih. Authentic volumetric avatars from a phone scan. *ACM Trans. Graph.*, 41, 2022. 2
- [19] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. Pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5799–5809, 2021. 3
- [20] 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. 3, 4, 5
- [21] Chuhan Chen, Matthew O’Toole, Gaurav Bharaj, and Pablo Garrido. Implicit neural head synthesis via controllable local deformation fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 416–426, 2023. 2

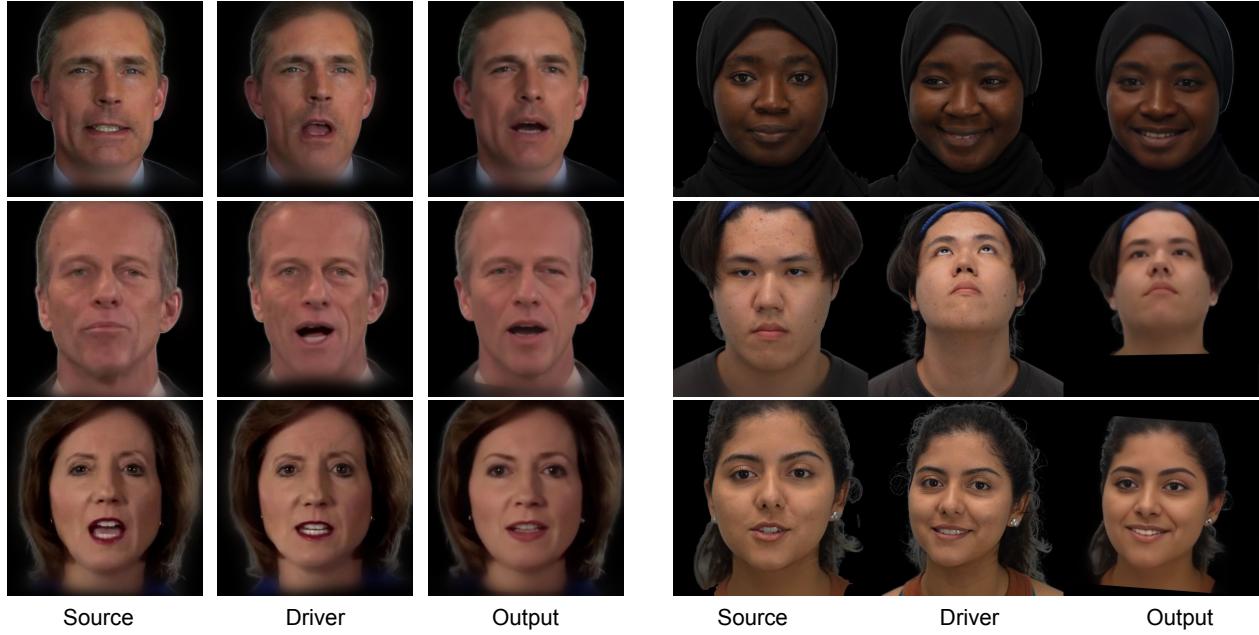


Figure 10. Qualitative results of our method on self-reenactment

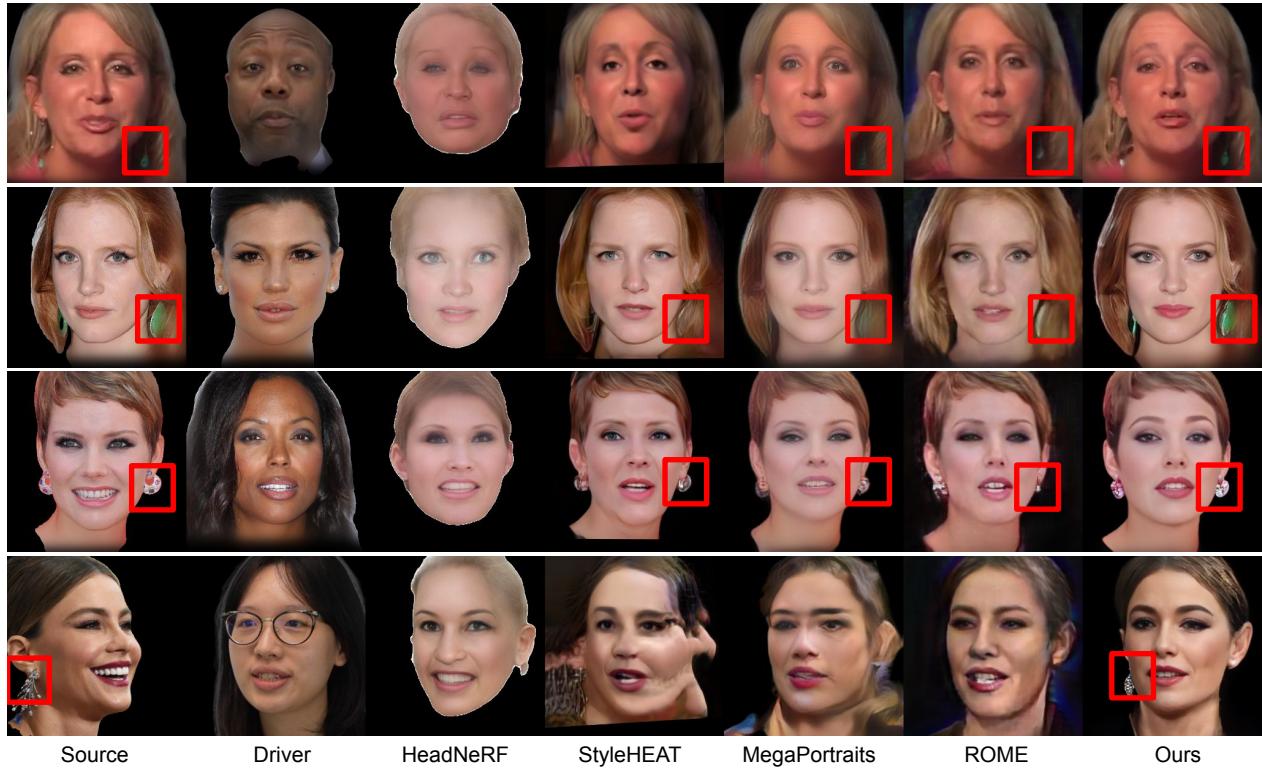


Figure 11. Our method faithfully retains the jewelries from the source image

- [22] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

- [23] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–



Figure 12. Additional Limitations: our method cannot handle the driver’s tongue and sometimes produces wrong accessories that are out-of-domain, such as exotic sunglasses. Also, our head pose uses a single rigid transformation instead of a multi-joint body rig, which leads to the shoulders always moving together with the head pose.



Figure 13. Our method can handle glass’s refraction

20322, 2022. 7, 1

- [24] 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. 5
