MoFaNeRF: Morphable Facial Neural Radiance Field

Yiyu Zhuang*, Hao Zhu*, Xusen Sun, and Xun Cao†

Nanjing University, Nanjing, China
{yiyu.zhuang, xusensun}@smail.nju.edu.cn {zhuhaoese, caoxun}@nju.edu.cn

Abstract. We propose a parametric model that maps free-view images into a vector space of coded facial shape, expression and appearance with a neural radiance field, namely Morphable Facial NeRF. Specifically, MoFaNeRF takes the coded facial shape, expression and appearance along with space coordinate and view direction as input to an MLP, and outputs the radiance of the space point for photo-realistic image synthesis. Compared with conventional 3D morphable models (3DMM), MoFaNeRF shows superiority in directly synthesizing photo-realistic facial details even for eyes, mouths, and beards. Also, continuous face morphing can be easily achieved by interpolating the input shape, expression and appearance codes. By introducing identity-specific modulation and texture encoder, our model synthesizes accurate photometric details and shows strong representation ability. Our model shows strong ability on multiple applications including image-based fitting, random generation, face rigging, face editing, and novel view synthesis. Experiments show that our method achieves higher representation ability than previous parametric models, and achieves competitive performance in several applications. To the best of our knowledge, our work is the first facial parametric model built upon a neural radiance field that can be used in fitting, generation and manipulation. The code and data is available at https://github.com/zhuhao-nju/mofanerf.

Keywords: neural radiance field, 3D morphable models, face synthesis.

1 Introduction

Modeling 3D face is a key problem to solve face-related vision tasks such as 3D face reconstruction, reenactment, parsing, and digital human. The 3D morphable model (3DMM)[2] has long been the key solution to this problem, which is a parametric model transforming the shape and texture of the faces into a vector space representation. 3DMMs are powerful in representing various shapes and appearances, but require a sophisticated rendering pipeline to produce photo-realistic images. Besides, 3DMMs struggled to model non-Lambertian objects

[*] These authors contributed equally to this work.

[†] Xun Cao is the corresponding author.

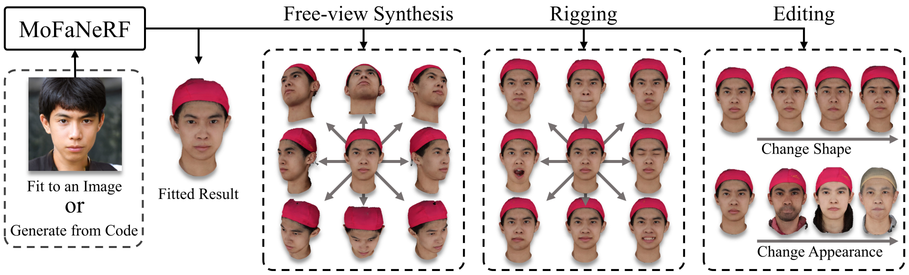

Fig. 1. We propose MoFaNeRF, which is a parametric model that can synthesize free-view images by fitting to a single image or generating from a random code. The synthesized face is *morphable* that can be rigged to a certain expression and be edited to a certain shape and appearance.

like pupils and beards. Recently, the neural radiance field (NeRF)[31] was proposed to represent the shapes and appearances of a static scene using an implicit function, which shows superiority in the task of photo-realistic free-view synthesis. The most recent progress shows that the modified NeRF can model a dynamic face[11,52,35,34], or generate diversified 3D-aware images[42,5,15]. However, there is still no method to enable NeRF with the abilities of single-view fitting, controllable generation, face rigging and editing at the same time. In summary, conventional 3DMMs are powerful in representing large-scale editable 3D faces but lack the ability of photo-realistic rendering, while NeRFs are the opposite.

To combine the best of 3DMM and NeRF, we aim at creating a facial parametric model based on the neural radiance field to have the powerful representation ability as well as excellent free-view rendering performance. However, achieving such a goal is non-trivial. The challenges come from two aspects: firstly, how to memorize and parse the very large-scale face database using a neural radiance field; secondly, how to effectively disentangle the parameters (e.g. shape, appearance, expression), which are important to support very valuable applications like face rigging and editing.

