

Learning Dense Correspondence for NeRF-Based Face Reenactment

Songlin Yang^{1,2}, Wei Wang^{2*}, Yushi Lan³, Xiangyu Fan⁴, Bo Peng², Lei Yang⁴, Jing Dong²

¹School of Artificial Intelligence, University of Chinese Academy of Sciences, China

²CRIPAC & MAIS, Institute of Automation, Chinese Academy of Sciences, China

³S-Lab, Nanyang Technological University, Singapore

⁴SenseTime, China

yangsonglin2021@ia.ac.cn, {wwang, bo.peng, jdong}@nlpr.ia.ac.cn,

yushi001@e.ntu.edu.sg, {fanxiangyu, yanglei}@sensetime.com

Abstract

Face reenactment is challenging due to the need to establish dense correspondence between various face representations for motion transfer. Recent studies have utilized Neural Radiance Field (NeRF) as fundamental representation, which further enhanced the performance of multi-view face reenactment in photo-realism and 3D consistency. However, establishing dense correspondence between different face NeRFs is non-trivial, because implicit representations lack ground-truth correspondence annotations like mesh-based 3D parametric models (e.g., 3DMM) with index-aligned vertexes. Although aligning 3DMM space with NeRF-based face representations can realize motion control, it is sub-optimal for their limited face-only modeling and low identity fidelity. Therefore, we are inspired to ask: *Can we learn the dense correspondence between different NeRF-based face representations without a 3D parametric model prior?* To address this challenge, we propose a novel framework, which adopts tri-planes as fundamental NeRF representation and decomposes face tri-planes into three components: canonical tri-planes, identity deformations, and motion. In terms of motion control, our key contribution is proposing a Plane Dictionary (**PlaneDict**) module, which efficiently maps the motion conditions to a linear weighted addition of learnable orthogonal plane bases. To the best of our knowledge, our framework is the first method that achieves one-shot multi-view face reenactment without a 3D parametric model prior. Extensive experiments demonstrate that we produce better results in fine-grained motion control and identity preservation than previous methods.

Introduction

One-shot face reenactment (Hong et al. 2022) aims to utilize motion conditions from the driving image, such as facial expressions and head poses, to animate the face in the source image. The main challenge is establishing dense correspondence between different face representations to transfer motion conditions. Recent studies (Li et al. 2023b; Ma et al. 2023b) have utilized Neural Radiance Field (NeRF) (Mildenhall et al. 2021) as fundamental representation, which further enhanced the performance of multi-view face reenactment in photo-realism and 3D consistency.

However, establishing dense correspondence between different face NeRFs is non-trivial. Unlike mesh-based representations which have index-aligned vertexes as ground-truth correspondence annotations, the NeRF-based representations lack an explicit surface descriptor that constructs correspondence of spatial points (Lan, Loy, and Dai 2022). Although introducing 3D parametric models (e.g., 3DMM (Blaiz and Vetter 1999), FLAME (Li et al. 2017), and DECA (Feng et al. 2021)) as motion conditions make it feasible to achieve explicit motion control for cross-identity face reenactment (Zeng et al. 2022; Ma et al. 2023b; Li et al. 2023b), aligning mesh-based parametric space with latent space of NeRF-based generative models brings a significant optimization burden. Additionally, the 3D parametric models themselves have some limitations, such as their focus being predominantly on the facial region, requiring additional processing for hair and eyes. These limitations inspire us to ask: Can we learn the dense correspondence between different NeRF-based face representations without a 3D parametric model prior?

To address the challenge of learning dense correspondences between NeRF-based face representations, the first issue is the selection of NeRF representations. The vanilla NeRF (Mildenhall et al. 2021) employs an MLP network to capture the spatial information of the target object, which tends to suffer from overfitting and can lead to a loss of 3D consistency when animating the representation network. In order to strike a balance between 3D consistency and animation capabilities, we have opted to utilize the tri-plane representation proposed by EG3D (Chan et al. 2022) as our fundamental NeRF representation, which adopts three spatially-orthogonal plane feature maps to represent an object. This choice allows us to maintain 3D consistency within the tri-plane representation itself, while also leveraging the strong modeling capacity of the StyleGAN-based (Karras et al. 2020) generator to handle feature deformations.

In this work, we propose a novel framework that can realize one-shot multi-view face reenactment, as shown in Fig. 1. Our framework utilizes tri-planes as fundamental NeRF representation and decomposes face tri-planes into three components: canonical tri-planes, identity deformations, and motion. The plane feature deformations regarding motion conditions differ from those caused by identity conditions since the rules governing motion are shared be-