- [25] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10673–10683, 2022. 3
- [26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An im-



Figure 14. Qualitative comparisons with LPR [55] on HDTF dataset.

age is worth  $16 \times 16$  words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 2

- [27] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoria Sharmanska. Headgan: One-shot neural head synthesis and editing. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2
- [28] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. *arXiv preprint arXiv:2207.07621*, 2022. 2, 3, 5, 7
- [29] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 3
- [30] Yonggan Fu, Yuecheng Li, Chenghui Li, Jason Saragih, Peizhao Zhang, Xiaoliang Dai, and Yingyan (Celine) Lin. Auto-card: Efficient and robust codec avatar driving for real-time mobile telepresence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21036–21045, 2023. 2
- [31] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 2
- [32] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d

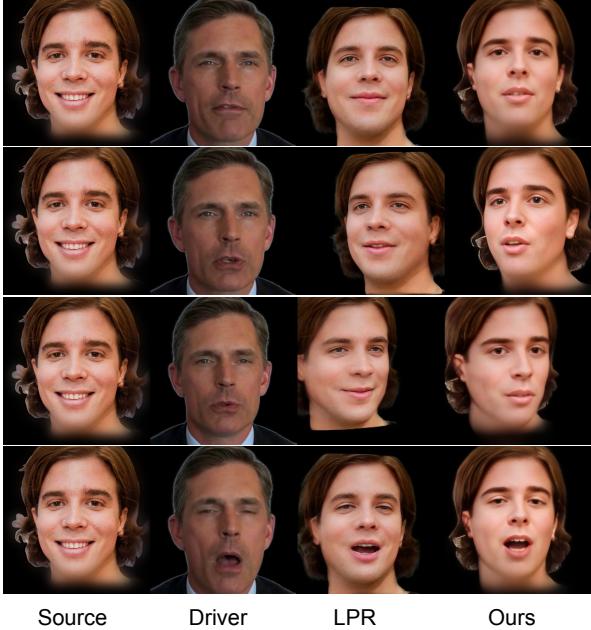


Figure 15. Novel view synthesis comparison with LPR. In this example, LPR fails to remove the smiling expression from the source while our method successfully transfer the expression from the driver to the source due to better disentanglement.

