

One-Shot High-Fidelity Talking-Head Synthesis with Deformable Neural Radiance Field

Weichuang Li^{1*}
Mulin Chen^{1,3}

Longhao Zhang^{2*}
Bang Zhang²

Dong Wang^{1*}
Zhongjian Wang²

Bin Zhao^{1,3†}
Liefeng Bo²

Zhigang Wang¹
Xuelong Li^{1,3 †}

¹Shanghai AI Laboratory ²DAMO Academy, Alibaba Group ³Northwestern Polytechnical University
{liweichuang, wangdong, zhaobin}@pjlab.org.cn, longhao.zlh@alibaba-inc.com, li@nwpu.edu.cn

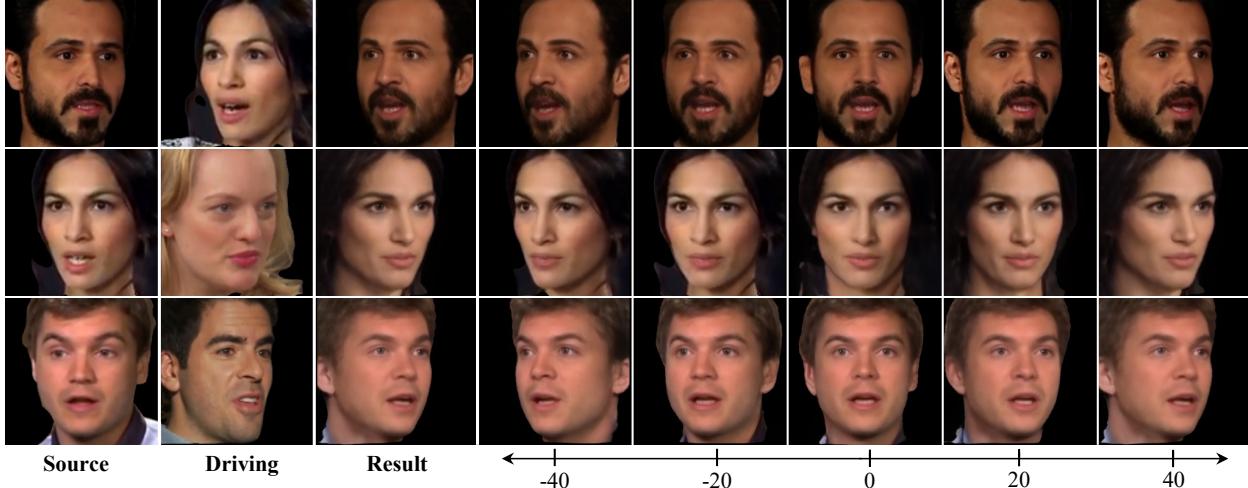


Figure 1. **Representative results of our method.** The first three columns exhibit the source, driving, generated images, respectively. The rest columns show the exploration of the generated images to different yaw angles.

Abstract

Talking head generation aims to generate faces that maintain the identity information of the source image and imitate the motion of the driving image. Most pioneering methods rely primarily on 2D representations and thus will inevitably suffer from face distortion when large head rotations are encountered. Recent works instead employ explicit 3D structural representations or implicit neural rendering to improve performance under large pose changes. Nevertheless, the fidelity of identity and expression is not so desirable, especially for novel-view synthesis. In this paper, we propose HiDe-NeRF, which achieves high-fidelity and free-view talking-head synthesis. Drawing on the recently proposed Deformable Neural Radiance Fields, HiDe-NeRF represents the 3D dynamic scene into a canonical appearance field and an implicit deformation field, where the former comprises the canonical source face and the latter models the driving pose and expression. In particular, we improve fidelity from two aspects: (i) to enhance identity expressiveness, we design a generalized appearance module

that leverages multi-scale volume features to preserve face shape and details; (ii) to improve expression precisionness, we propose a lightweight deformation module that explicitly decouples the pose and expression to enable precise expression modeling. Extensive experiments demonstrate that our proposed approach can generate better results than previous works. Project page: <https://www.waytron.net/hidenerf/>

1. Introduction

Talking-head synthesis aims to preserve the identity information of the source image and imitate the motion of the driving image. Synthesizing talking faces of a given person driven by other speaker is of great importance to various applications, such as film production, virtual reality, and digital human. Existing talking head methods are not capable of

* denotes equal contribution

† denotes the corresponding authors

generating high-fidelity results, they cannot precisely preserve the source identity or mimic the driving expression.

Most pioneering approaches [11, 17, 25, 36, 38, 45, 45] learn source-to-driving motion to warp the source face to the desired pose and expression. According to the warping types, previous works can be roughly divided into: 2D warping-based methods, mesh-based methods, and neural rendering-based methods. 2D warping-based methods [17, 36, 38] warps source feature based on the motion field estimated from the sparse keypoints. However, these methods encounter the collapse of facial structure and expression under large head rotations. Moreover, they cannot fully disentangle the motion with identity information of the driving image, resulting in a misguided face shape. Mesh-based methods [13] are proposed to tackle the problem of facial collapse by using 3D Morphable Models (3DMM) [4, 31] to explicitly model the geometry. Limited by non-rigid deformation modeling ability of 3DMM, such implementation leads to rough and unnatural facial expressions. Besides, it ignores the influence of vertex offset on face shape, resulting in low identity fidelity. With the superior capability in multi-view image synthesis of Neural Radiance Fields (NeRF) [26], a concurrent work named FNeVR [48] takes the merits of 2D warping and 3D neural rendering. It learns a 2D motion field to warp the source face and utilizes volume rendering to refine the warped features to obtain final results. Therefore, it inherits the same problem as other warping-based methods.

To address above issues and improve the fidelity of talking head synthesis, we propose **H**igh-fidelitiy and **D**eformable NeRF, dubbed HiDe-NeRF. Drawing on the idea of recently emerged Deformable NeRF [1, 28, 29]. HiDe-NeRF represents the 3D dynamic scene into a canonical appearance field and an implicit deformation field. The former is a radiance field of the source face in canonical pose, and the latter learns the backward deformation for each observed point to shift them into the canonical field and retrieve their volume features for neural rendering. On this basis, we devise a *Multi-scale Generalized Appearance module (MGA)* to ensure identity expressiveness and a *Lightweight Expression-aware Deformation module (LED)* to improve expression preciseness.

To elaborate, *MGA* encodes the source image into multi-scale canonical volume features, which integrate high-level face shape information and low-level facial details, for better identity preservation. We employ the tri-plane [7, 32] as volume feature representation in this work for two reasons: (i) it enables generalization across 3D representations; (ii) it is fast and scales efficiently with resolution, facilitating us to build hierarchical feature structures. Moreover, we modify the ill-posed tri-plane representation by integrating a camera-to-world feature transformation, so that we can extract the planes from the source image with full control of

identity. This distinguishes our model from those identity-uncontrollable approaches [3, 7] that generate the planes from noise with StyleGAN2-based generators [21]. Notably, the *MGA* enables our proposed HiDe-NeRF to be implemented in a subject-agnostic manner, breaking the limitation that existing Deformable NeRFs can only be trained for a specific subject.

The deformations in talking-head scenes could be decomposed into the global pose and local expression deformation. The former is rigid and easy to handle, while the latter is non-rigid and difficult to model. Existing Deformable NeRFs predict them as a whole, hence failing to capture precise expression. Instead, our proposed *LED* could explicitly decouple the expression and pose in deformation prediction, thus significantly improving the expression fidelity. Specifically, it uses a pose-agnostic expression encoder and a position encoder to obtain the latent expression embeddings and latent position embeddings, where the former models the expression independently and the latter encodes positions of points sampled from rays under arbitrary observation views. Then, a deformation decoder takes the combination of two latent embeddings as input and outputs point-wise deformation. In this way, our work achieves precise expression manipulation and maintains expression consistency for free-view rendering (as shown in Fig. 1).

To summarize, the contributions of our approach are:

- Firstly, we introduce the HiDe-NeRF for high-fidelity and free-view talking head synthesis. To the best of our knowledge, HiDe-NeRF is the first *one-shot* and *subject-agnostic* Deformable Neural Radiance Fields.
- Secondly, we propose the *Multi-scale Generalized Appearance module (MGA)* and the *Lightweight Expression-aware Deformation module (LED)* to significantly improve the fidelity of identity and expression in talking-head synthesis.
- Lastly, extensive experiments demonstrate that our proposed approach can generate more realistic results than state-of-the-art in terms of capturing the driving motion and preserving the source identity information.

2. Related Work

Talking-Head Synthesis. Previous talking-head synthesis methods can be divided into warping-based, mesh-based, NeRF-based methods. Warping-based methods [10–12, 16, 23, 50] are the most popular methods among 2D generation methods [2, 6, 14, 33, 47]. These methods warp the source features by estimated motion field to transport driving pose and expression into source face. For instance, Monkey-Net [37] builds a 2D motion field from the sparse keypoints detected by an unsupervised trained detector. DaGAN [17] incorporates the depth estimation to supplement the missing 3D geometry information in 2D motion field.