*Corresponding author.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

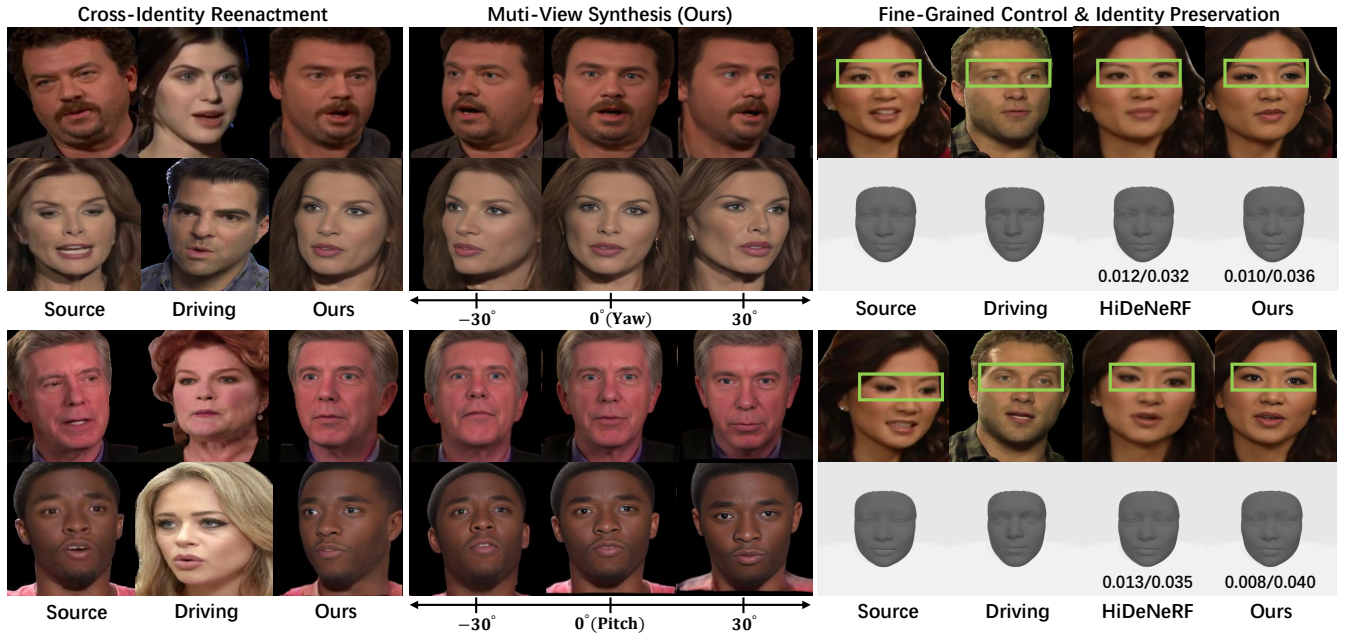


Figure 1: Results of one-shot multi-view face reenactment. We present both cross-identity reenactment and multi-view synthesis at various yaw and pitch angles. Through comparisons with state-of-the-art HiDeNeRF, we illustrate that our PlaneDict module excels in fine-grained motion control, particularly for non-facial elements such as eyes, and offers better identity preservation (vertex distance from source identity↓/ vertex distance from driving identity↑) than utilizing 3DMM as correspondence.

tween various identities. Thus, we design a Plane Dictionary module, referred to as **PlaneDict**, to efficiently maps motion conditions to a linear weighted addition of learnable orthogonal plane bases. Extensive experiments demonstrate that our method achieves better results in fine-grained motion control and identity preservation than previous work.

To summarize, the contributions of our approach are:

- We propose a novel decomposition method of face tri-plane representation, making it suitable for learning the dense correspondence between different face tri-planes and realizing motion transfer.
- We propose a Plane Dictionary (**PlaneDict**) module for tri-plane representation, which efficiently maps motion conditions to a linear weighted addition of learnable orthogonal plane bases.
- To the best of our knowledge, we propose the first method to achieve one-shot multi-view face reenactment without a 3DMM prior, which achieves better results in fine-grained motion control and identity preservation than previous work.

Related Work

Face Implicit Representation

Compared with 2D (Liu et al. 2015) and 3D parametric representation (Li et al. 2017), 3D implicit representation has advantages of photo-realism and 3D-consistency. Previous work based on 3D scene representation has tried to use Neural Radiance Field (Gu et al. 2022; Yang et al. 2023a), Signed Distance Field (Or-El et al. 2022; Ma et al. 2023a),

and Tri-Planes (Chan et al. 2022) to model face as static objects. Considering the dynamic synthesis requirements of faces, two strategies have been proposed: First, constructing the deformable neural radiance fields, such as NeRFies (Park et al. 2021a) and HyperNeRF (Park et al. 2021b), which maps each observed point into a canonical space through a continuous deformation field, but it tends to handle small movements or person-specific rendering. Second, adopting NeRF with 3DMM (Li et al. 2017) prior for explicit motion control, such as RigNeRF (Athar et al. 2022), NeR-Face (Gafni et al. 2021), MoFaNeRF (Zhuang et al. 2022), OmniAvatar (Xu et al. 2023), and some 3D GAN Inversion methods (Lin et al. 2022; Lan et al. 2023; Yang et al. 2023b). However, the dense correspondence provided by 3DMM has limitations in non-facial regions (e.g., eyes and hair) and brought an optimization burden to align the 3DMM representation and NeRF-based latent space. Therefore, a better dense correspondence of different 3D implicit representations is needed.

One-Shot Face Reenactment

Previous face reenactment methods can be divided into warping-based, mesh-based, and NeRF-based. Warping-based methods (Dong et al. 2018; Geng et al. 2018; Liu et al. 2019; Ha et al. 2020; Drobyshev et al. 2022; Zhao and Zhang 2022; Wang et al. 2022) warp the source features by estimated motion field to transport driving expressions and poses into the source face for 2D generation. Among them, FOMM (Siarohin et al. 2019) builds a 2D motion field from the sparse keypoints detected by an unsupervised trained de-