To address these challenges, we propose the Morphable Facial NeRF (MoFaNeRF) that maps free-view images into a vector space of coded facial identity, expression, and appearance using a neural radiance field. Our model is trained on two large-scale 3D face datasets, FaceScape[55,61] and HeadSpcae[8] separately. FaceScape contains 359 available faces with 20 expressions each, and HeadSpace contains 1004 faces in the neutral expression. The training strategy is elaborately designed to disentangle the shape, appearance and expression in the parametric space. The identity-specific modulation and texture encoder are proposed to maximize the representation ability of the neural network. Compared to traditional 3DMMs, MoFaNeRF shows superiority in synthesizing photo-realistic images even for pupils, mouth, and beards which can not be modeled well by 3D mesh models. Furthermore, we also propose the methods to use our model to achieve image-based fitting, random face generation, face rigging, face editing, and view extrapolation. Our contributions can be summarized as follows:

- To the best of our knowledge, we propose the first parametric model that maps free-view facial images into a vector space using a neural radiance field and is free from the traditional 3D morphable model.
- The neural network for parametric mapping is elaborately designed to maximize the solution space to represent diverse identities and expressions. The disentangled parameters of shape, appearance and expression can be interpolated to achieve a continuous and morphable facial synthesis.
- We present to use our model for multiple applications like image-based fitting, view extrapolation, face editing, and face rigging. Our model achieves competitive performance compared to state-of-the-art methods.

2 Related Work

As our work is a parametric model based on neural radiance field, we will review the related work of 3D morphable model and neural radiance field respectively.
3D Morphable Model. 3DMM is a statistical model which transforms the shape and texture of the faces into a vector space representation[2]. By optimizing and editing parameters, 3DMMs can be used in multiple applications like 3D face reconstruction[54], alignment[20], animation[59], etc. We recommend referring to the recent survey[10] for a comprehensive review of 3DMM. To build a 3DMM, traditional approaches first capture a large number of 3D facial meshes, then align them into a uniform topology representation, and finally process them with principal component analysis algorithm[55,61,48,3,26,18]. The parameter of the 3DMM can be further disengaged into multiple dimensions like identity, expression, appearance, and poses. In recent years, several works tried to enhance the representation power of 3DMM by using a non-linear mapping[1,43,46,45,7,44], which is more powerful in representing detailed shape and appearance than transitional linear mapping. However, they still suffer from the mesh representation which is hard to model fine geometry of pupils, eyelashes and hairs. Besides, traditional 3DMMs require sophisticated rendering pipelines to render photo-realistic images. By contrast, our model doesn't explicitly generate shape but directly synthesizes photo-realistic free-view images even for pupils, inner-mouth and beards.

Very recently, Yenamandra *et al.* [56] proposed to build the 3DMM with an implicit function representing facial shape and appearance. They used a neural network to learn a signed distance field(SDF) of 64 faces, which can model the whole head with hair. Similarly, our model is also formulated as an implicit function but very different from SDF. SDF still models shape while our method focuses on view synthesis and releases constraints of the shape, outperforming SDF in rendering performance by a large margin.
Neural Radiance Field. NeRF[31] was proposed to model the object or scene with an impressive performance in free-view synthesis. NeRF synthesizes novel views by optimizing an underlying continuous volumetric scene function that is learned from multi-view images.

As the original NeRF is designed only for a static scene, many efforts have been devoted to reconstructing deformable objects. Aiming at the human face

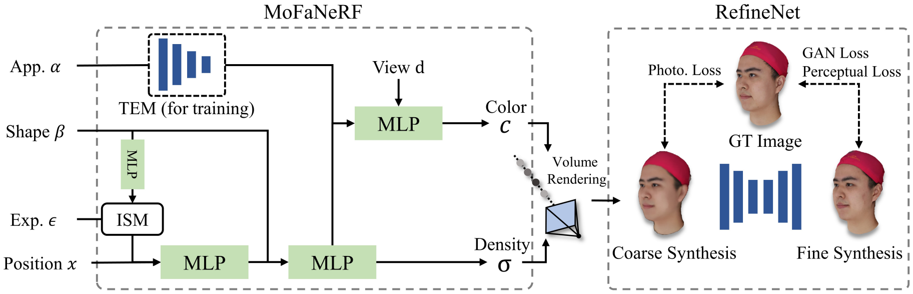

Fig. 2. MoFaNeRF takes appearance code α, shape code β, expression code ϵ, position code $\mathbf{x}$ and view direction $\mathbf{d}$ as input, synthesizing a coarse result which is then refined by a RefineNet. As shown in the right bottom corner, MoFaNeRF can be used in generating (synthesize free-view images given parameters) or fitting (optimize for parameters given a single image).

many methods[11,52,35] modeled the motion of a single human head with a designed conditional neural radiance field, extending NeRF to handle dynamic scenes from monocular or multi-view videos. Aiming at human body, several methods have been proposed by introducing human parametric model (e.g. SMPL)[33,6,29,37] or skeleton[36] as prior to build NeRF for human body. For a wide range of dynamic scenarios, Park *et al.* [34] proposed to augment NeRF by optimizing an additional continuous volumetric deformation field, while Pumarola *et al.* [39] optimized an underlying deformable volumetric function. Another group of works [42,5,15] turned NeRF into a generative model that is trained or conditioned on certain priors, which achieves 3D-aware images synthesis from a collection of unposed 2D images. To reduce the image amount for training, many works [57,49,40,12] trained the model across multiple scenes to learn a scene prior, which achieved reasonable novel view synthesis from a sparse set of views.