- facial avatar reconstruction. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2021. 2
- [33] Yue Gao, Yuan Zhou, Jinglu Wang, Xiao Li, Xiang Ming, and Yan Lu. High-fidelity and freely controllable talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5609–5619, 2023. 2
- [34] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos P Zafeiriou. Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1
- [35] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18653–18664, 2022. 2
- [36] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. 2022. 2
- [37] 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, 7
- [38] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jae-woo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Trans. Graph.*, 36(6), 2017. 1
- [39] Yiyu huang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. In *ECCV*, 2022. 3

- [40] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, 2022. 2
- [41] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 7, 2
- [42] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5
- [43] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3
- [44] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *European Conference of Computer vision (ECCV)*, 2022. 2, 3, 7
- [45] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Péerez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep video portraits. *ACM Transactions on Graphics 2018 (TOG)*, 2018. 2
- [46] Jaehoon Ko, Kyusun Cho, Daewon Choi, Kwangrok Ryoo, and Seungryong Kim. 3d gan inversion with pose optimization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2967–2976, 2023. 3
- [47] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos P Zafeiriou. Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1
- [48] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Fitme: Deep photorealistic 3d morphable model avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8629–8640, 2023. 1
- [49] Jason Lawrence, Dan B Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G. Desloge, Tommy Fortes, Eric M. Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, Claude Knaus, Brian Kuschak, Ricardo Martin-Brualla, Harris Nover, Andrew Ian Russell, Steven M. Seitz, and Kevin Tong. Project starline: A high-fidelity telepresence system. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 40(6), 2021. 2
- [50] Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 394–403, 2023. 1
- [51] Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. Facial performance sensing head-mounted display. *ACM Transactions on Graphics (Proceedings SIGGRAPH 2015)*, 34(4), 2015. 2