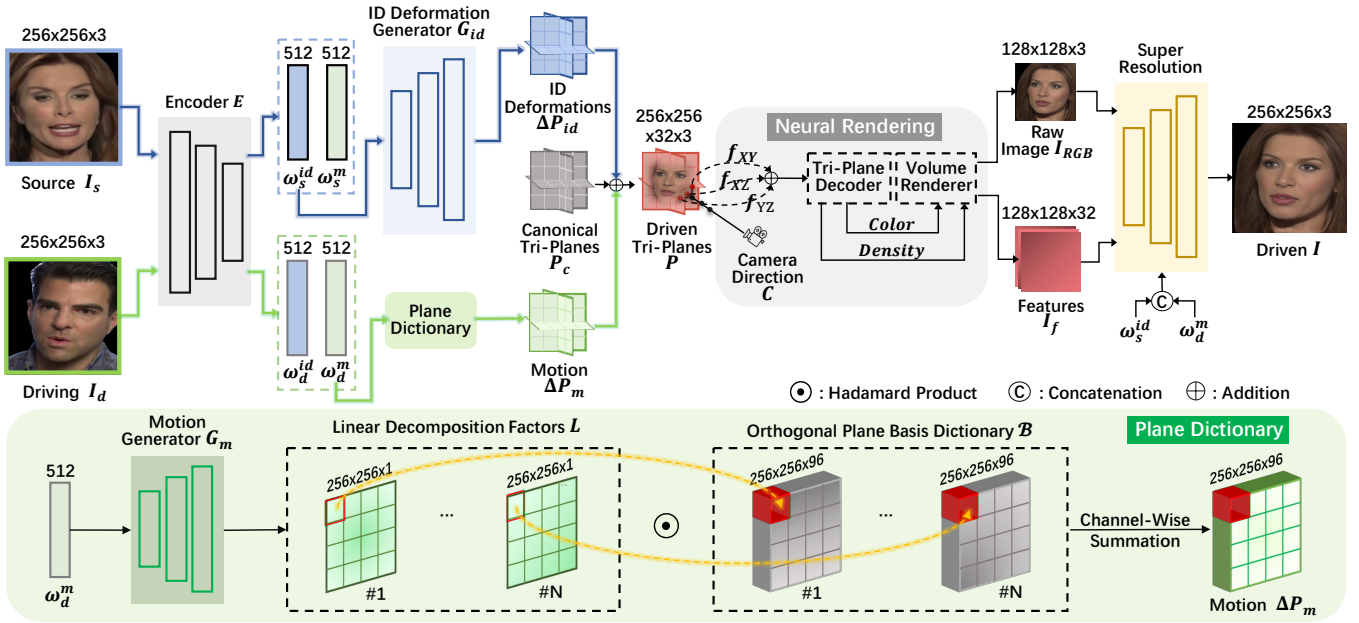


Figure 2: Method overview. We decompose face tri-planes into three components: canonical tri-planes, identity deformations, and motion. In terms of motion control, which is the topology transformation rules shared by different identities, we propose a Plane Dictionary (PlaneDict) module to transfer motion conditions between different face tri-planes for face reenactment.

tector, while DaGAN (Hong et al. 2022) incorporates the depth estimation to supplement the missing 3D geometry information in 2D motion field. Mesh-based methods employ 3DMM uses a single image to create realistic photos in a rigged mesh format such as ROME (Khakhulin et al. 2022). In terms of NeRF-based methods (Guo et al. 2021; Liu et al. 2022; Shen et al. 2022; Li et al. 2023a), using one image to build a 3D implicit representation is an ill-posed problem, because the lack of multi-view information makes the failure of learning the dense spatial information from one image. FNeRF (Zhang et al. 2022) relaxes the constraint to the required number of images, while FNeVR (Zeng et al. 2022) takes the merits of 2D warping and 3D neural rendering. As for the motion control of tri-plane representation, OTAvatar (Ma et al. 2023b) designs a motion encoding strategy for pre-trained EG3D (Chan et al. 2022) with the 3DMM prior, while HiDeNeRF (Li et al. 2023b) proposes a multi-scale tri-plane feature extractor, as well as 3DMM-based implicit motion-conditioned deformation field, to train a generative model from scratch. However, these 3DMM-based methods still suffer from the limitations brought by 3DMM itself. Therefore, we aim to tackle the more tricky challenge that is learning the dense correspondence between different tri-planes without 3DMM prior and achieving matchable or even better results than previous methods, which can have the potential of animating arbitrary objects which lack a sophisticated 3D parametric modeling like human faces.

Method

We propose a novel framework to achieve one-shot multi-view face reenactment, which can learn the dense correspondence between different face tri-planes and realize motion

transfer. In terms of motion control, our key contribution is to construct a Plane Dictionary (**PlaneDict**) module to efficiently map motion conditions to feature deformations of tri-planes, which realizes fine-grained motion transfer.

Overview

Tri-Plane Representation. Previous studies have utilized NeRF (Mildenhall et al. 2021) as fundamental implicit representation. Notably, 3D generative models like StyleNeRF (Gu et al. 2022) and EG3D (Chan et al. 2022) combine NeRF-based representation with StyleGAN-based (Karras et al. 2020) generator. These models extend identity-specific overfitting of vanilla NeRFs to a GAN space which can render 3D-consistent multi-identity face images with diverse expressions and poses. Among these NeRF variants, the tri-planes proposed by EG3D achieve a superior balance between information density and photo-realism, while also enabling the construction of a diverse latent space for manipulation. In contrast to the vanilla NeRF, which employs an MLP network to record spatial points in the space, EG3D adopts a latent vector to represent it and maps it to three spatially-orthogonal plane feature maps (i.e., tri-planes) through a StyleGAN-based generator. The tri-planes effectively provide sufficient information for rendering a spatial point and can be queried efficiently. Consequently, we adopt tri-planes as our fundamental representation.

Decomposition Strategy. The computer graphics researchers (Blanz and Vetter 1999; Li et al. 2017) typically decompose the canonical space (also known as the template mesh) and vertex deformations caused by identity and motion conditions in order to enhance the stability and interpretability of the learning process used in the mesh-

based pipeline for modeling human faces. Unfortunately, the EG3D pipeline did not incorporate this decomposition structure, resulting in a lack of explicit control over the identity and motion of human faces. This limitation poses challenges for downstream applications such as facial attribute editing and face reenactment. To address this issue, previous methods (Lin et al. 2022; Ma et al. 2023b) have employed GAN inversion to embed real face images into the latent space of EG3D, which tend to utilize the encoder of a face parametric model to obtain explicit control over identities, expressions, and poses. However, this roadmap only distills the correspondence from the parametric models and inherits their limitations. Therefore, we decompose the canonical space, identity deformations, and motion to learn dense correspondence of tri-planes for more flexible motion transfer.