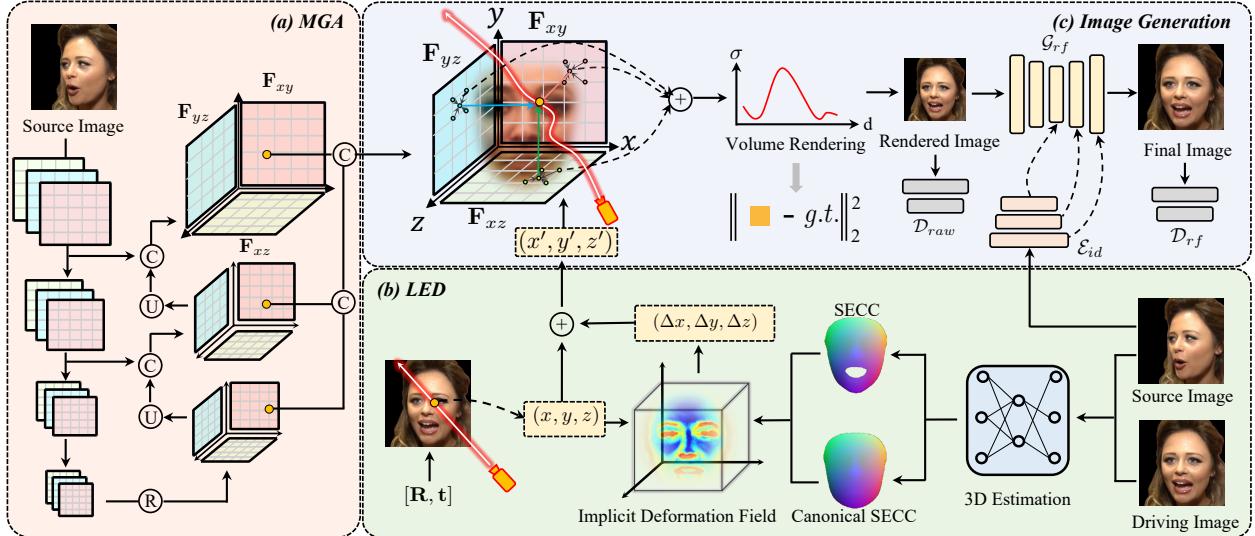


Figure 2. Illustration of the proposed approach. The *MGA* (Pink) encodes the source image into multi-scale canonical volume features. Notably, the skip connection is devised to preserve hierarchical tri-plane properties. The *LED* (Green) predicts backward deformation for each observed point to shift them into the canonical space and retrieve their volume features. The deformation is learned from paired SECCs, conditioned on the positions of points sampled from the rays. The image generation module (Blue) takes as input the deformed points to sample features from different scales of tri-planes. These multi-scale features are composed for the following neural rendering. Here we also design a refine network to further refine the texture details (e.g., teeth, skin, hair, etc.), and to enhance the resolution of rendered images. Notable, both rendered images and refined images are supervised by reconstruction loss and adversarial loss. The symbol \odot , \circledcirc , \circledR , \oplus indicates channel-wise concatenation, upsample, resize and upsample, element-wise sum, respectively.

OSFV [45] tries to extract 3D appearance features and predict a 3D motion field for free-view synthesis. Some conventional works [39, 42–44] employ 3D Morphable Models (3DMM) [4, 31], which support a diverse range of animations via disentangled shape, expression and rigid motions. StyleRig [40] and PIE [41] are proposed to exploit the semantic information in the latent space of StyleGAN [20] and modulate the expressions using 3DMM. PIRender [34] uses 3DMM to predict the flow and warp the source image. ROME [22] is the first 3DMM-based method that uses a single image to create realistic photo in a rigged mesh format. It learns the offset for each mesh vertex and renders the rigged mesh with predicted neural texture. Moreover, it introduces a U-Net generator with adversarial loss to refine the rendered images. We also use 3DMM in this work, but unlike other approaches, we use it to generate the Shape- and Expression-aware Coordinate Code (SECC) to build our deformation field, which is detailed in Sec. 3.2. NeRF is a recently proposed implicit 3D scene representation method that renders the static scene with points along different view directions. NeRF first flourished in the audio-driven approaches [15, 24, 35, 46], as they can easily combined with the latent code learned from the audio. But they are restricted to subject-dependent generation.

Deformable Neural Radiance Field. Deformable NeRF is proposed for rendering of dynamic scene. NeRFies [28] maps each observed point into a canonical space through

a continuous deformation field represented by a scene-specific MLP, and it can only handle small non-rigid movements. HyperNeRF [29] inherits NeRFies and uses an ambient dimension to model the topological changes in the deformation field. Conversely, RigNeRF [1] uses 3DMM to learn the deformation by finding the closest driving mesh vertex for each sampled point and explicitly calculating its distance from the corresponding canonical mesh vertex, which is not efficient or precise enough. Notably, all the above-mentioned Deformable NeRFs are subject-dependent, which implies they need to train an individual model for each subject. Besides, they treat the global pose and local expression deformation as a whole, thus cannot achieve precise expression manipulation. In this work, we draw on the idea of Deformable NeRF and propose the HiDe-NeRF. Different from existing methods, our approach is an *one-shot* and *subject-agnostic* Deformable NeRF specially designed for talking head synthesis. Besides, our *LED*-based deformation field is computationally much more efficient than other MLP-based deformation fields.

3. Methods

In this section, we describe our method, HiDe-NeRF, that enables high-fidelity talking head synthesis. The overall procedure can be illustrated in Fig. 2. Then, we expound the proposed *Multi-scale Generalized Appearance module (MGA)* and *Lightweight Expression-aware Deformation*

module (LED) in Sec. 3.1 and Sec. 3.2. Afterward, we describe the image generation module in Sec. 3.3, including volume rendering and texture refinement. The training details can be found in the Supplementary Materials.

3.1. Multi-Scale Generalized Appearance Module

As discussed, Deformable NeRF [1, 28, 29] typically targets novel view synthesis based on multi-view inputs under a single-scene scenario. Different from the settings of conventional Deformable NeRF, talking-head synthesis aims to generate high-fidelity images of any particular person with a single image, which can be summarized as subject-agnostic and high-fidelity preservation under one-shot setting.

In this work, we introduce tri-plane representation as our appearance field to accommodate the attribute of subject-agnostic. Tri-plane hybrid representation [7, 32] is recently proposed, which builds three orthogonal planes with feature maps. Given a 3D point $\mathbf{p} = (x, y, z)$, it is projected onto $\mathbf{F}_{xy}, \mathbf{F}_{xz}, \mathbf{F}_{yz}$ three planes to query the feature vectors via bilinear interpolation. The queried features from three planes are averaged as the representation of point $F(\mathbf{p}) = (F_{xy}(\mathbf{p}) + F_{xz}(\mathbf{p}) + F_{yz}(\mathbf{p})) / 3$, where $F_{ij} : \mathbb{R}^3 \mapsto \mathbb{R}^C$ denotes sampling the feature of 3D coordinates from the planar feature map \mathbf{F}_{ij} . Although the tri-plane representation has been applied in different areas [3, 18, 52], there is no method that directly extract planes from an image. We find one core issue in learning tri-plane representation from the image is that the camera coordinate system and the world coordinate system are oriented in different directions. The definition of tri-plane is based on the world coordinate system, but the axes of the image are aligned with the camera coordinate system. Due to this problem, the predicted volume features $\{\mathbf{V}_i, i = 1, 2, 3\}$, where $\mathbf{V}_i \in \mathbb{R}^{c \times h \times w}$ from deep network are mismatched with the definition of tri-plane representation, which makes the representation difficult to learn from the image directly. Therefore, we use source camera parameters $\{\mathbf{R}_{src}, \mathbf{t}_{src}\}$ to transfer the predicted volume features into tri-plane representation. Concretely, this transformation could be formulated as below,

$$\mathbf{F}_{plane} = \mathcal{T}(\{\mathbf{R}_{src}, \mathbf{t}_{src}\}, \mathbf{V}_i), i = 1, 2, 3, \quad (1)$$

where \mathcal{T} denotes the camera-to-world transformation function and $plane \in [xy, xz, yz]$.

Another challenging issue of talking-head generation is preserving identity fidelity under the one-shot setting. It is known that different scales of feature provide different information, high-level feature maps comprise facial shape information while low-level feature maps contain facial details, for instance, skin texture, makeup, etc. To improve the expressiveness of tri-plane representation, we adopt multi-scale tri-plane representation instead, which integrates different levels of semantic information. As shown in Fig. 2(a), we first employ a deep feature extractor to

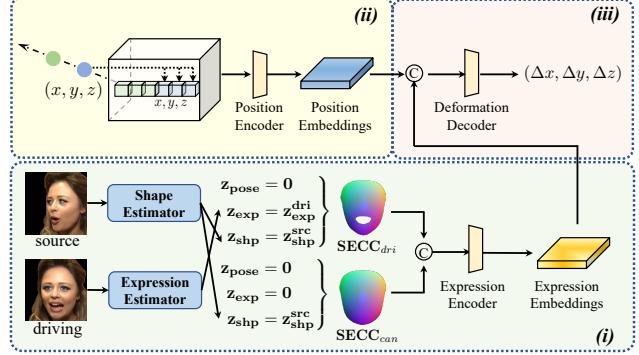


Figure 3. **The illustration of the proposed LED.** The symbol \odot indicates channel-wise concatenation.

derive the pyramid feature maps $[\mathbf{M}^0, \dots, \mathbf{M}^n]$ from the source image I_{src} . For the lowest-resolution feature map \mathbf{M}^0 , a small convolutional decoder ψ^0 is used to predict volume features \mathbf{V}^0 , and the corresponding lowest-scale tri-plane representation \mathbf{F}^0 is obtained by applying camera-to-world transformation in Eq. 1. Based on this, the multi-scale tri-plane representations are formulated as:

$$\begin{aligned} \mathbf{V}_k^{j+1} &= \psi_k^{j+1}([\mathbf{M}_k^{j+1}, \mathbf{F}_k^j \uparrow]), \\ \mathbf{F}_k^{j+1} &= \mathcal{T}(\{\mathbf{R}_{src}, \mathbf{t}_{src}\}, \mathbf{V}_k^{j+1}), \end{aligned} \quad (2)$$

where $k \in \{1, 2, 3\}$ represent different planes, ψ_k^{j+1} denotes a convolutional network, \uparrow denotes the up-sampling operation. \mathbf{V}_k^j is the j -th level of the predicted volume feature. Based on this multi-scale tri-plane representation, the final representation of a single point \mathbf{p} can be formulated as $F(\mathbf{p}) = [F^0(\mathbf{p}), \dots, F^n(\mathbf{p})]$, where $[\dots]$ denotes channel-wise concatenation.