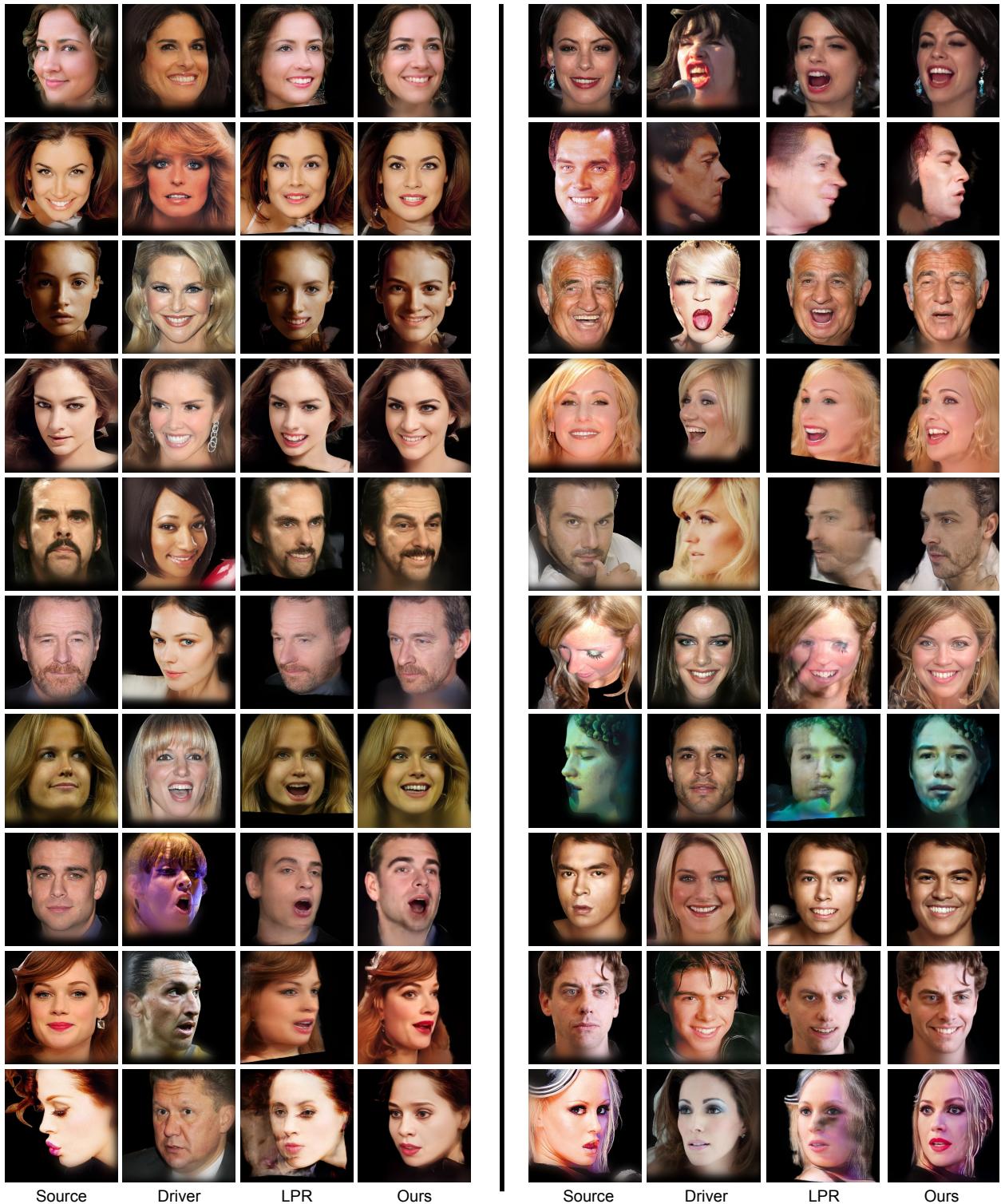


Figure 16. Qualitative comparisons with LPR [55] on CelebA-HQ dataset.

- [52] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*,

- (Proc. SIGGRAPH Asia), 36(6):194:1–194:17, 2017. 3  
[53] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and ex-



Figure 17. Synthesizing novel views using our method.

pression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 2, 7

- [54] Weichuang Li, Longhao Zhang, Dong Wang, Bin Zhao, Zhigang Wang, Mulin Chen, Bang Zhang, Zhongjian Wang, Liefeng Bo, and Xuelong Li. One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17969–17978, 2023. 2, 3

- [55] Xuetong Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot neural head avatar. *arXiv preprint arXiv:2306.08768*, 2023. 2, 3, 4, 5, 7, 8

- [56] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv*



Figure 18. Qualitative results on various datasets.