Specifically, as shown in Fig. 2, we adopt an encoder E to extract the style codes of the input image I_{input} , which can be embedded as identity code ω^{id} and motion code ω^m :

$$(\omega^{id}, \omega^m) = E(I_{input}). \quad (1)$$

The dense correspondence and motion transfer are achieved by identity deformations ΔP_{id} and motion ΔP_m . We feed the identity code ω^{id} into a StyleGAN-based generator G_{id} to obtain the identity deformations ΔP_{id} . However, the motion ΔP_m aims to achieve motion transfer between different face tri-planes, which means a shared deformation method should be proposed to handle this challenge. Therefore, we propose our **PlaneDict** module to obtain the motion ΔP_m , which will be presented in the next section.

The tri-planes P which represents the driven face image to be rendered consist of the learnable canonical tri-planes P_c , the identity deformations ΔP_{id} from the source face image I_s , and the motion ΔP_m from the driving face image I_d , which can be formulated as follows:

$$P(\omega_s^{id}, \omega_d^m) = P_c + \Delta P_{id} + \Delta P_m, \quad (2)$$

$$\Delta P_{id} = G_{id}(\omega_s^{id}), \quad (3)$$

$$\Delta P_m = \text{PlaneDict}(G_m(\omega_d^m); \mathcal{B}), \quad (4)$$

where G_m and \mathcal{B} are the motion deformation generator and the learnable plane bases of the PlaneDict module.

Finally, when we query a spatial point according to its location (x, y, z) and the camera direction C , we sample features $(f_{XY}, f_{XZ}, \text{ and } f_{YZ})$ from the driven tri-planes P , aggregate by summation, and process the aggregated features with a lightweight tri-plane decoder. This decoder outputs a scalar density and a 32-channel feature, and both of them are then processed by a volume renderer to project the 3D feature volume into a 2D feature image. For training efficiency, we render 32-channel feature maps I_f at a resolution of 128^2 , with 96 total depth samples per ray. And we adopt the Super Resolution module to increase the final image size to 256^2 , which utilizes two blocks of StyleGAN-modulated convolution layers that upsample and refine the 32-channel feature maps I_f into the final RGB image I .

Plane Dictionary (PlaneDict)

Preliminary. In the 3D Morphable Model (3DMM) (Blaiz and Vetter 1999), every face is represented by a shared topology which consists of vertexes with the aligned index. The hundreds of these 3D faces are high dimensional data and then reduced through Principal Component Analysis (PCA) to several orthogonal vector bases. These bases are further divided into identity-related and expression-related bases. When we fit the mesh model to a target face, we only need to linearly add these orthogonal bases to obtain the identity and expression deformations of every vertex relative to the template mesh, and finally get the 3D representation of the target face. It is worth noting that the expression representation of different faces can be obtained through linear addition of expression bases, which means that through the above modeling method, we can achieve dense correspondence between different face representations.

Motivation. The modeling approach of graphics inspires us to establish dense correspondence and realize motion transfer between different face tri-planes. However, if we have achieved such dense correspondence, how can we transform these motion conditions into deformations of each implicit spatial point? The previous methods (Lin et al. 2022; Li et al. 2023b; Ma et al. 2023b) either skipped this issue and directly learned the mapping relationship between 3DMM and their latent space of generative models, and then used dense correspondence of 3DMM to control their own generative model; or conducted the learning of dense correspondence and motion transfer in the latent vector space corresponding to each implicit representation. The former is limited by 3DMM and cannot handle non-facial areas such as hair and eyes; The latter, which ignores the inherent characteristics of implicit representation, cannot perform more fine-grained expression control. Therefore, we propose a Plane Dictionary (**PlaneDict**) module, which can obtain the plane feature deformations driven by motion conditions by linearly adding a set of orthogonal plane bases.

Specifically, as shown in Fig. 2 and Eqn. (4), the driving motion code ω_d^m is first fed into the motion generator G_m to obtain the linear decomposition factors L , which consists of N feature maps. These decomposition factors are then multiplied by the orthogonal plane bases in the plane dictionary \mathcal{B} through the Hadamard product. The plane bases in \mathcal{B} are channel-wise orthogonal, i.e., N vectors that have the same channel index in these orthogonal plane feature maps are reduced by QR decomposition to maintain the orthogonality, and they are learnable in the training stage. Finally, the Hadamard product of L and \mathcal{B} is channel-wisely summed to output the motion ΔP_m . Note that our face motion conditions include facial expressions and head poses.

Optimization

The goal of proposing our framework is to learn the dense correspondence. We have the assumption that different identities have different topological structures which is suitable for modeling by a StyleGAN-based generator, and the rules of topological transformations are shared among different identities which can be modeled by our PlaneDict module.

So we only adopt a self-supervision manner to train our framework. If this self-supervision successfully disentangles the deformations brought by identity and motion conditions, it precisely indicates that our aforementioned assumptions are valid, and our framework can exactly learn dense correspondence without a 3D parametric model prior.