3.2. Lightweight Expression-Aware Deformation Module

The deformations in talking-head scenes could be decomposed into the global rigid pose and local non-rigid expression deformation. Existing Deformable NeRFs predict them as a whole, hence failing to accurately model complex and delicate expressions. In this work, we propose the *Lightweight Expression-aware Deformation Module (LED)*, which explicitly decouples the expression and pose in deformation prediction, significantly improving the expression fidelity. Moreover, the decoupling of expression and pose ensures expression consistency for free-view rendering. As illustrated in Fig. 3, the *LED* could be divided into three steps:

(i) **Expression Encoding.** First, we introduce the Shape-and Expression-aware Coordinate Code (SECC) to learn the pose-agnostic expression deformation for precise expression manipulation. SECC is obtained by rendering a 3DMM face [31] through Z-Buffer with Normalized Coordinate Code (NCC) [51] as its colormap. It could be

formulated as:

$$\begin{aligned} \text{SECC} &= \text{Z-Buffer}(V_{3d}(\mathbf{p}), \text{NCC}), \\ V_{3d}(\mathbf{p}) &= \mathbf{R}(\bar{\mathbf{S}} + \mathbf{A}_{shp}\mathbf{z}_{shp} + \mathbf{A}_{exp}\mathbf{z}_{exp}) + \mathbf{t}, \end{aligned} \quad (3)$$

where $\bar{\mathbf{S}}$ is the template shape, \mathbf{A}_{shp} and \mathbf{A}_{exp} are principle axes for shape and expression. We set $\mathbf{R} = \mathbf{1}$ and $\mathbf{t} = \mathbf{0}$ to eliminate the pose.

As shown in Fig. 3(i), we use a pair of SECC to model the shape-aware expression changes from driving to canonical. Specifically, we predict the expression coefficient \mathbf{z}_{exp}^{dri} from the driving image \mathbf{I}_{dri} and the shape coefficient \mathbf{z}_{shp}^{src} from the source image \mathbf{I}_{src} using the 3D Estimator [9], and we form the driving $\text{SECC}_{dri} \in \mathbb{R}^{H \times W \times 3}$ with source shape \mathbf{z}_{shp}^{src} and driving expression \mathbf{z}_{exp}^{dri} , and the canonical $\text{SECC}_{can} \in \mathbb{R}^{H \times W \times 3}$ with source shape \mathbf{z}_{shp}^{src} and zero expression \mathbf{z}_{exp}^0 . Since the *rgb* value of each point in NCC [51] corresponds to the *xyz* coordinate of a specific mesh vertex, it establishes the vertex-to-pixel correspondence between 3D and 2D. Therefore, we directly apply a 2D convolutional encoder on paired SECCs to learn the latent expression embeddings that contains 3D expression deformation.

(ii) Position Encoding. In order to learn the point-wise deformation under the observation view (can be the driving image view or arbitrary views), we encode the 3D coordinate of points sampled from the rays as the positional condition. Specifically, we first reshape the points $\mathbf{P} \in \mathbb{R}^{H \times W \times N \times 3}$ to $\mathbf{P} \in \mathbb{R}^{H \times W \times (3 \times N)}$, where N is the number of sampled points along each ray, H and W denote the rendering resolution. It is then fed into a fully-convolutional position encoder to get the latent position embeddings.

(iii) Deformation Prediction. The latent expression embeddings and the latent position embeddings are concatenated in channel-wise and fed into a deformation decoder to predict the point-level deformation $\Delta\mathbf{P} \in \mathbb{R}^{(H \times W) \times (3 \times N)}$.

In summary, for a deformation module parameterized by Φ , the implicit function can be formulated as:

$$\mathcal{F}_\Phi^{\text{deform}} : (\mathbf{P}, \text{SECC}_{dri}, \text{SECC}_{can}) \rightarrow \Delta\mathbf{P}. \quad (4)$$

As mentioned above, our proposed *LED* employs the vertex-to-pixel correspondence and the positional encoding to learn point-wise 3D deformations. It is lightweight yet efficient since it doesn't need to find the closest driving mesh vertex for each sampled point and explicitly calculate its distance from the corresponding canonical mesh vertex like [1]. Besides, the encoder and decoder network in *LED* are fully-convolutional and very shallow, thus is computationally much more efficient than other MLP-based deformation fields [1, 28, 29].

3.3. Image Generation Module

The image generation module is composed of volume rendering and texture refinement.

Volume Rendering. Given camera intrinsic parameters and camera pose of driving image, we calculate the view direction \mathbf{d} of a pixel coordinate (h, w) . We first sample N points along this ray for stratified sampling. Formally, let $\mathbf{p}_i = \mathbf{o} + t_i \mathbf{d}, i \in \{1, \dots, N\}$ denote the sampling points on the ray given camera origin \mathbf{o} , and t_i corresponds to the depth value of \mathbf{p}_i along this ray. For every point \mathbf{p}_i , we first apply positional encoding $\gamma(q) = \langle \sin(2^l \pi q), \cos(2^l \pi q) \rangle$ to it, and sample volume feature $F(\mathbf{p}'_i)$ from multi-scale tri-planes at deformed point position $\mathbf{p}'_i = \mathbf{p}_i + \Delta\mathbf{p}_i$, where $\Delta\mathbf{p} \in \mathbb{R}^{3 \times N}$ is the $(h \times H + w)$ -th row vector of $\Delta\mathbf{P}$. Then $F(\mathbf{p}'_i)$ and $\gamma(\mathbf{p}_i)$ are concatenated as $\mathbf{f}(\mathbf{p}_i) = [F(\mathbf{p}'_i), \gamma(\mathbf{p}_i)]$ and fed into a two-layer MLP to predict color \mathbf{c} and density σ of point \mathbf{p}_i . Finally, the color of this pixel can be rendered as:

$$\begin{aligned} \mathbf{C}(\mathbf{r}) &= \int_{t_n}^{t_f} T(t) \sigma(\mathbf{r}(t)) \mathbf{f}(\mathbf{r}(t)) dt, \\ T(t) &= \exp \left(- \int_{t_n}^{t_f} \sigma(\mathbf{r}(s)) ds \right), \end{aligned} \quad (5)$$

where t_n and t_f indicate near and far bounds along the ray.

Texture Refinement. Since the texture details (e.g., teeth, skin, hair, etc.) and the resolution of the rendered images \mathbf{I}_{raw} are limited, we design an refine network \mathcal{G}_{rf} with encoder-decoder structure to improve them and generate the final image \mathbf{I}_{rf} . Specifically, we use an identity extractor \mathcal{E}_{id} to extract multi-scale texture features from the source image, and inject them to the decoder of \mathcal{G}_{rf} through SPADE [30]. Notably, \mathbf{I}_{rf} and \mathbf{I}_{raw} are fed into two separate discriminators for adversarial training. Details of the refine network are illustrated in the supplementary.

4. Experiments and Results

4.1. Dataset Preprocessing and Metrics

Dataset Preprocessing. We conduct experiments over three commonly used talking-head generation datasets (*i.e.* VoxCeleb1 [27], VoxCeleb2 [8], TalkingHead-1KH [45]). Each frame is cropped and aligned into 256×256 to center the talking portraits, their rotation angles are predicted with Face-Alignment [5].

Metrics. We measure the quality of synthetic images using structured similarity (**SSIM**), **PSNR** (masked, only compare the region of face, hair, and torso), **LPIPS** [49], and **FID**. Following the previous work [16, 17], we adopt fidelity metrics **CSIM**, **PRMSE**, **AUCON** to evaluate the identity preservation of the source image, the accuracy of head poses and the precision of expression, respectively.