Different from previous NeRFs, our method is the first parametric model for facial neural radiance field trained on a large-scale multi-view face dataset. Our model supports multiple applications including random face generation, image-based fitting and facial editing, which is unavailable for previous NeRFs.

3 Morphable Facial NeRF

Morphable facial NeRF is a parametric model that maps free-view facial portraits into a continuous morphable parametric space, which is formulated as:

$$\mathcal{M} : (\mathbf{x}, \mathbf{d}, \beta, \alpha, \epsilon) \rightarrow \{\mathbf{c}, \sigma\}, \tag{1}$$

where $\mathbf{x}$ is the 3D position of a sample point; $\mathbf{d}$ is the viewing direction consisting of pitch and yaw angles; β, α, ϵ are the parameters denoting facial shape, appearance, and expression respectively; $\mathbf{c}$ and σ are the RGB color and the density used to represent the neural radiance field. In the next, we will explain $\mathbf{x}, \mathbf{d}, \mathbf{c}, \sigma$ that are referred from NeRF in Section 3.1, then introduce β, α, ϵ in Section 3.2. The network design is illustrated in Section 3.3 and the training details are explained in Section 3.4.

3.1 Neural Radiance Field

As defined in NeRF[31], the radiance field is represented as volumetric density σ and color $\mathbf{c} = (R, G, B)$. An MLP is used to predict σ and $\mathbf{c}$ from a 3D point $\mathbf{x} = (x, y, z)$ and viewing direction $\mathbf{d} = (\theta, \phi)$. Position encoding is introduced to transform the continuous inputs $\mathbf{x}$ and $\mathbf{d}$ into a high-dimensional space, which is also used in our model. The field of σ and $\mathbf{c}$ can be rendered to images using a differentiable volume rendering module. For a pixel in the posed image, a ray $\mathbf{r}$ is cast through the neural volume field from the ray origin $\mathbf{o}$ along the ray direction $\mathbf{d}$ according to the camera parameters, which is formulated as $\mathbf{r}(z) = \mathbf{o} + z\mathbf{d}$. Through sampling points along this ray, and accumulating the sampled density $\sigma(\cdot)$ and RGB values $\mathbf{c}(\cdot)$ computed by $\mathcal{F}$, the final output color $\mathbf{C}(\mathbf{r})$ of this pixel can be evaluated by:

$$\mathbf{C}(\mathbf{r}) = \int_{z_n}^{z_f} T(z)\sigma(\mathbf{r}(z))\mathbf{c}(\mathbf{r}(z), \mathbf{d})dz \text{ , where } T(z) = \exp\left(-\int_{z_n}^{z} \sigma(\mathbf{r}(s))\mathrm{d}s\right) . \quad (2)$$

$T(t)$ is defined as the accumulated transmittance along the ray from z_n to z, where z_n and z_f are near and far bounds. Through the rendered color, a photometric loss can be applied to supervise the training of the MLP.

3.2 Parametric Mapping

Our model is conditioned on the parameters to represent the identity and facial expression ϵ, and the identity is further divided into shape β and appearance α. Initially, we consider integrating β and α into a single identity code, however, we find it is hard for an MLP to memorize the huge amount of appearance information. Therefore, we propose to decouple the identity into shape and appearance. These parameters need to be disentangled to support valuable applications like face rigging and editing.

Shape parameter β represents the 3D shape of the face that is only related to the identity of the subject, like the geometry and position of the nose, eyes, mouth and overall face. A straightforward idea is to use one-hot encoding to parameterize β, while we find it suffers from redundant parameters because the similarity of large-amount faces is repeatedly expressed in one-hot code. Instead, we adopt the identity parameters of the bilinear model of FaceScape[55] as shape parameter, which is the PCA factors of the 3D mesh for each subject. The numerical variation of the identity parameter reflects the similarity between face shapes, which makes the solution space of facial shapes more efficient.