The cross-identity driving is our ultimate goal, and we should fully utilize the paired data and learn the motion deformations with ground truth. Therefore, we adopt disentanglement loss \mathcal{L}_{dis} to model the different deformations brought identity and motion conditions, and reconstruction loss \mathcal{L}_{recons} to improve the image quality and rendering details as follows:

$$\mathcal{L} = \mathcal{L}_{dis} + \lambda_1 \mathcal{L}_{recons}. \quad (5)$$

Disentanglement Loss. We denote the input source image as I_s , the input driving image as I_d and the output driven image as I . We adopt the encoder E to extract ω_s^{id} & ω_d^{id} and ω_s^m & ω_d^m from I_s and I_d respectively. Moreover, we use E to extract the ω^{id} & ω^m from I . Our optimization goal is to minimize the distance between ω_s^{id} and ω^{id} , and the distance between ω_d^m and ω^m . Therefore, we propose our disentanglement loss \mathcal{L}_{dis} :

$$\mathcal{L}_{dis} = \|\omega_s^{id} - \omega^{id}\|_2 + \lambda_2 \|\omega_d^m - \omega^m\|_2. \quad (6)$$

Reconstruction Loss. When we conduct the experiments, we found that the regions of eyes and mouths take longer training time to learn the distribution, which are high-frequency details with small areas and big variations. Therefore, as for the reconstruction loss \mathcal{L}_{recons} , we include \mathcal{L}_1 loss and mask loss, as follows:

$$\mathcal{L}_{recons} = \|I_d - I\|_1 + \lambda_3 \|\mathcal{M}(I_d) - \mathcal{M}(I)\|_1, \quad (7)$$

where \mathcal{M} is a mask that indicates the eye and mouth part of the face image.

Training Strategy. We use two types of paired data and use different losses to optimize the network parameters. One type is to use the source image and target image from the same identity, and we adopt \mathcal{L} for training. The other is to use the source image and target image from different identities, and we only use \mathcal{L}_{dis} for training.

Experiments

Experimental Settings

Implementation Details. The encoder E is a ResNet-10 (He et al. 2016) network. The generators G_{id} and G_m are two StyleGAN-based generators (Karras et al. 2020). The Super Resolution module utilizes two blocks of StyleGAN-modulated layers. To further improve the image resolution and quality, we adopt the pre-trained dual discriminator from EG3D (Chan et al. 2022) and sample data from FFHQ (Karras, Laine, and Aila 2019) as real images to fine-tune our networks in a GAN manner. The λ_1 , λ_2 , and λ_3 are set as 1.0, 1.5, and 10. The iteration ratio of self-reenactment and cross-identity reenactment is better set at 2:1. Using Adam optimizer (set learning rate as 0.0001), the training takes about 4 days on 8 Tesla V100 GPUs while the fine-tuning takes 1 day.



Figure 3: Qualitative results of self-reenactment.

Baselines. We select five state-of-the-art methods from different perspectives, including 2D-warping-based FOMM (Siarohin et al. 2019) & DaGAN (Hong et al. 2022), mesh-based ROME (Khakhulin et al. 2022), NeRF-based FNeVR (Zeng et al. 2022), and tri-plane-based HiDeNeRF (Li et al. 2023b). For fair comparisons, these methods are trained with VoxCeleb dataset (Nagrani, Chung, and Zisserman 2017; Chung, Nagrani, and Zisserman 2018).

Datasets. We conduct experiments over three commonly used datasets: VoxCeleb1 (Nagrani, Chung, and Zisserman 2017), VoxCeleb2 (Chung, Nagrani, and Zisserman 2018), and TalkingHead-1KH (Wang, Mallya, and Liu 2021). We follow the FOMM to pre-process these videos, in which each frame is aligned and cropped into 256^2 resolution. We follow the EG3D (Chan et al. 2022) to extract camera extrinsics, which is based on an off-the-shelf pose estimator (Deng et al. 2019). Furthermore, we use face-parsing.Pytorch (zll-running 2019) to provide region masks of face, hair, and torso, and set the background region as black, which can reduce the impact of complex backgrounds. The selected videos for the test are not overlapped with the training videos.

Metrics. We evaluate different methods from three perspectives: (1) Visual quality: We adopt SSIM (Wang et al. 2004), PSNR, LPIPS (Zhang et al. 2018), and FID (Heusel et al. 2017) as quality metrics. (2) Identity fidelity and motion accuracy: Following the previous works (Ha et al. 2020; Hong

	SSIM \uparrow	PSNR \uparrow	LPIPS \downarrow
FOMM	0.690	19.2	0.112
DaGAN	0.807	23.2	0.088
FNeVR	0.901	21.1	0.092
ROME	0.833	21.6	0.085
HiDeNeRF	0.862	21.9	0.084
Ours	0.870	22.1	0.079

Table 1: Visual quality evaluation of self-reenactment.

	CSIM \uparrow	AUCON \uparrow	PRMSE \downarrow	AVD \downarrow	ET \downarrow
FOMM	0.837	0.872	2.88	0.021	1.98
DaGAN	0.875	0.921	1.79	0.016	4.08
FNeVR	0.880	0.929	2.22	0.016	2.01
ROME	0.906	0.918	1.68	0.013	5.28
HiDeNeRF	0.931	0.956	1.66	0.010	5.44
Ours	0.946	0.961	1.60	0.009	1.72

Table 2: Identity fidelity and motion accuracy evaluation of self-reenactment.

et al. 2022), we adopt CSIM, PRMSE, and AUCON to evaluate the identity preservation of the source image, the accuracy of head poses, and the precision of expression. (3) Multi-view consistency: We adopt the AVD proposed by HiDeNeRF (Li et al. 2023b) to evaluate multi-view identity preservation. Furthermore, we propose the ET (Eye Tracking) metric to evaluate fine-grained motion control of gaze, which calculate the error of eye locations (Antoine et al. 2022) between the source image and driving image (they are all aligned face images).

Self-Reenactment

The self-reenactment experiments are using the source and driving images of the same identity, which have the ground truth of the synthesized results for comparisons. As shown in Fig. 3, we show the qualitative results of different methods. Because we use the motion features extracted from the driving images, instead of 3DMM parameters, we can not only realize more fine-grained motion control than 3DMM-based methods but also handle special regions like hair and eyes. We list the quantitative results in Tab. 1 and Tab. 2, and we achieve matched or better scores than other state-of-the-art methods which are based on the correspondence from the 3DMM prior. These results show that, instead of using 3DMM parameter control at the cost of missing details, we can directly establish dense correspondence between different tri-plane representations, which overcomes the optimization burden of aligning 3DMM space and the latent space of NeRF-based generative models.