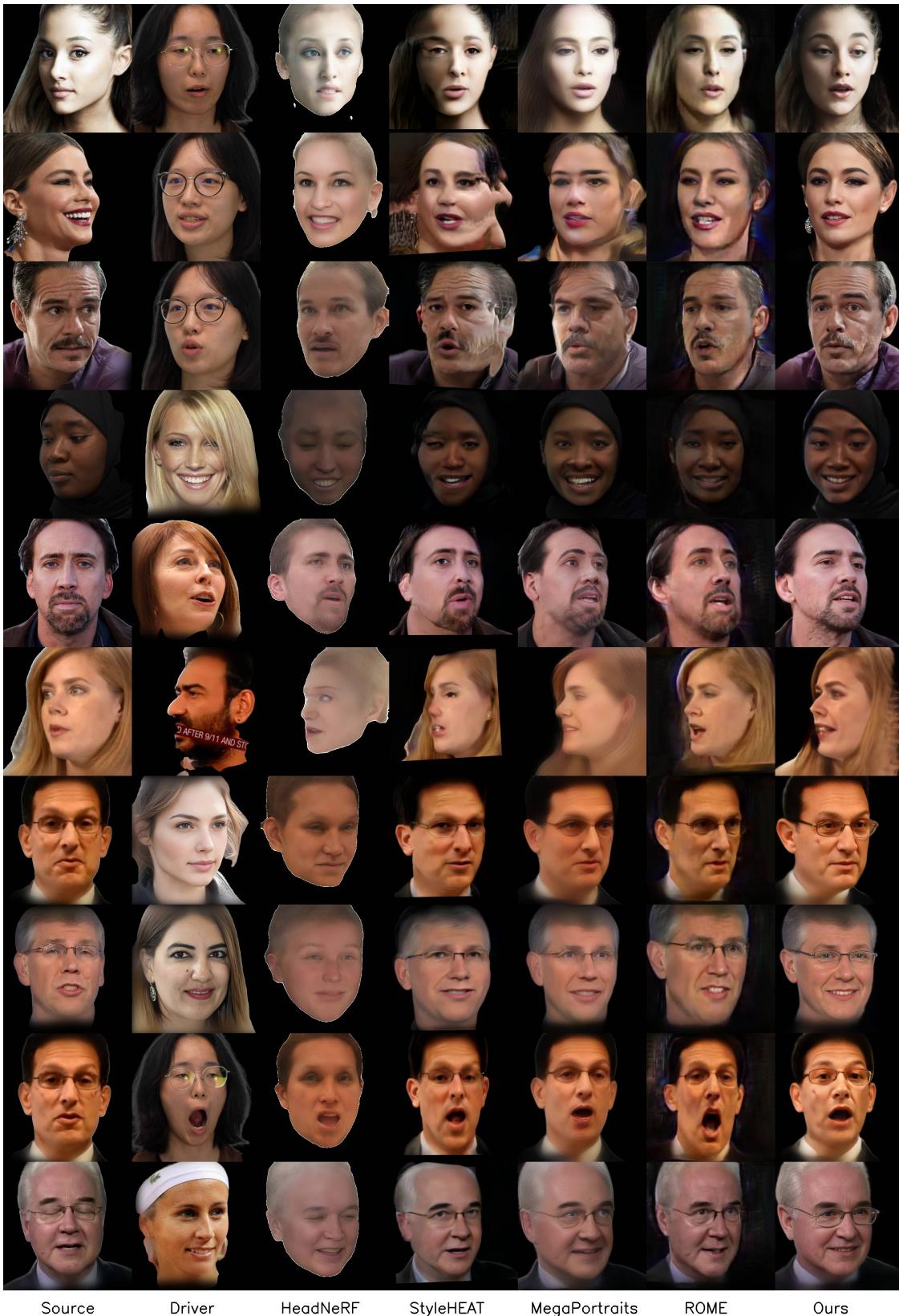


Figure 19. Qualitative results on various datasets.



Figure 20. Qualitative results on various datasets.



Figure 21. Qualitative results on various datasets.



Figure 22. Qualitative results on various datasets.



Figure 23. Qualitative results on various datasets.



Figure 24. Qualitative results on various datasets.