Appearance parameter α reflects photometric features like the colors of skin, lips, and pupils. Some fine-grained features are also reflected by appearance parameters, such as beard and eyelashes. Considering that the UV texture provided by FaceScape dataset is the ideal carrier to convey the appearance in a spatial-aligned UV space, we propose to encode the UV texture maps into α for training. The texture encoding module (TEM) is proposed to transfer the coded appearance information into the MLP, which is a CNN based encoder network. TEM is only used in the training phase, and we find it significantly improves the quality of synthesized images. We consider the reason is that the appearance details are well disentangled from shape and spatial-aligned, which relieves the burden of memorizing appearances for the MLP.

Expression parameter ϵ is corresponding to the motions caused by facial expressions. Previous methods[35,47] try to model the dynamic face by adding a warping vector to the position code $\mathbf{x}$, namely deformable volume. However, our experiments show that the deformable volume doesn't work in our task where too many subjects are involved in a single model. More importantly, our training data are not videos but images with discrete 20 expressions, which makes it even harder to learn a continuous warping field. By contrast, we find directly concatenating expression parameters with the position code as [27,11] causes fewer artifacts, and our identity-specific modulation (detailed in Section 3.3) further enhances the representation ability of expression. We are surprised to find that MLP without a warping module can still synthesize continuous and plausible interpolation for large-scale motions. We believe this is the inherent advantage of the neural radiance field over 2D-based synthesis methods.

3.3 Network Design

As shown in Figure 2, the backbone of MoFaNeRF mainly consists of MLPs, identity-specific modulation (ISM) module and texture encoding module(TEM). These networks transform the parameters α, β, ϵ, position code $\mathbf{x}$ and viewing direction $\mathbf{d}$ into the color $\mathbf{c}$ and density σ. The predicted colors are then synthesized from $\mathbf{c}$ and σ through volume rendering. Considering that the appearance code α is only related to the color c, it is only fed into the color decoder. The expression code ϵ is concatenated to the position code after the identity-specific modulation, as it mainly reflects the motions that are intuitively modulated by shape β. The RefineNet takes the coarse image predicted by MoFaNeRF as input and synthesizes a refined face. The results presented in this paper are the refined results by default. The additional texture encoding module (TEM) is used only in the training phase, which consists of 7 convolution layers and 5 full connected layers . The detailed parameters of our network are shown in the supplementary.

To represent a large-scale multi-view face database, the capacity of the network needs to be improved by increasing the number of layers in MLP and the number of nodes in each hidden layer. The generated images indeed gets improved after enlarging the model size, but is still blurry and contains artifacts in the expressions with large motions. To further improve the performance, we present the identity-specific modulation and RefineNet.

Identity-specific modulation (ISM). Intuitively, facial expressions of different individuals differ from each other as individuals have their unique expression idiosyncrasies. However, we observed that the MLPs erase most of these unique characteristics after the disentanglement, homogenizing the expressions from different subjects. Motivated by AdaIN[21,22], we consider the unique expression of individuals as a modulation relationship between β and ϵ, which can be formulated as:

$$\epsilon' = M_s(\beta) \cdot \epsilon + M_b(\beta), \tag{3}$$

where ϵ' is the updated value to the expression code, M_s and M_b are the shallow MLPs to transform β into an identity-specific code to adjust ϵ. Both M_s and M_b output tensors with the same length as ϵ. Our experiments show that ISM improves the representation ability of the network especially for various expressions.

RefineNet. We propose to take advantage generative adversial networks to further improve the synthesis of the facial details. We use Pix2PixHD[50] as the backbone of RefineNet, which refine the results of MoFaNeRF with GAN loss[14] and perceptual loss[19]. The input of RefineNet is the coarse image rendered by MoFaNeRF, and the output is a refined image with high-frequency details. We find that RefineNet significantly improves details and realism with less impact on identity-consistency. The influence of RefineNet on identity-consistency are validated in the ablation study in Section 4.2.

3.4 Training

Data preparation. We use 7180 models released by FaceScape[55] and 1004 models released by HeadSpace[8] to train two models respectively. In FaceScape, the models are captured from 359 different subjects with 20 expressions each. For FaceScape, we randomly select 300 subjects (6000 scans) as training data, leaving 59 subjects (1180 scans) for testing. For HeadSpace, we randomly select 904 subjects as training data, leaving 100 subjects for testing. As HeadSpace only consists of a single expression for each subjects, the expression input part of the network to train HeadSpace data is removed. All these models are aligned in a canonical space, and the area below the shoulder is removed. We render 120 images in different views for each subjects. The details about the rendering setting are shown in the supplementary.

Landmark-based sampling. In the training phase, the frequency of ray-sampling is modified besides the uniform sampling to make the network focus on the facial region. Specifically, we dectect 64 2D key-points of the mouth, nose, eyes, and eyebrows, and the inverse-projecting rays are sampled around each keypoint based on a Gaussian distribution. The standard deviation of the Gaussian distribution is set to 0.025 of the image size all our experiments. The uniform sampling and the landmark-based sampling are combined with the ratio of 2:3.