Cross-Identity Reenactment

The cross-identity reenactment is using the source and driving images of different identities, which is a more difficult challenge for source identity preservation and fine-grained motion transfer between different face tri-planes. We first qualitatively compare different state-of-the-art methods and

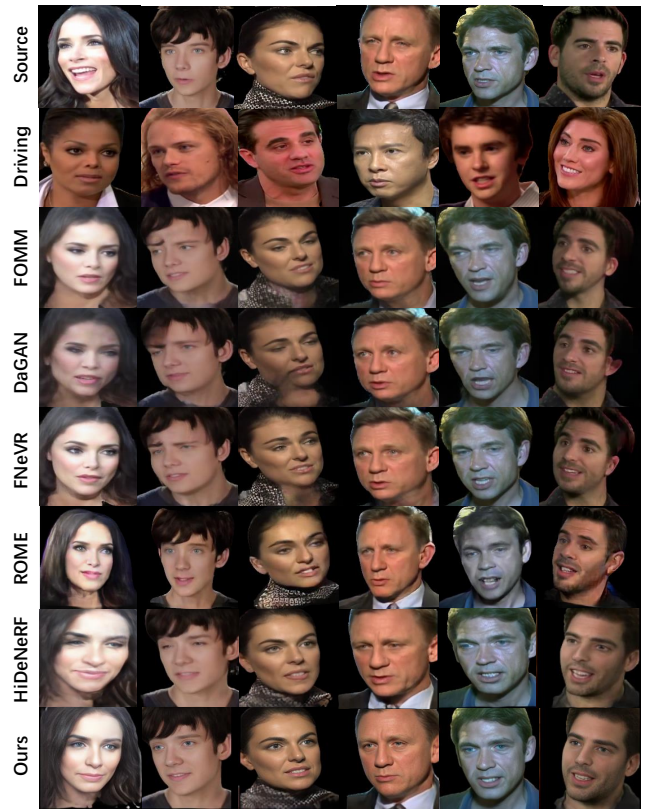
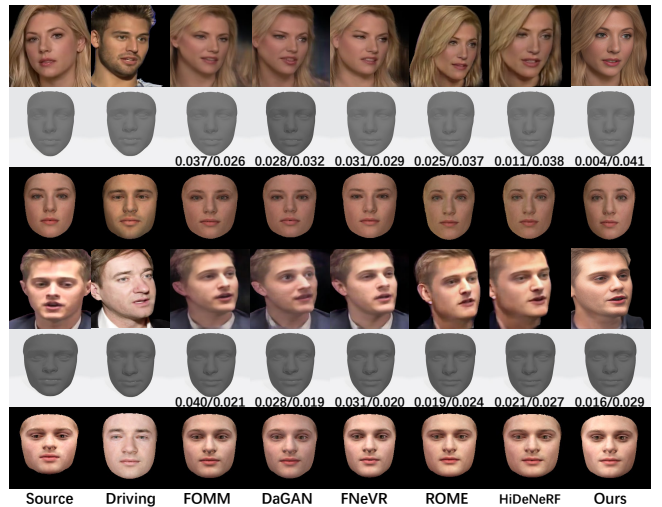


Figure 4: Qualitative results of cross-identity reenactment.

Figure 5: Mesh evaluation for identity preservation of cross-identity reenactment (vertex distance from source identity \downarrow /vertex distance from driving identity \uparrow).

show their synthesized results in Fig. 4. Although the fusion of identity and motion information is a hard problem, our framework with the PlaneDict module is able to generate cross-identity reenactment results with better image quality and identity fidelity without any artifacts.

	VoxCeleb1					
	CSIM \uparrow	AUCON \uparrow	PRMSE \downarrow	FID \downarrow	AVD \downarrow	ET \downarrow
FOMM	0.748	0.752	3.66	86	0.044	6.08
DaGAN	0.790	0.880	3.06	87	0.036	6.16
FNeVR	0.812	0.884	3.32	82	0.041	6.10
ROME	0.833	0.871	2.64	76	0.016	7.08
HiDeNeRF	0.876	0.917	2.62	57	0.012	7.02
Ours	0.911	0.928	2.50	49	0.011	5.18

	VoxCeleb2					
	CSIM \uparrow	AUCON \uparrow	PRMSE \downarrow	FID \downarrow	AVD \downarrow	ET \downarrow
FOMM	0.680	0.707	4.16	85	0.047	6.23
DaGAN	0.693	0.815	3.93	86	0.040	6.62
FNeVR	0.699	0.829	3.90	84	0.047	5.99
ROME	0.710	0.821	3.08	76	0.019	7.29
HiDeNeRF	0.787	0.889	2.91	61	0.014	7.30
Ours	0.790	0.894	2.83	58	0.012	5.33

	TalkingHead-1KH					
	CSIM \uparrow	AUCON \uparrow	PRMSE \downarrow	FID \downarrow	AVD \downarrow	ET \downarrow
FOMM	0.723	0.741	3.71	76	0.039	6.17
DaGAN	0.766	0.872	2.98	73	0.035	6.59
FNeVR	0.775	0.879	3.39	73	0.037	6.03
ROME	0.781	0.864	2.66	68	0.017	6.97
HiDeNeRF	0.828	0.901	2.60	52	0.011	7.09
Ours	0.831	0.912	2.55	49	0.010	5.42

Table 3: Cross-identity reenactment evaluation.