Furthermore, we propose a new metric named average vertices distance (**AVD**) for better identity preservation evaluation. To do this, we first obtain face meshes using [9] and neutralize the impact of pose and expression by setting the corresponding coefficients to 0. Then we calcu-

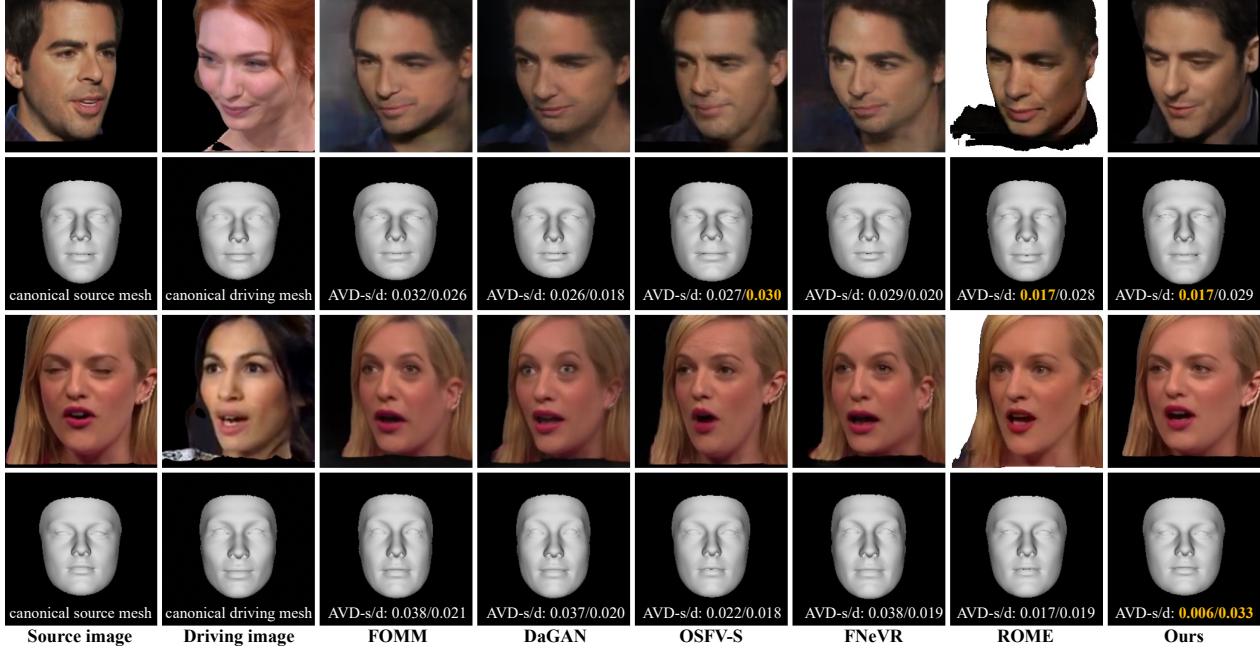


Figure 4. **Comparison of shape preservation with prior works.** The first and third row contains the source, driving, and generated images. The second and fourth row includes the corresponding canonical mesh of their above image. Best results are marked with yellow.

AVD-s/d indicate average vertices distance with source and driving mesh.

late the mesh vertices distance between generated face and source face as AVD-s, and that between synthesized face and driving face as AVD-d. As Fig. 4 illustrates, our approach outperforms other methods by exhibiting a lower AVD-s and a higher AVD-d, which indicates that we can preserve the source face shape effectively and not be swayed by the driving face shape. Since AVD-d does not reflect identity preservation, we only report AVD-s and abbreviate it as AVD in future discussions.

4.2. Talking-Head Synthesis

Baselines. We compare our method against five state-of-the-art methods, including 2D-warping-based methods: FOMM [36], OSFV-S [45] (“-S” indicates the model with SPADE [30] layers, which produces better results), DaGAN [17] and 3D-based methods: ROME [22] (Mesh-based), FNeVR [48] (NeRF-based). All results are obtained by evaluating these method with their official code.

Self-Reenactment. We first compare the synthesized results when the source and driving images are of the same person. The quantitative results concerning the quality and fidelity are listed in Tab. 1 and Tab. 2. It can be observed that our method outperforms other state-of-the-arts on all fidelity metrics. Fig. 5 displays the qualitative comparisons. When head rotation is minimal, all generated results are of similar quality, but as rotation increases, our method produces images with significantly better quality.

Cross-Identity Reenactment. The cross-identity motion transfer was performed on VoxCeleb1 [27], VoxCeleb2 [8]

	SSIM \uparrow	PSNR-M \uparrow	LPIPS \downarrow
FOMM [36]	0.690	19.2	0.112
OSFV-S [45]	0.807	23.2	0.088
DaGAN [17]	0.748	21.8	0.092
ROME [22]	0.833	21.6	0.085
FNeVR [48]	0.801	21.1	0.092
Ours	0.862	21.9	0.084

Table 1. Comparisons with prior works on self-reenactment (**quality metrics**) on the VoxCeleb1 dataset [27]. (\uparrow indicates larger is better, while \downarrow indicates smaller is better.)

	CSIM \uparrow	AUCON \uparrow	PRMSE \downarrow	AVD \downarrow
FOMM [36]	0.837	0.872	2.88	0.021
OSFV-S [45]	0.911	0.934	1.81	0.014
DaGAN [17]	0.875	0.921	1.79	0.016
ROME [22]	0.906	0.918	1.68	0.013
FNeVR [48]	0.880	0.929	2.22	0.016
Ours	0.931	0.956	1.66	0.010

Table 2. Comparisons with prior works on self-reenactment (**fidelity metrics**) on the VoxCeleb1 dataset [27].

and TalkingHead-1KH [45], where the source image and the driving image contains different persons. The quantitative results, as detailed in Tab. 3, show that our method outperforms other methods substantially, conclusively affirming the positive impact of our proposed *MGA* and *LED*.

	VoxCeleb1 [27]					VoxCeleb2 [8]					TalkingHead-1KH [45]				
	CSIM ↑	AUCON ↑	PRMSE ↓	FID ↓	AVD ↓	CSIM ↑	AUCON ↑	PRMSE ↓	FID ↓	AVD ↓	CSIM ↑	AUCON ↑	PRMSE ↓	FID ↓	AVD ↓
FOMM [36]	0.748	0.752	3.66	86	0.044	0.680	0.707	4.16	85	0.047	0.723	0.741	3.71	76	0.039
OSFV-S [45]	0.791	0.893	3.01	74	0.028	0.711	0.833	3.84	72	0.033	0.787	0.884	3.03	67	0.025
DaGAN [17]	0.790	0.880	3.06	87	0.036	0.693	0.815	3.93	86	0.040	0.766	0.872	2.98	73	0.035
ROME [22]	0.833	0.871	2.64	76	0.016	0.710	0.821	3.08	76	0.019	0.781	0.864	2.66	68	0.017
FNeVR [48]	0.812	0.884	3.32	82	0.041	0.699	0.829	3.90	84	0.047	0.775	0.879	3.39	73	0.037
Ours	0.876	0.917	2.62	57	0.012	0.787	0.889	2.91	61	0.014	0.828	0.901	2.60	52	0.011

Table 3. Comparisons with prior works on cross-identity reenactment with different datasets.

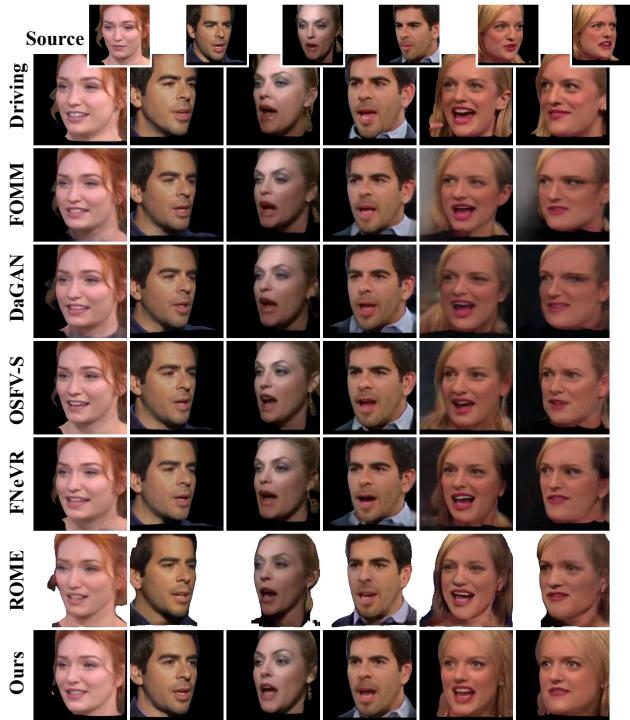


Figure 5. Qualitative comparisons of self-reenactment on the VoxCeleb1 dataset [27].

on image fidelity. Furthermore, Fig. 6 presents qualitative results, where the source and driving images differ significantly in several attributes (head orientation, gender, facial shape, skin tone, *etc.*). As shown in the second and third row of Fig. 6, previous works might generate images with artifacts when the source and driving image have large head rotations, while our method can still produce high-fidelity results. Furthermore, the generated facial shape of other methods is influenced by the driving image, while our method effectively preserves the source face shape, showing improved identity fidelity. Finally, our method exhibits remarkable precision in imitating the expression of the driving image, as demonstrated in the last two rows of Fig. 6.