Loss function. The loss function to train MoFaNeRF is formulated as:

$$L = \sum_{\mathbf{r}\in\mathcal{R}} \left[\left\| \hat{C}_c(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2 + \left\| \hat{C}_f(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2 \right], \tag{4}$$

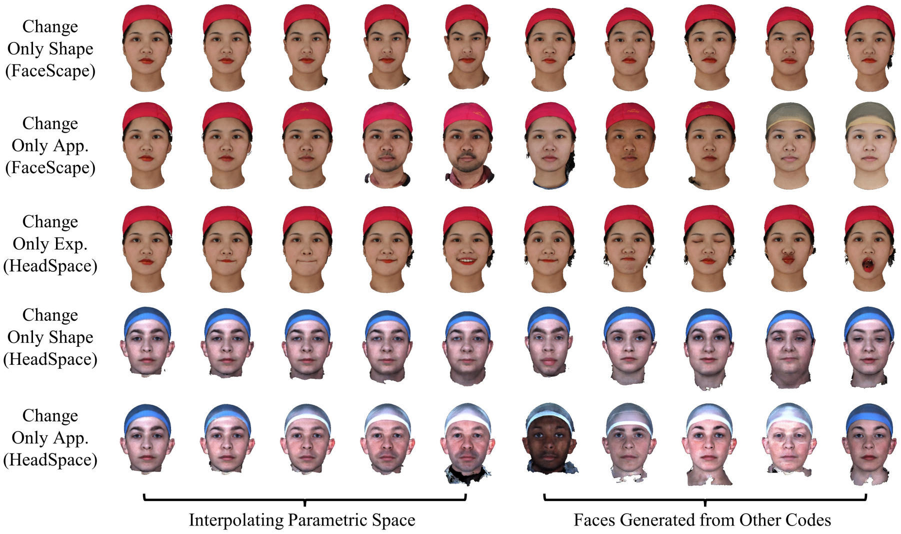
Change Only Shape (FaceScape)
Change Only App. (FaceScape)
Change Only Exp. (HeadSpace)
Change Only Shape (HeadSpace)
Change Only App. (HeadSpace)
Interpolating Parametric Space
Faces Generated from Other Codes

Fig. 3. Our model is able to synthesize diverse appearance, shape and expressions, while these three dimensions are well disentangled. The interpolation in the parametric space shows that the face can morph smoothly between two parameters.

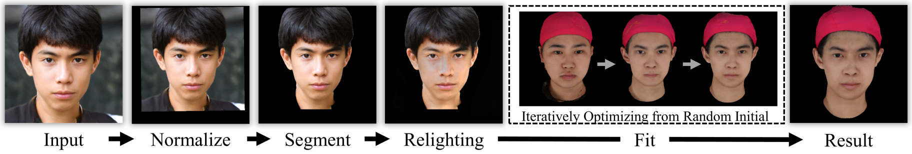

Fig. 4. The pipeline for fitting our model to a single image.

where $\mathcal{R}$ is the set of rays in each batch, $C(\mathbf{r})$ is the ground-truth color, $\hat{C}_c(\mathbf{r})$ and $\hat{C}_f(\mathbf{r})$ are the colors predicted by coarse volume and fine volume along ray $\mathbf{r}$ respectively. It is worth noting that the expression and appearance parameters are updated according to the back-propagated gradient in the training, while the shape parameters remain unchanged. We firstly train the network of MoFaNeRF, then keep the model fixed and train the RefineNet. The RefineNet is trained with the loss function following Pix2PixHD[50], which is the combination of GAN loss[14] and perceptual loss[9,13,19]. The implementation details can be found in the supplementary material.

3.5 Application

In addition to directly generating faces from a certain or random vector, MoFaNeRF can also used in image-based fitting, face rigging and editing.

Image-based fitting. As shown in Figure 4, we propose to fit our model to an input image. Firstly, we normalize the image to the canonical image with an affine transformation. Specifically, we first extract the 2D landmarks L_t from the target image with [23] , then align L_t to the predefined 3D landmarks L_c of the canonical face by solving:

$$\mathbf{d},\ \mathbf{s} = \arg\min \left\|(\mathit{\Pi}(L_c, \mathbf{d})) \cdot \mathbf{s} - L_t\right\|_2 , \tag{5}$$

where $\mathbf{d}$ is the view direction, $\mathbf{s}$ is the scale. $\mathit{\Pi}(L_c,\mathbf{d})$ is the function to project 3D points to the 2D image plane according to the view direction $\mathbf{d}$. The scale $\mathbf{s}$ is applied to the target image, and $\mathbf{d}$ is used in the fitting and remains constant. Then we use EHANet [30,24] to segment the background out, and normalize the lighting with the relighting method [60]. In practice, we find it important to eliminate the influence of light because our model cannot model complex lighting well.