To further compare the identity preservation ability of our framework, as shown in Fig. 5, we evaluate and visualize the identity fidelity (i.e., the shape preservation of the source identity) using 3D face reconstruction models (Deng et al. 2019). The quantitative experiments are performed with VoxCeleb1 (Nagrani, Chung, and Zisserman 2017), VoxCeleb2 (Chung, Nagrani, and Zisserman 2018), and TalkingHead-1KH (Wang, Mallya, and Liu 2021), which are shown in Tab. 3 respectively.

Multi-View Synthesis

The one-shot multi-view cross-identity reenactment is the most challenging task. It requires not only using one face image from the source identity to construct a 3D face representation for multi-view rendering but also this representation can be controlled by motion conditions for novel expression and pose reenactment. We adopt the state-of-the-art HiDeNeRF (Li et al. 2023b) as the baseline method for comparison. As shown in Fig. 6, we render the driven results from different view directions. Our method achieves better image quality than HiDeNeRF and does not show artifacts or 3D inconsistency in some angles. As shown in Tab. 4, we further evaluate ours and HiDeNeRF in quantitative experiments. Our framework with the PlaneDict module can replace the 3DMM to model the dense correspondence between different tri-plane representations. In this way, instead of aligning the implicit spaces of the two generative models, we learn the dense correspondence without any loss of the 3D consistency of the target NeRF-based model.

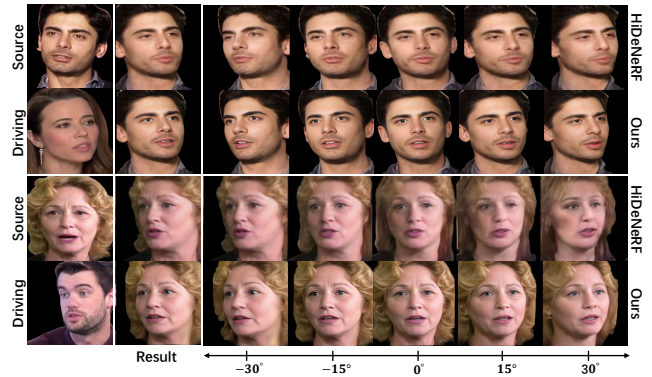


Figure 6: Qualitative results of multi-view synthesis.

	CSIM \uparrow	AUCON \uparrow	PRMSE \downarrow	AVD \downarrow
HiDeNeRF	0.829	0.864	3.78	0.014
Ours	0.840	0.881	3.53	0.008

Table 4: Quantitative evaluation of multi-view synthesis.

	CSIM \uparrow	AUCON \uparrow	PRMSE \downarrow	AVD \downarrow	ET \downarrow
w/o PlaneDict	0.763	0.809	3.10	0.035	7.58
w PlaneDict (5)	0.679	0.718	3.93	0.058	9.92
w PlaneDict (10)	0.802	0.824	3.18	0.038	7.36
w PlaneDict (15)	0.899	0.871	2.92	0.019	6.69
w PlaneDict (20)	0.911	0.928	2.50	0.011	5.18

Table 5: Ablation study.

Ablation Study

As shown in Tab. 5, in the ablation study of whether to use the PlaneDict module, we adopt the same structure as identity deformations for obtaining motion and they are trained with the same time and dataset. Limited by the hardware, the max number of plane bases is 23. However, since 20, there has been almost no improvement. Therefore, we adopt 20 as the number of plane bases in our PlaneDict module to balance the quality and optimization difficulty.

Conclusions

In this paper, we propose a novel framework to learn the dense correspondence between different face tri-planes without a 3D parametric model prior. With the PlaneDict module, our framework can achieve fine-grained motion driving of face tri-planes without any 3D inconsistency. Extensive experiments demonstrate our better image quality, fine-grained motion control, and identity fidelity of one-shot multi-view face reenactment than previous methods.

Limitations and Ethical Concerns. Due to the inherent biases in the datasets, we are not able to handle extreme poses and expressions. We strongly oppose any misuse of our technology but we believe it has the potential to achieve multi-view animation of diverse objects without relying on sophisticated 3D parametric models like human faces.

Acknowledgments

This work is supported by the National Key Research and Development Program of China under Grant No. 2021YFC3320103, the National Natural Science Foundation of China (NSFC) under Grants 62372452, U19B2038.