4.3. Free-View Synthesis

We also benchmark the face redirection capability of our proposed Hide-NeRF with other free-view talking

	CSIM ↑	AUCON ↑	PRMSE ↓	AVD ↓
OSFV-S [45]	0.727	0.808	4.82	0.039
Ours	0.829	0.864	3.78	0.014

Table 4. Comparisons with OSFV [45] on cross-identity reenactment under free-view generation with the VoxCeleb1 dataset [27].

head synthesis method [45]. Specifically, we render the generated results with different view angles. In Tab. 4, we evaluate by rendering results from varying views and averaging the corresponding metrics. Our method overtakes OSFV-S in free-view synthesis to a greater extent than cross-identity reenactment, indicating superior face redirection capability compared to OSFV-S. The difference in CSIM is the most noticeable, where our method surpasses OSFV-S by 0.102, demonstrating that our method can well preserve identity information under different views. We also exhibit some qualitative comparisons with OSFV-S on the yaw angle extrapolation in Fig. 7. Our method produces more realistic results. As seen in Fig. 7, our method accurately imitates the mouth shape of the driving image and preserves realistic details such as teeth. Despite large divergence from the source image in the leftmost column, our method maintains skin texture, while OSFV-S fails to do so. Our method also retains expression consistency under differing view angles, unlike OSFV-S, which has mismatched eye gaze. Please consult the supplementary for more qualitative comparisons.

4.4. Ablation Study

We also benchmark our performance gain upon different modules. Specifically, we conduct four ablations about our proposed *MGA* and *LED*. As for the *MGA*, we replace our proposed multi-scale tri-plane representation with single-scale tri-plane representation (w/o multi-scale). We also test the effectiveness of the camera to world transformation by deprecating it. Concerning the *LED*, we deploy a pose-coupled SECC (w/o SECC), with $\mathbf{R} = \mathbf{R}^{dri}$, $\mathbf{t} = \mathbf{t}^{dri}$.

Effectiveness of Multi-Scale Tri-Plane Representations.

Multi-scale representation brought more delicate features concerning the identity (CSIM increased by 0.062) and expression details (AUCON increased by 0.065), but not sig-



Figure 6. **Qualitative comparisons of cross-identity reenactment on the VoxCeleb1 dataset [27].** Our method captures the driving motion and preserves the identity information better.



Figure 7. **Qualitative comparisons of free-view generation on the VoxCeleb1 dataset [27].** (Best view when zoomed in).

	CSIM ↑	AUCON ↑	PRMSE ↓	AVD ↓
w/o multi-scale	0.814	0.852	2.62	0.017
w/o cam2world	0.760	0.864	3.58	0.024
w/o SECC	0.802	0.879	3.06	0.018
Full	0.876	0.917	2.62	0.012

Table 5. Ablation Study over different modules.

nificant improvement for the head orientation(PRMSE). As discussed, the camera-to-world transformation will relieve the learning difficulty of the tri-plane representation, thus brought much performance gain regarding different metrics.

Effectiveness of Decoupling Pose. As shown in Tab. 5, coupling the pose with other information will be harmful for the identity preservation (CSIM dropped by 0.074, AVD increased by 0.006) and expression preciseness (AUCON dropped by 0.038). These results verify that our proposed module benefits the fidelity of talking-head generation. Please consult the Supplementary Materials for more

qualitative comparisons about the ablation studies.

5. Conclusion

In this paper, we propose High-fidelity and Deformable Neural Radiance Field (HiDe-NeRF) for high-fidelity and free-view talking head synthesis. HiDe-NeRF learns a multi-scale neural radiance field from one source image to preserve identity information, and use a expression-aware deformation field to model local non-rigid expression. Ablation studies clearly show that the proposed modules can benefit the motion transfer between two faces. We demonstrate that our approach can achieve the state-of-the-art synthesis quality on multiple benchmark datasets. Moreover, our model can be easily applied to other modality-driven (audio, text, etc.) talking head synthesis, by replacing the inputs of the expression-aware deformation module. We consider this virtue as fruitful avenues for future work.

Limitations. Our method has certain limitations. First, we can't handle the obvious facial occlusions in the source image. Second, due to the pose bias in the training datasets, we can not obtain satisfactory results with extreme poses. Also, considering the social impact, being a face reenactment method has the risk of misuse for “DeepFakes”.

Acknowledgement. This research is supported in part by Shanghai AI Laboratory, National Natural Science Foundation of China under Grant 62106183 and 62106182, Natural Science Basic Research Program of Shaanxi under Grant 2021JQ-204, and Alibaba Group.

References

- [1] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 20364–20373, 2022. [2](#), [3](#), [4](#), [5](#)
- [2] Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen. Bringing portraits to life. In *ACM Trans. Graph.*, 2017. [2](#)
- [3] Alexander W. Bergman, Petr Kellnhofer, Wang Yifan, Eric R. Chan, David B. Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. In *Adv. Neural Inform. Process. Syst.*, 2022. [2](#), [4](#)
- [4] 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*, pages 187–194, 1999. [2](#), [3](#)
- [5] 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 *Int. Conf. Comput. Vis.*, 2017. [5](#)
- [6] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. [2](#)
- [7] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. [2](#), [4](#)
- [8] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. [5](#), [6](#), [7](#), [1](#)
- [9] 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 *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2019. [5](#), [1](#)
- [10] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-GAN for pose-guided person image synthesis. In *Adv. Neural Inform. Process. Syst.*, 2018. [2](#)
- [11] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Alekssei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *ACM Int. Conf. Multimedia*, 2022. [2](#)
- [12] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided GANs for single-photo facial animation. In *ACM Trans. Graph.*, 2018. [2](#)
- [13] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18653–18664, 2022. [2](#)
- [14] Kuangxiao Gu, Yuqian Zhou, and Thomas S Huang. FLNet: Landmark driven fetching and learning network for faithful talking facial animation synthesis. In *AAAI*, 2020. [2](#)
- [15] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Int. Conf. Comput. Vis.*, 2021. [3](#)
- [16] Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. MarioNETte: Few-shot face reenactment preserving identity of unseen targets. In *AAAI*, 2020. [2](#), [5](#)
- [17] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3397–3406, 2022. [2](#), [5](#), [6](#), [7](#)
- [18] Kaiwen Jiang, Shu-Yu Chen, Feng-Lin Liu, Hongbo Fu, and Lin Gao. NeRFFaceEditing: Disentangled face editing in neural radiance fields. In *SIGGRAPH Asia*, 2022. [4](#)
- [19] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Adv. Neural Inform. Process. Syst.*, 2020. [1](#)
- [20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4401–4410, 2019. [3](#)
- [21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8110–8119, 2020. [2](#)
- [22] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *Eur. Conf. Comput. Vis.*, 2022. [3](#), [6](#), [7](#)
- [23] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid Warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Int. Conf. Comput. Vis.*, 2019. [2](#)
- [24] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. In *Eur. Conf. Comput. Vis.*, 2022. [3](#)
- [25] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Implicit Warping for Animation with Image Sets. In *Adv. Neural Inform. Process. Syst.*, 2022. [2](#)
- [26] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Eur. Conf. Comput. Vis.*, 2020. [2](#)
- [27] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017. [5](#), [6](#), [7](#), [8](#), [1](#), [3](#), [4](#)
- [28] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Int. Conf. Comput. Vis.*, pages 5865–5874, 2021. [2](#), [3](#), [4](#), [5](#)
- [29] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. HyperNeRF: A higher-dimensional representation for topologically varying neural radiance fields. In *ACM Trans. Graph.* ACM, dec 2021. [2](#), [3](#), [4](#), [5](#)

- [30] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2337–2346, 2019. 5, 6
- [31] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 2, 3, 4
- [32] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Eur. Conf. Comput. Vis.*, pages 523–540. Springer, 2020. 2, 4
- [33] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2
- [34] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H. Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Int. Conf. Comput. Vis.*, 2021. 3
- [35] Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. In *Eur. Conf. Comput. Vis.*, 2022. 3
- [36] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Adv. Neural Inform. Process. Syst.*, 32, 2019. 2, 6, 7
- [37] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2019. 2
- [38] Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2
- [39] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing Obama: learning lip sync from audio. In *ACM Trans. Graph.*, 2017. 3
- [40] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. StyleRig: Rigging stylegan for 3d control over portrait images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 3
- [41] Ayush Tewari, Mohamed Elgharib, Mallikarjun B R, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. In *ACM Trans. Graph.*, 2020. 3
- [42] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. In *ACM Trans. Graph.*, 2019. 3
- [43] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. In *ACM Trans. Graph.*, 2015. 3
- [44] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2Face: Real-time face capture and reenactment of rgb videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 3
- [45] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2, 3, 5, 6, 7
- [46] Shunyu Yao, RuiZhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. DFA-NeRF: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv preprint arXiv:2201.00791*, 2022. 3
- [47] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *Eur. Conf. Comput. Vis.*, August 2020. 2
- [48] Bohan Zeng, Boyu Liu, Hong Li, Xuhui Liu, Jianzhuang Liu, Dapeng Chen, Wei Peng, and Baochang Zhang. FNeVR: Neural volume rendering for face animation. In *Adv. Neural Inform. Process. Syst.*, 2022. 2, 6, 7
- [49] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 586–595, 2018. 5, 2
- [50] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3657–3666, 2022. 2
- [51] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017. 4, 5
- [52] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12786–12796, 2022. 4
- [53] zllrunning. face-parsing.pytorch. <https://github.com/zllrunning/face-parsing.PyTorch>, 2019. 1

Supplemental Material

One-Shot High-Fidelity Talking-Head Synthesis with Deformable Neural Radiance Field

In this supplement, we first provide additional training details and network architecture implementation for completeness and reproducibility. We discuss experiment details, such as datasets and metrics, and provide more qualitative results. Lastly, we also show some failure cases that may be targets of future work.