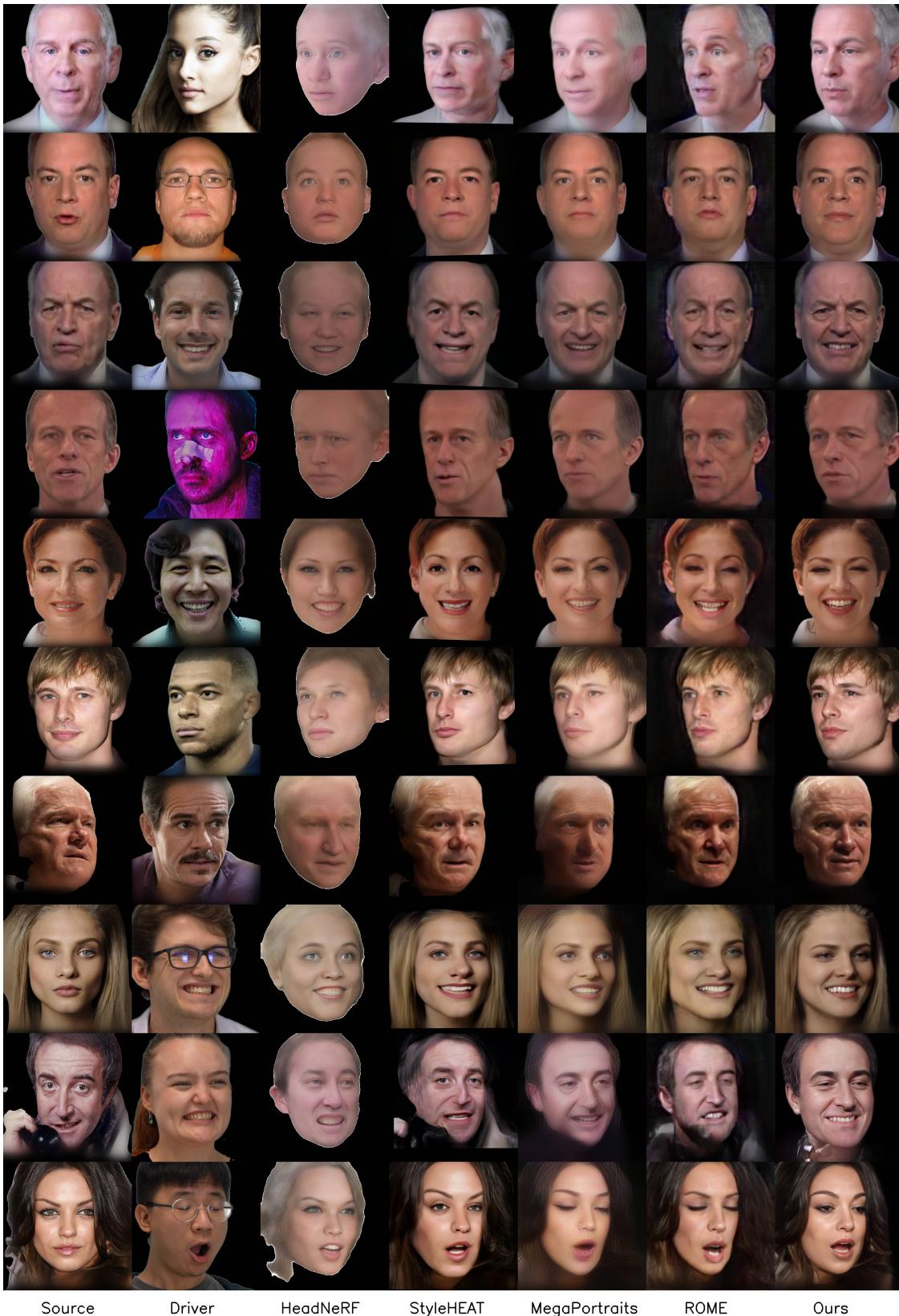


Figure 25. Qualitative results on various datasets.



Figure 26. Qualitative results on various datasets.



Figure 27. Qualitative results on various datasets.