References

- Antoine, L.; Shuvam, K.; Labintsev; Wing-Fung, K.; and Vazquez, R. 2022. Gaze Tracking. <https://github.com/antoinelame/GazeTracking>. Eye Tracking library.
- Athar, S.; Xu, Z.; Sunkavalli, K.; Shechtman, E.; and Shu, Z. 2022. Rignerf: Fully controllable neural 3d portraits. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 20364–20373.
- Blanz, V.; and Vetter, T. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 187–194.
- Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16123–16133.
- Chung, J. S.; Nagrani, A.; and Zisserman, A. 2018. Voxceleb2: Deep speaker recognition. *INTERSPEECH*.
- Deng, Y.; Yang, J.; Xu, S.; Chen, D.; Jia, Y.; and Tong, X. 2019. 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*, 0–0.
- Dong, H.; Liang, X.; Gong, K.; Lai, H.; Zhu, J.; and Yin, J. 2018. Soft-gated warping-gan for pose-guided person image synthesis. *Advances in neural information processing systems*, 31.
- Drobyshev, N.; Chelishev, J.; Khakhulin, T.; Ivakhnenko, A.; Lempitsky, V.; and Zakharov, E. 2022. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2663–2671.
- Feng, Y.; Feng, H.; Black, M. J.; and Bolkart, T. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4): 1–13.
- Gafni, G.; Thies, J.; Zollhofer, M.; and Nießner, M. 2021. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8649–8658.
- Geng, J.; Shao, T.; Zheng, Y.; Weng, Y.; and Zhou, K. 2018. Warp-guided gans for single-photo facial animation. *ACM Transactions on Graphics (ToG)*, 37(6): 1–12.
- Gu, J.; Liu, L.; Wang, P.; and Theobalt, C. 2022. StyleNeRF: A style-based 3d-aware generator for high-resolution image synthesis. *ICLR*.
- Guo, Y.; Chen, K.; Liang, S.; Liu, Y.-J.; Bao, H.; and Zhang, J. 2021. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5784–5794.
- Ha, S.; Kersner, M.; Kim, B.; Seo, S.; and Kim, D. 2020. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 10893–10900.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.
- Hong, F.-T.; Zhang, L.; Shen, L.; and Xu, D. 2022. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3397–3406.
- Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.
- Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.
- Khakhulin, T.; Sklyarova, V.; Lempitsky, V.; and Zakharov, E. 2022. Realistic one-shot mesh-based head avatars. In *European Conference on Computer Vision*, 345–362. Springer.
- Lan, Y.; Loy, C. C.; and Dai, B. 2022. Correspondence Distillation from NeRF-based GAN. *arXiv preprint arXiv:2212.09735*.
- Lan, Y.; Meng, X.; Yang, S.; Loy, C. C.; and Dai, B. 2023. Self-Supervised Geometry-Aware Encoder for Style-Based 3D GAN Inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20940–20949.
- Li, J.; Zhang, J.; Bai, X.; Zhou, J.; and Gu, L. 2023a. Efficient Region-Aware Neural Radiance Fields for High-Fidelity Talking Portrait Synthesis. *arXiv preprint arXiv:2307.09323*.
- Li, T.; Bolkart, T.; Black, M. J.; Li, H.; and Romero, J. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 36(6): 194–1.
- Li, W.; Zhang, L.; Wang, D.; Zhao, B.; Wang, Z.; Chen, M.; Zhang, B.; Wang, Z.; Bo, L.; and Li, X. 2023b. 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)*, 17969–17978.
- Lin, C. Z.; Lindell, D. B.; Chan, E. R.; and Wetzstein, G. 2022. 3d gan inversion for controllable portrait image animation. *arXiv preprint arXiv:2203.13441*.

- Liu, W.; Piao, Z.; Min, J.; Luo, W.; Ma, L.; and Gao, S. 2019. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5904–5913.
- Liu, X.; Xu, Y.; Wu, Q.; Zhou, H.; Wu, W.; and Zhou, B. 2022. Semantic-aware implicit neural audio-driven video portrait generation. *ECCV*.
- Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.
- Ma, T.; Li, B.; He, Q.; Dong, J.; and Tan, T. 2023a. Semantic 3D-aware Portrait Synthesis and Manipulation Based on Compositional Neural Radiance Field. *arXiv preprint arXiv:2302.01579*.
- Ma, Z.; Zhu, X.; Qi, G.-J.; Lei, Z.; and Zhang, L. 2023b. OTAvatar: One-shot Talking Face Avatar with Controllable Tri-plane Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16901–16910.
- Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.
- Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. Voxceleb: a large-scale speaker identification dataset. *INTER-SPEECH*.
- Or-El, R.; Luo, X.; Shan, M.; Shechtman, E.; Park, J. J.; and Kemelmacher-Shlizerman, I. 2022. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13503–13513.
- Park, K.; Sinha, U.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Seitz, S. M.; and Martin-Brualla, R. 2021a. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5865–5874.
- Park, K.; Sinha, U.; Hedman, P.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Martin-Brualla, R.; and Seitz, S. M. 2021b. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*.
- Shen, S.; Li, W.; Zhu, Z.; Duan, Y.; Zhou, J.; and Lu, J. 2022. Learning Dynamic Facial Radiance Fields for Few-Shot Talking Head Synthesis. In *ECCV*.
- Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. First order motion model for image animation. *Advances in neural information processing systems*, 32.
- Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10039–10049.
- Wang, Y.; Yang, D.; Bremond, F.; and Dantcheva, A. 2022. Latent Image Animator: Learning to Animate Images via Latent Space Navigation. In *International Conference on Learning Representations*.
- Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.
- Xu, H.; Song, G.; Jiang, Z.; Zhang, J.; Shi, Y.; Liu, J.; Ma, W.; Feng, J.; and Luo, L. 2023. OmniAvatar: Geometry-Guided Controllable 3D Head Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12814–12824.
- Yang, S.; Wang, W.; Ling, J.; Peng, B.; Tan, X.; and Dong, J. 2023a. Context-Aware Talking-Head Video Editing. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7718–7727.
- Yang, S.; Wang, W.; Peng, B.; and Dong, J. 2023b. Designing A 3d-Aware Stylenerf Encoder for Face Editing. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.
- Zeng, B.; Liu, B.; Li, H.; Liu, X.; Liu, J.; Chen, D.; Peng, W.; and Zhang, B. 2022. FNeVR: Neural volume rendering for face animation. *Advances in Neural Information Processing Systems*, 35: 22451–22462.
- Zhang, J.; Li, X.; Wan, Z.; Wang, C.; and Liao, J. 2022. Fd-nerf: Few-shot dynamic neural radiance fields for face reconstruction and expression editing. In *SIGGRAPH Asia 2022 Conference Papers*, 1–9.
- Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.
- Zhao, J.; and Zhang, H. 2022. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3657–3666.
- Zhuang, Y.; Zhu, H.; Sun, X.; and Cao, X. 2022. Mofanerf: Morphable facial neural radiance field. In *European Conference on Computer Vision*, 268–285. Springer.
- zllrunning. 2019. face-parsing.PyTorch. <https://github.com/zllrunning/face-parsing.PyTorch>. Using modified BiSeNet for face parsing in PyTorch.