After the pre-processing, we can optimize for β, α, ϵ through the network. Specifically, β and α are randomly initialized by Gaussian distribution, and ϵ is initialized with the learned value from the training. Then we freeze the pre-trained network weights and optimize α, β, ϵ through the network by minimizing only the MSE loss function between the predicted color and the target color. Only points around landmarks are sampled in fitting.

Face rigging and editing. The generated or fitted face can be rigged by interpolating in expression dimension with controllable view-point. The expression vector can be obtained by fitting to a video or manually set. Currently, we only use the basic 20 expressions provided by FaceScape to generate simple expression-changing animation. By improving the rigging of the face to higher dimensions[25], our model has the potential to perform more complex expressions. The rigged results are shown in Figure 1, Figure 3 and the supplementary materials.

The generated or fitted face can be edited by manipulating the shape and appearance code. As explained in Section 3.2, shape coder refers to the shape of the face, the geometry and position of the nose, eyes, and mouth; while appearance refers to the color of skin, lips, pupils, and fine-grained features like beard and eyelashes. These features can be replaced from face A to face B by simply replacing the shape or appearance code, as shown in Figure 1. Our model supports manually editing by painting texture map, then using TEM to generate appearance code for a generation. However, we find only large-scale features of the edited content in the texture map will take effect, like skin color and beard, while small-scale features like moles won't be transferred to the synthesized face. We also demonstrate that the face can morph smoothly by interpolating in the vector space, as shown in Figure 3.

4 Experiment

We firstly compare our model with previous parametric models in representation ability, then show the effectiveness of the parameter disentanglement and the network design in the ablation study. Finally, we evaluate the performance of MoFaNeRF in single-view image-based fitting, view extrapolation, and face manipulation.

Table 1. Quantitative evaluation of representation ability.

Model	PSNR(dB)↑	SSIM*↑	LPIPS*↓
FaceScape[55]	27.96±1.34	0.932±0.012	0.069±0.009
FaceScape*[55]	27.07±1.46	0.933±0.011	0.080±0.014
i3DMM[56]	24.45±1.58	0.904±0.014	0.112±0.015
MoFaNeRF	**31.49±1.75**	**0.951±0.010**	0.061±0.011
MoFaNeRF-fine	30.17±1.71	0.935±0.013	**0.034±0.007**

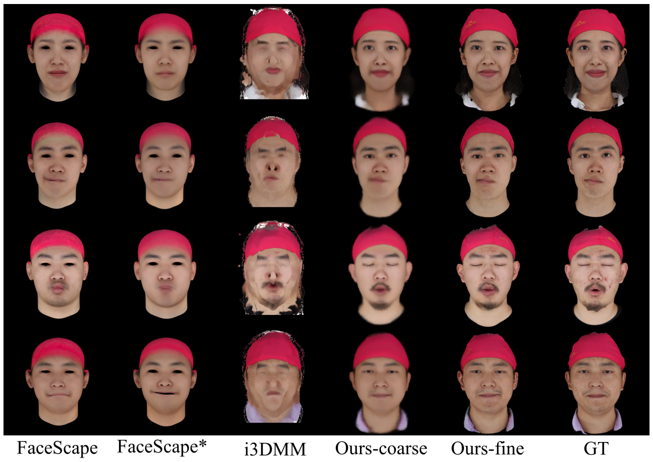

Fig. 5. Visual comparison of representation ability. Facescape* is the smaller version with comparable model size to our model(≈ 120M).

4.1 Comparison of Representation Ability

We compare the representation ability of our MoFaNeRF with two SOTA facial parametric models - FaceScape bilinear model[55] and i3DMM[56]. FaceScape is the traditional 3DMM that applies PCA to 3D triangle mesh, while i3DMM is the learning-based 3DMM that represents shape via SDF. Both models are trained on FaceScape dataset as described in Section 3.4. The default generated number of parameters for FaceScape is very large(≈ 630M), so to be fair, we also generated a model with a similar number of parameters to our model(≈ 120M), labeled as FaceScape*. PSNR[17], SSIM[51] and LPIPS[58] are used to measure the objective, structural, and perceptual similarity respectively. The better performance in similarity between the generated face image and ground truth indicates better representation ability.