6. Additional Details about Networks and Training Details

6.1. Training Details

Perturbation Regularization We observe that directly feeding the point positions into the deformation field leads to over-fitting and alias. To this end, we devise a perturbation regularization, which adds very small random perturbation into the point positions \mathbf{P} fed into the *LED*. However, the predicted deltas $\Delta\mathbf{P}$ are added to the original points instead of the perturbed points to form the \mathbf{P}' . The intuition of our design lies in the fact that nearby points should have similar deformation, and this perturbation regularization enables our deformation module to learn the “*local consistency*”.

Data Augmentation. NeRF is known to benefit from different view inputs, but there are subtle pose changes in most short video clips. To improve the diversity of head poses, we set the flipping probability $p_{flip} = 0.5$ to randomly flip the source image horizontally. Considering flipping leads to changes in the background, we only preserve the head and torso parts in the video frames via an off-the-shelf segmentation predictor [53].

6.2. Loss Functions

In the training stage, we can take different frames of the same video as the source and driving images, as they share the same identity. Our proposed HiDe-NeRF is combined of NeRF and a refine module, thus it will generate two outputs \mathbf{I}_{raw} and \mathbf{I}_{rf} , we take the driving image \mathbf{I}_{dri} and its downsampled result $\mathbf{I}_{\text{dri}}^{\text{down}}$ as their corresponding ground-truth. It is optimized with the following loss functions:

Mean Square Error Loss \mathcal{L}_M . We minimize the mean square error of the refined image \mathbf{I}_{rf} and rendered image \mathbf{I}_{raw} w.r.t. their corresponding ground-truth.

Perceptual Loss \mathcal{L}_P . We minimize the perceptual loss of the refined image \mathbf{I}_{rf} and rendered image \mathbf{I}_{raw} w.r.t. their corresponding ground-truth to get a more realistic result.

Adversarial Loss \mathcal{L}_G . We deploy conditional discriminator as in StyleGAN2-ADA [19]. Specifically, \mathbf{I}_{rf} and \mathbf{I}_{raw} are fed into two separate discriminators, and the camera parameters are used as the condition.

Deformation Regularization \mathcal{R}_D . We add a deformation regularization as a loss to enforce the deformation module find the “shortest” path for each single point. The deformation regularization is the sum of $l1$ norm for the predicted point-wise delta. Concretely, it is calculated by $\mathcal{R}_D = \|\mathcal{F}_{\Phi}^{\text{deform}}(\mathbf{P}, \text{SECC}_{\text{dri}}, \text{SECC}_{\text{can}})\|_1$.

MSE loss ensures the rendered and refined images be similar to their ground-truth. While the perceptual and adversarial loss ensures the images are more realistic. And the last regularization loss makes the training stage more stable. The overall loss can be summarized as below:

$$\begin{aligned} \mathcal{L} = & \lambda_M (\mathcal{L}_M(\mathbf{I}_{\text{rf}}, \mathbf{I}_{\text{dri}}) + \mathcal{L}_M(\{\mathbf{I}_{\text{raw}}, \mathbf{I}_{\text{dri}}^{\text{down}}\}_{n=1}^K)) \\ & + \lambda_P (\mathcal{L}_P(\mathbf{I}_{\text{rf}}, \mathbf{I}_{\text{dri}}) + \mathcal{L}_P(\{\mathbf{I}_{\text{rf}}, \mathbf{I}_{\text{dri}}^{\text{down}}\}_{n=1}^K)) \\ & + \lambda_G (\mathcal{L}_G(\mathbf{I}_{\text{rf}}, \mathbf{I}_{\text{dri}}) + \mathcal{L}_G(\{\mathbf{I}_{\text{rf}}, \mathbf{I}_{\text{dri}}^{\text{down}}\}_{n=1}^K)) \\ & + \lambda_R \mathcal{R}_D. \end{aligned} \quad (\text{S1})$$

6.3. Network Architecture details

The implementation details of some of the sub-networks in our work are illustrated in Fig. S1 and we will introduce them accordingly. We also show the architecture of SPADE and different blocks in Fig. S2 and Fig. S3.

Feature Extractor. The feature extractor neural network extracts appearance features from the source image. It consists of a number of downsampling blocks and a number of upsampling blocks to compute the final tri-plane volume features.

Refine Network. In the training process, we feed the rendered image and features extracted by ID Encoder into the refine network, and get the refined image. The detailed structure of the refine network is shown in Fig. S1(b).

ID Encoder. We utilize another convolutional neural network to extract identity information from the source image, and inject the extracted information into the refine network with SPADEBlock. The detailed connections between the refine network and the ID encoder is shown in Fig. S1(b-c).

7. Additional Details about Experiments

7.1. Dataset Details

We use the following datasets in our evaluations. During testing, we first align [9] faces and segment the facial parts out as the input.

VoxCeleb1 [27]. The VoxCeleb dataset contains about 20,000 videos. For pre-processing, we extract an initial bounding box in the first video frame. The validation dataset contains about 500 videos.

VoxCeleb2 [8]. This dataset contains about 1M talking head videos of different celebrities. We follow the training and test split proposed in the original paper and report

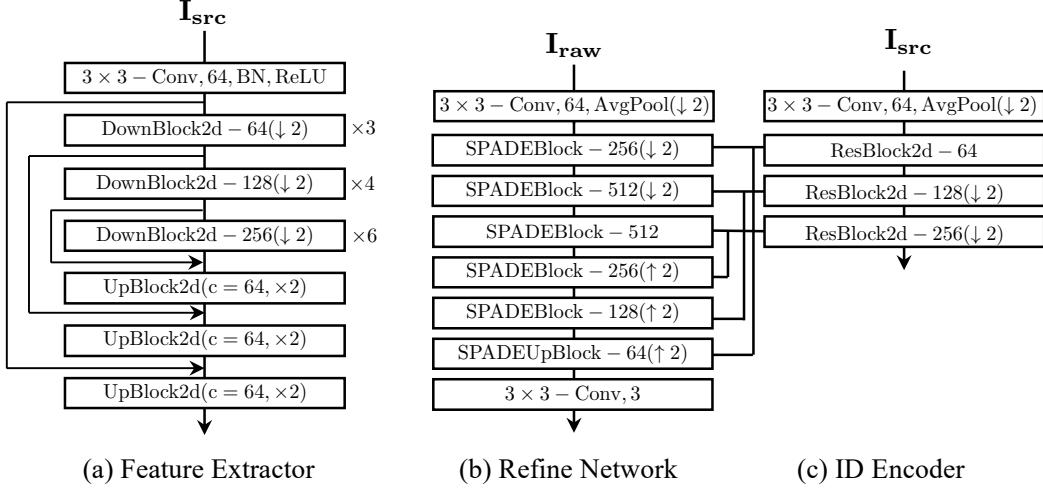


Figure S1. Illustration of sub-networks.

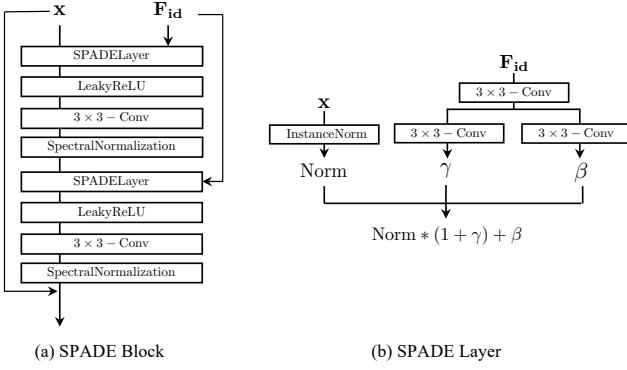


Figure S2. Illustration of SPADE block and layer.

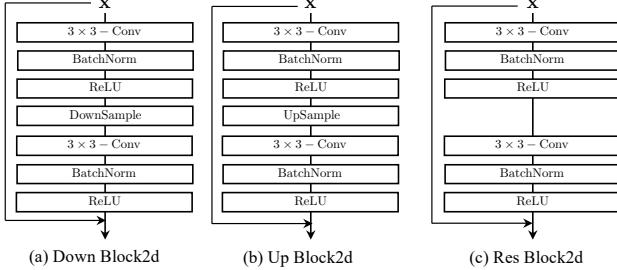


Figure S3. Illustration of different building blocks.

our results on the validation set, which contains about 36k videos.

TalkingHead-1KH [45]. This dataset is composed of about 1000 hours of videos from various sources. This dataset is in general with higher resolutions and better image quality than those in the VoxCeleb2. For fair comparison, we report results on the resized faces at the 256×256 resolution.

7.2. Metrics

PSNR is numerically related to the mean squared error (MSE) between the ground truth and the reconstructed image, it is used to measure the image reconstruction quality.

SSIM measures the structural similarity between patches of

the input images. As a result, it is more robust to changes in the global illumination than PSNR.

LPIPS [49] calculates the cosine distances between the network features of the two images layer by layer and averages them to estimate the perceived distance of the generated image from the ground truth image.

CSIM [49] To evaluate the effectiveness of identity preservation, we compute the cosine similarity using embedded vectors created by the pre-trained face recognition model.

AUCON is used to calculate the ratio of the same facial action unit values between the generated images and the driving images.