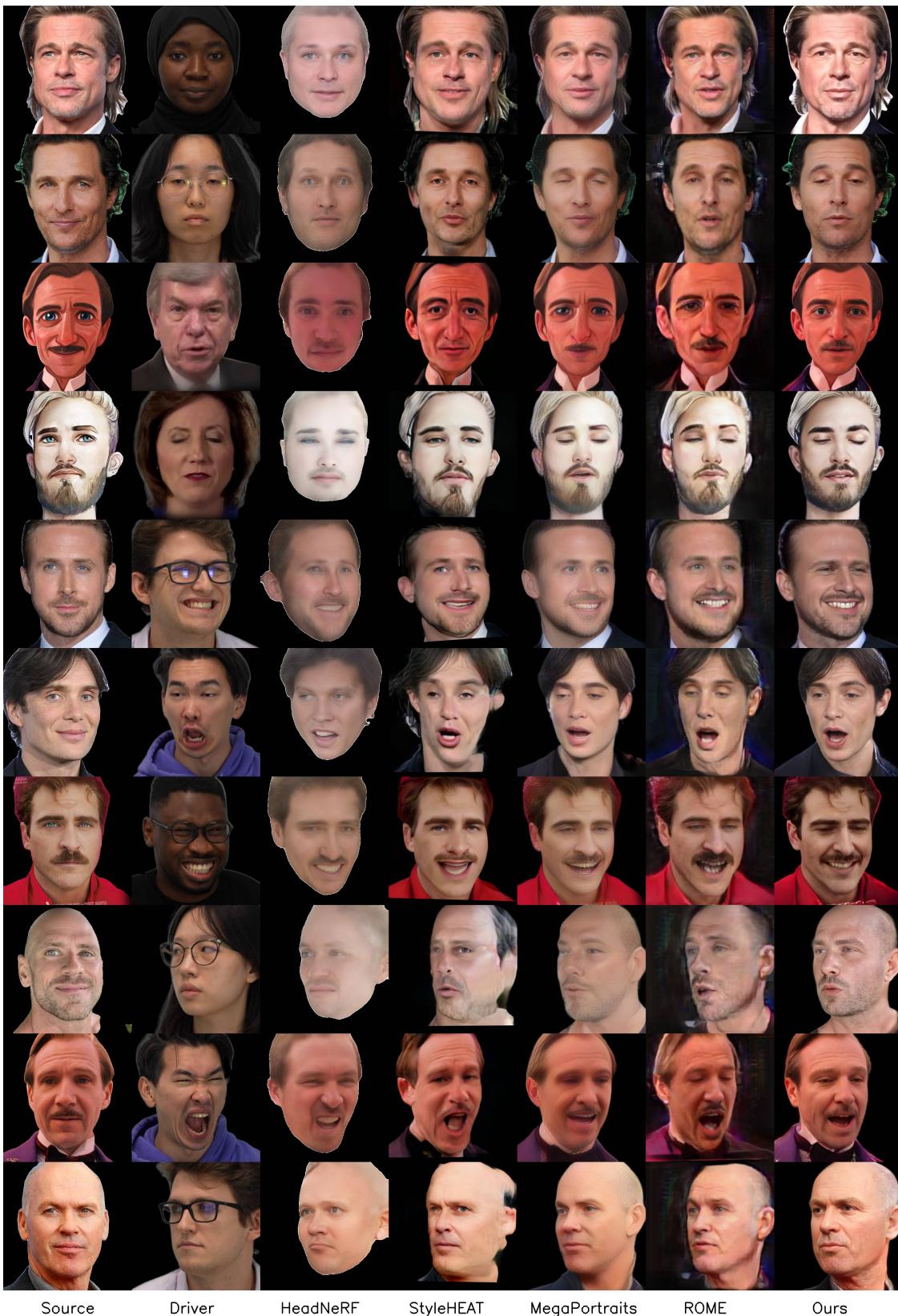


Figure 28. Qualitative results on various datasets.



Figure 29. Qualitative results on various datasets.



Figure 30. Qualitative results on various datasets.



Figure 31. Qualitative results on various datasets.

- preprint arXiv:1705.02894, 2017.* 2
- [57] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Trans. Graph.*, 37(4), 2018. 2
- [58] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 5
- [59] Huiwen Luo, Koki Nagano, Han-Wei Kung, Mclean Goldwhite, Qingguo Xu, Zejian Wang, Lingyu Wei, Liwen Hu, and Hao Li. Normalized avatar synthesis using stylegan and perceptual refinement. *CorR*, abs/2106.11423, 2021. 1
- [60] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De la Torre, and Yaser Sheikh. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 64–73, 2021. 2

- [61] Zhiyuan Ma, Xiangyu Zhu, Guo-Jun Qi, Zhen Lei, and Lei Zhang. Otavator: One-shot talking face avatar with controllable tri-plane rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16910, 2023. 2, 3, 7
- [62] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020. 4
- [63] 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. *Communications of the ACM*, 65(1):99–106, 2021. 3
- [64] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature

- fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11453–11464, 2021. 3
- [65] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7184–7193, 2019. 2
- [66] Kyle Olszewski, Joseph J. Lim, Shunsuke Saito, and Hao Li. High-fidelity facial and speech animation for vr hmds. *ACM Transactions on Graphics (TOG)*, 35:1 – 14, 2016. 2
- [67] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13503–13513, 2022. 3
- [68] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H. Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13759–13768, 2021. 2
- [69] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3
- [70] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021. 2, 3
- [71] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. *Advances in Neural Information Processing Systems*, 34:17480–17492, 2021. 2
- [72] Gabriel Schwartz, Shih-En Wei, Te-Li Wang, Stephen Lombardi, Tomas Simon, Jason Saragih, and Yaser Sheikh. The eyes have it: An integrated eye and face model for photorealistic facial animation. *ACM Trans. Graph.*, 39(4), 2020. 2
- [73] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020. 3
- [74] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. *Advances in Neural Information Processing Systems*, 35:33999–34011, 2022. 3
- [75] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *CVPR*, 2019. 2
- [76] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.
- [77] Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021. 2
- [78] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. EpiGRAF: Rethinking training of 3d GANs. In *Advances in Neural Information Processing Systems*, 2022. 3
- [79] Linsen Song, Wayne Wu, Chaoyou Fu, Chen Qian, Chen Change Loy, and Ran He. Pareidolia face reenactment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [80] Jiale Tao, Biao Wang, Borun Xu, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. Structure-aware motion transfer with deformable anchor model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3637–3646, 2022. 2
- [81] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2016. 2
- [82] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Headon: Real-time reenactment of human portrait videos. *ACM Transactions on Graphics 2018 (TOG)*, 2018. 2
- [83] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 3
- [84] Alex Trevithick, Matthew Chan, Michael Stengel, Eric Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. Real-time radiance fields for single-image portrait view synthesis. *ACM Transactions on Graphics (TOG)*, 42(4):1–15, 2023. 2, 3, 4, 7, 1
- [85] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 2
- [86] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5
- [87] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9168–9178, 2021. 4, 1
- [88] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *International Conference on Learning Representations*, 2022. 2
- [89] O. Wiles, A.S. Koepke, and A. Zisserman. X2face: A network for controlling face generation by using images, audio, and pose codes. In *European Conference on Computer Vision*, 2018. 2
- [90] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2195–2205, 2023. 3
- [91] Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 3, 1
- [92] Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. High-fidelity 3d gan inversion by pseudo-multi-view optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 321–331, 2023. 3