From the visual comparison in Figure 5, we can see that the FaceScape bilinear model doesn't model pupils and inner mouth, as it is hard to capture accurate 3D geometry for these regions. The rendered texture is blurry due to the misalignment in the registration phase and the limited representation ability of the linear PCA model. i3DMM is able to synthesize the complete head, but the rendering result is also blurry. We observed that the performance of i3DMM

Table 2. Validation of identity consistency.

Setting	before RefineNet	after RefineNet	ground-truth
changing view	0.687±0.027	0.707±0.028	0.569±0.048
changing exp, view	0.703±0.023	0.720±0.025	0.633±0.029

trained on our training set has degraded to some extent, and we think it is because our data amount is much larger than theirs (10 times larger), which makes the task more challenging. By contrast, our model yields the clearest rendering result, which is also verified in quantitative comparison shown in Table 1. The refinement improves the LPIPS but decrease PSNR and SSIM, we believe this it is because the GAN loss and perceptual loss focus on hallucinate plausible details but is less faithful to the original image.

4.2 Disentanglement Evaluation

We show the synthesis results of different parameters in the right side of Figure 3 to demonstrate that shape, appearance and expression are well disentangled, and shown the interpolation of different attributes in the left side of Figure 3 to demonstrate that the face can morph continuously.

We further validate identity-consistency among different views and different expressions. The distance in facial identity feature space (DFID) defined in FaceNet [41] is used to measure how well the identity is preserved. Following the standard in FaceNet, two facial images with DFID ≤ 1.1 are judged to be the same person. We use a subset of our training set for this experiment, containing 10 subjects with 10 expressions each. We evaluate DFID between the ground truth and the fitted face rendered in 50 views with heading angle in $0 \sim 90°$. As reported in Table 2, the DFID scores after RefineNet are slightly increased, but are still comparable to the DFID scores of the ground-truth images. The DFID scores of both changing view and changing view and expression are much lower than 1.1, which demonstrates that the RefineNet doesn't cause severe identity inconstancy across rendering view-points.

4.3 Ablation Study

We provide ablation studies on coding strategy, morphable NeRF architecture, and sampling strategy:

• (a) Ablation Study for different coding strategy: one-hot code, PCA code, learnable code. PCA code is the PCA weights generated by the bilinear models[3,55]; Learnable code is optimized in the training of our MoFaNeRF model, which is initialized by a normal distribution. Our method adopts learnable expression code and PCA shape code, so the other 2 choices for expression and shape code are compared. Considering the appearance cannot be coded as one-hot or PCA code, we only compare with and without TEM module and coded appearance

Table 3. Ablation study.

Label	PSNR(dB)↑	SSIM↑	LPIPS↓
(a.1)One-hot expression code ϵ	25.59±2.25	0.888±0.025	0.184±0.039
(a.2)PCA expression code ϵ	25.79±2.25	0.886±0.025	0.187±0.039
(a.3)One-hot shape code β	26.27±2.25	0.895±0.024	0.174±0.039
(a.4)Leanable shape code β	25.24±2.13	0.883±0.024	0.200±0.041
(a.5)w/o appearance code α & TEM	25.73±2.03	0.889±0.025	0.184±0.039
(b.1)Deformation	24.22±2.33	0.863±0.027	0.231+0.041
(b.2)Modulation	25.69±2.22	0.886±0.025	0.187±0.039
(b.3)Hybrid	24.42±2.13	0.857±0.027	0.241+0.039
(c)Uniform Sampling	25.67±2.09	0.888±0.024	0.185±0.037
(d)Ours with on changes	**26.57±2.08**	**0.897±0.025**	**0.166±0.037**

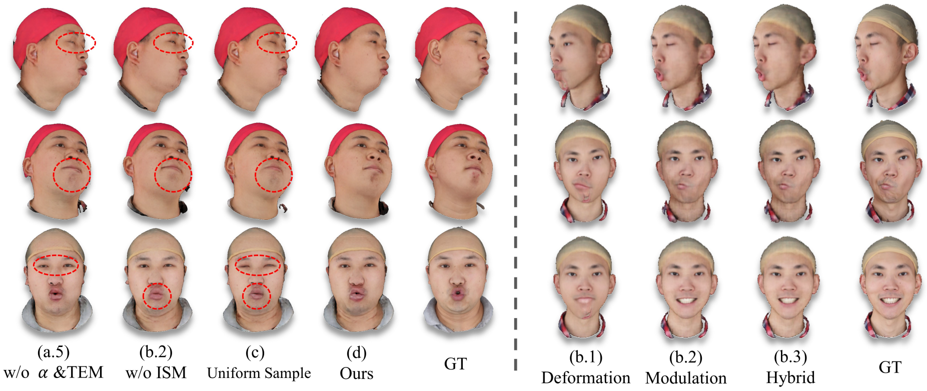

Fig. 6. Visual comparison of ablation study.

in the ablation study. The coding strategy and the TEM module are described in Section 3.2 and Section 3.3 respectively.