7.3. Ablation Study

We use self-reenactment for qualitative comparison as the driving image serves as the ground truth, which will better exhibit identity preservation. As shown in the Fig. S4, our proposed method can preserve the identity information from the source image quite well. However, when removing the multi-scale module, the identity information is largely change, the skin tone and texture cannot be maintained quite well, but the generated result can still mimic the driving motion. If we remove the camera-to-world transformation, the pose of the generated image cannot be changed precisely. We found that the rendered image is quite blurry and cannot guarantee large pose changes. As shown in the last row of Fig. S4, entangling the pose with expression will lead to inaccurate expression, there is a problem with the eye and mouth movements in the generated image.

7.4. Additional Qualitative Results

In this section, we will show more qualitative results of cross-identity reenactment on different datasets. We show some representative results in this supplementary, Fig. S6 and Fig. S7 contains cross-identity reenactment on the VoxCeleb1 dataset. Fig. S8 and Fig. S9 shows the cross-



Figure S4. Ablation Study.

identity reenactment on the VoxCeleb2 and TalkingHead-1KH dataset, respectively.

There is a diversity of disparities in the presented images (*e.g.*, gender, face shape, skin color, beard, *etc.*). Other methods have difficulty in maintaining the identity information of the source image under such disparities. The facial shape of images generated with warping-based methods are easily affected by the driving image, especially when the face shape difference is large. However, our method is able to maintain better identity information under all these conditions.

7.5. Free-view Rendering

Fig. S10 and Fig. S11 provides a visual comparison of our method against OSFV-S [45] for generating views from steep camera poses, the first two columns/rows contains the generated results of VoxCeleb1 [27] and the other two columns/rows contains the results of TalkingHead-1KH [45]. We observe that front-facing photos make up most of the commonly used talking-head datasets. Severe yaw and pitch angles are rarely shown in photographs, and neither is extreme yaw angles. Yet, reasonable extrapolation to different poses is a critical trait for real-world talking head synthesis.

As shown in Fig. S10, our method is able to cope with different angles, showing satisfactory results with different view directions. However, OSFV struggles to maintain the expression-consistency when shifting the view angles, which is undesired in generating talking-heads in the real-world. At the same time, when shifting the yaw angle, our

method is able to rotate more accurately without changing other angles. Meanwhile, as shown in Fig. S11, OSFV-S has difficulty in rotating the pitch angle. Even though we change the viewing angle considerably, there was no significant change in the visual effect of the generated images. This phenomenon aligns with our observation that the variation in pitch angle in the dataset is small.

7.6. Failure Cases

As shown in Fig. S5, our method generates results with degraded quality when there are occlusions in the images, such as microphones, and sunglasses. Also, our method cannot faithfully preserve the identity information under extreme pose changes.



Figure S5. Example failure cases.

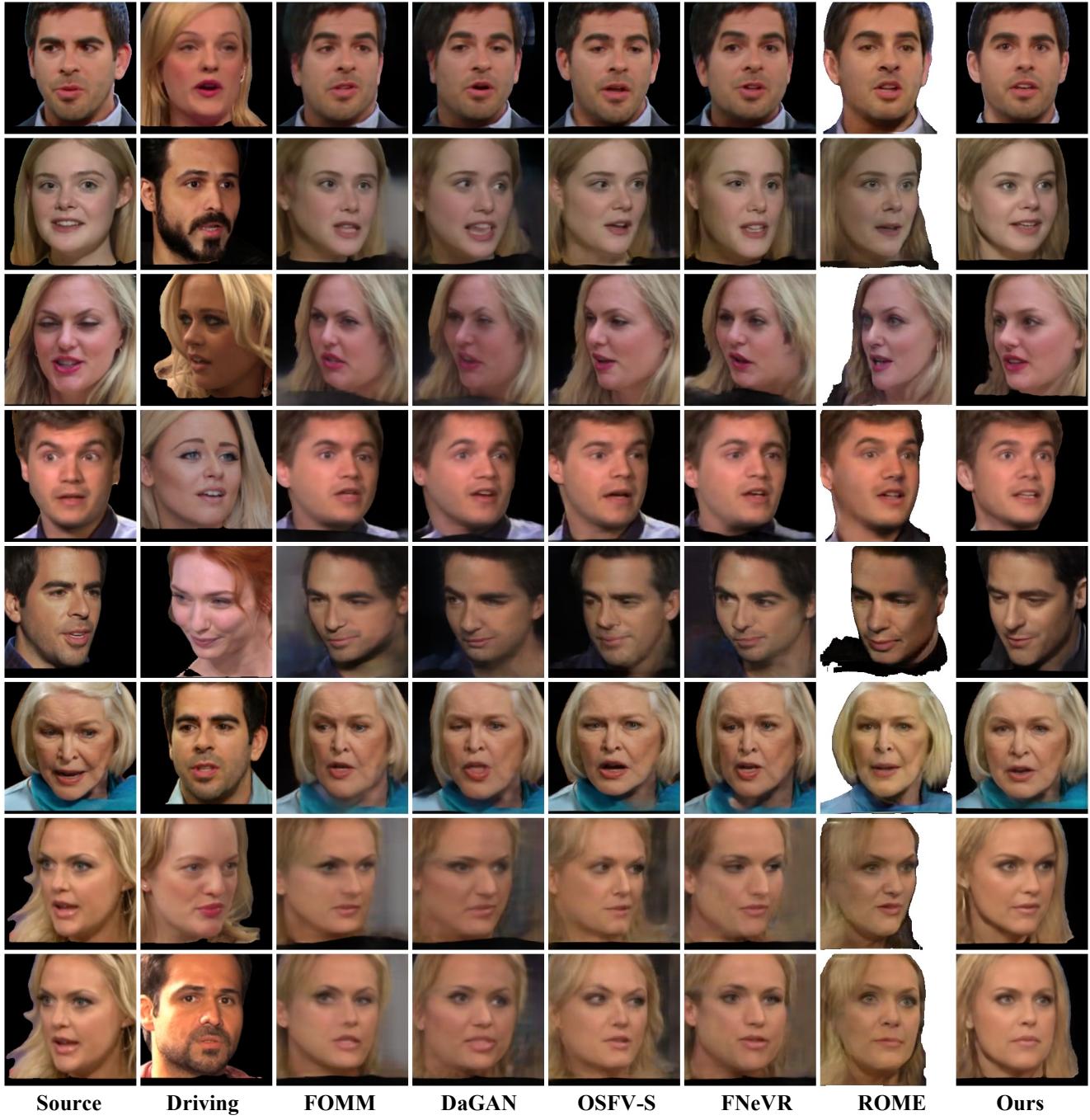


Figure S6. Qualitative comparisons of cross-identity reenactment on the VoxCeleb1 dataset [27].



Figure S7. Qualitative comparisons of cross-identity reenactment on the VoxCeleb1 dataset [27].



Figure S8. Qualitative comparisons of cross-identity reenactment on the VoxCeleb2 dataset [8].

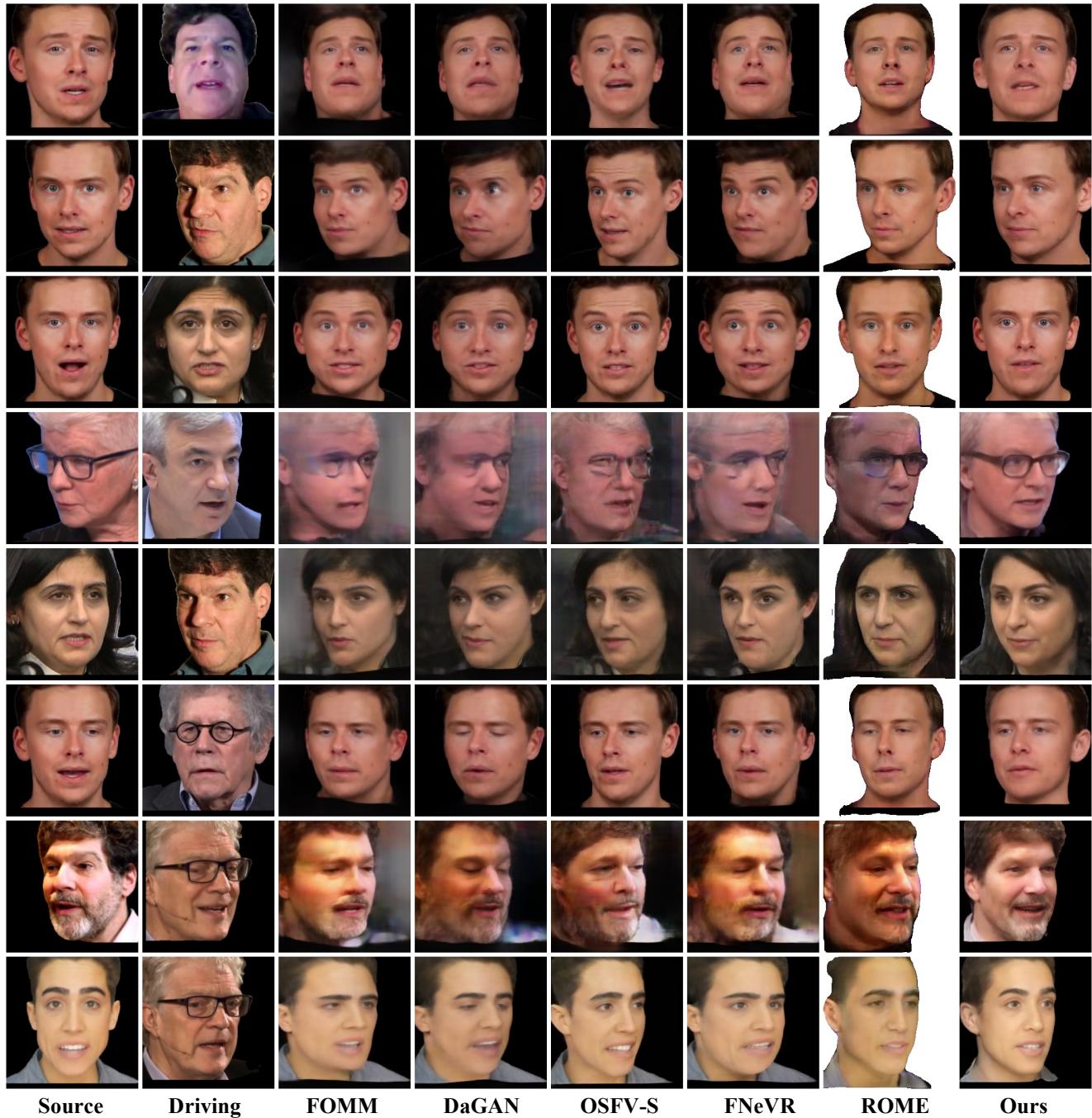


Figure S9. Qualitative comparisons of cross-identity reenactment on the TalkingHead-1KH dataset [45].