- [93] Hongyi Xu, Guoxian Song, Zihang Jiang, Jianfeng Zhang, Yichun Shi, Jing Liu, Wanchun Ma, Jiashi Feng, and Linjie Luo. Omniaavatar: Geometry-guided controllable 3d head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12814–12824, 2023. 3
- [94] Yuelang Xu, Hongwen Zhang, Lizhen Wang, Xiaochen Zhao, Han Huang, Guojun Qi, and Yebin Liu. Latentavatior: Learning latent expression code for expressive neural head avatar. In *ACM SIGGRAPH 2023 Conference Proceedings*. Association for Computing Machinery, 2023. 2
- [95] Zhongcong Xu, Jianfeng Zhang, Junhao Liew, Wenqing Zhang, Song Bai, Jiashi Feng, and Mike Zheng Shou. Pv3d: A 3d generative model for portrait video generation. In *The Tenth International Conference on Learning Representations*, 2023. 3
- [96] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. Giraffe hd: A high-resolution 3d-aware generative model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18440–18449, 2022. 3
- [97] Kewei Yang, Kang Chen, Daoliang Guo, Song-Hai Zhang, Yuan-Chen Guo, and Weidong Zhang. Face2face  $\rho$ : Real-time high-resolution one-shot face reenactment. 2022. 2
- [98] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *ECCV*, 2022. 2, 7
- [99] Yu Yin, Kamran Ghasedi, HsiangTao Wu, Jiaolong Yang, Xin Tong, and Yun Fu. Nerfinvertor: High fidelity nerf-gan inversion for single-shot real image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8539–8548, 2023. 3
- [100] Wangbo Yu, Yanbo Fan, Yong Zhang, Xuan Wang, Fei Yin, Yunpeng Bai, Yan-Pei Cao, Ying Shan, Yang Wu, Zhongqian Sun, and Baoyuan Wu. Nofa: Nerf-based one-shot facial avatar reconstruction. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 2, 3
- [101] Ziyang Yuan, Yiming Zhu, Yu Li, Hongyu Liu, and Chun Yuan. Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding. *arXiv preprint arXiv:2303.12326*, 2023. 3
- [102] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019. 2, 7
- [103] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference on Computer Vision*, pages 524–540. Springer, 2020. 2
- [104] Bowen Zhang, Chenyang Qi, Pan Zhang, Bo Zhang, HsiangTao Wu, Dong Chen, Qifeng Chen, Yong Wang, and Fang Wen. Metaportrait: Identity-preserving talking head generation with fast personalized adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22096–22105, 2023. 2
- [105] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Fdnerf: Few-shot dynamic neural radiance fields for face reconstruction and expression editing. *arXiv preprint arXiv:2208.05751*, 2022. 2
- [106] 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, 7
- [107] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 7, 2
- [108] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *CVPR*, pages 3657–3666, 2022. 2
- [109] Xiaochen Zhao, Lizhen Wang, Jingxiang Sun, Hongwen Zhang, Jinli Suo, and Yebin Liu. Havatar: High-fidelity head avatar via facial model conditioned neural radiance field. *ACM Trans. Graph.*, 2023. 2
- [110] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühlér, Xu Chen, Michael J. Black, and Otmar Hilliges. I m avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13545–13555, 2022.
- [111] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. Pointavator: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [112] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer, 2022. 5
- [113] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, 2022. 2, 5, 7, 1
- [114] Wojciech Zienolka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2