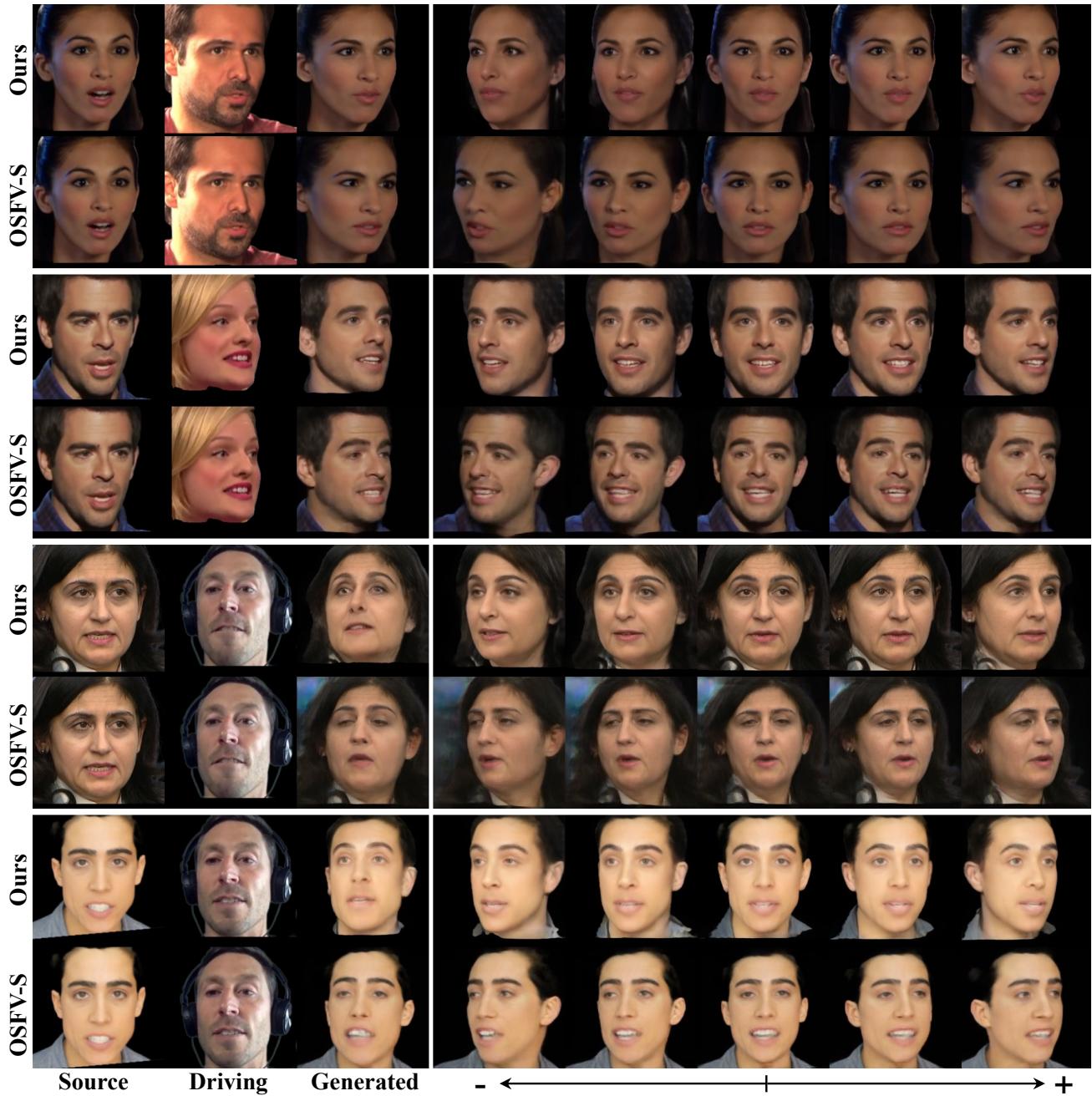


Figure S10. **Extrapolation to steep yaw angles.** The first two rows contains the generated images from VoxCeleb1 dataset, while the last two rows are from TalkingHead-1KH dataset.

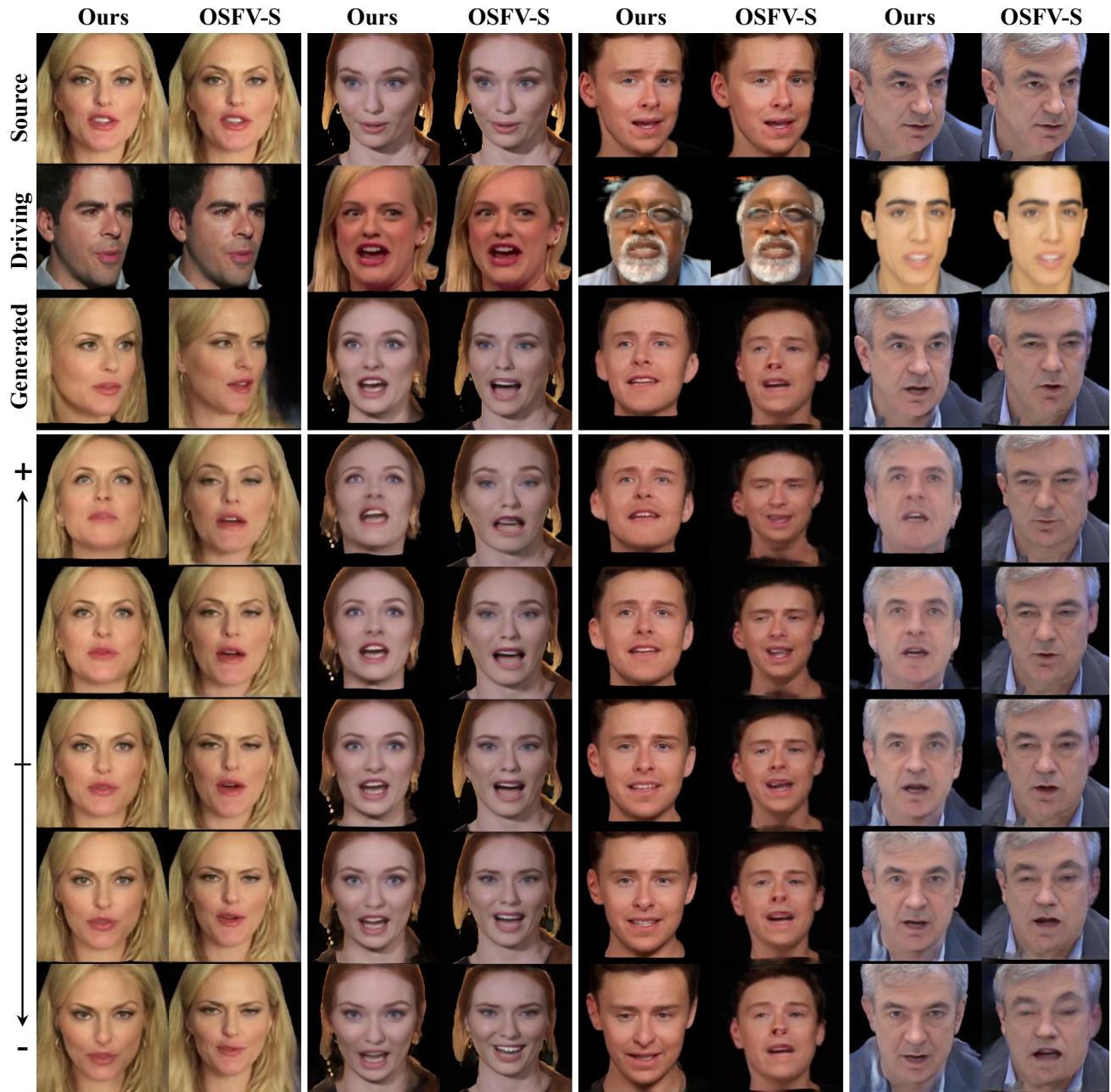


Figure S11. **Extrapolation to steep pitch angles.** The first two columns contains the generated images from VoxCeleb1 dataset, while the last two columns are from TalkingHead-1KH dataset.