- (b) Ablation study for morphable NeRF architecture. Previous NeRF variants that support morphable or dynamic objects can be divided into three distinct categories: deformation-based approaches[47,34,38], modulation-based approaches[27,53,28], and a hybrid of both[35]. All of these methods were only tested for a single or a small collections, and our ablation study aims to verify their representative ability for a large-scale face dataset. We select NR-NeRF[47], Dy-NeRF[27], Hyper-NeRF[35] to represent deformation-based, modulation-based, and hybird architecture respectively.
- (c) Ablation study to verify our sampling strategy (Section 3.4). We replace our landmark-based sampling strategy with uniform sampling strategy in the training phase.
- (d) Ours with no changes.

We reconstruct 300 images of the first 15 subjects in our training set for evaluation, with random view directions and expressions. The results are reported quantitatively in Table 3 and qualitatively in Figure 6.

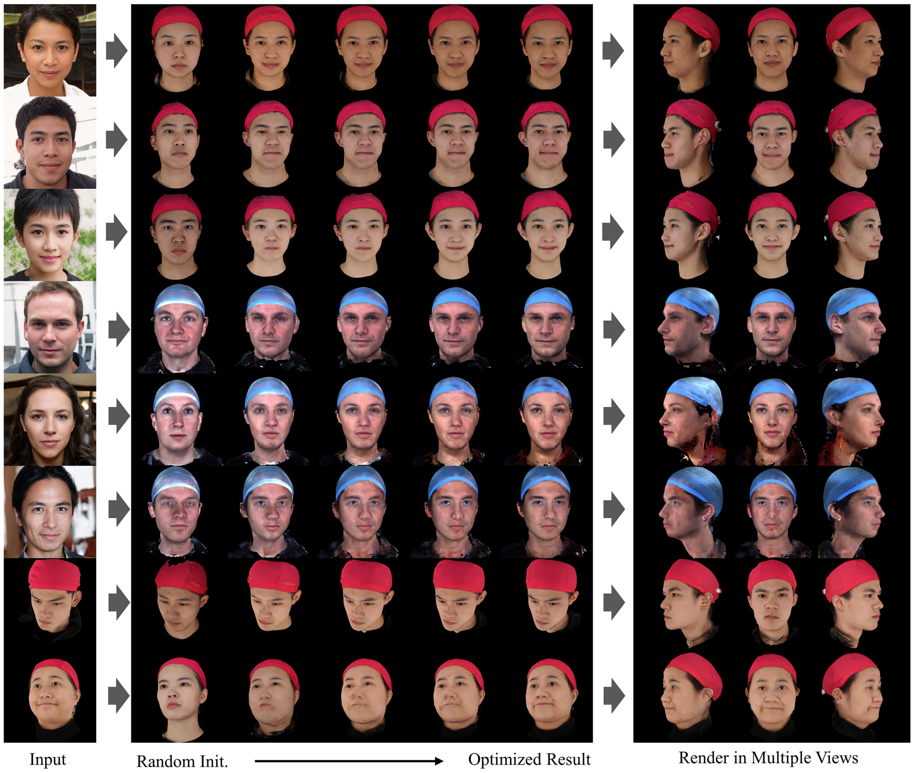

Fig. 7. The fitting results to a single-view image of MoFaNeRF. The testing image are from FaceScape testing set and in-the-wild images. The comparison with single-view face reconstruction and failure cases is shown in the supplementary material.

Discussion. As reported in Table 3, comparing (d) to (a.1) and (a.2), we find the PCA identity code most suitable for encoding shape, which reflects the shape similarity in the parameters space. Comparing (d) to (a.3) and (a.4), we can see that learnable code is most suitable for encoding expressions. We think the reason is that the categories of the expression are only 20, which is quite easy for the network to memorize and parse, while PCA code doesn't help for the few categories. By comparing items (b.1) - (b.3), we can see that modulation-based method (Dy-NeRF) shows better representative ability in modeling large-scale morphable faces, which explain the reason why our final model is based on modulation-based structure. By comparing (b.2) and (d), the positive effect of ISM module explained in Section 3.3 is verified. By comparing (c) and (d), we can see our sampling method further boost the performance. Comparing (a.5), (b.2), (c) to (d), we can see that our proposed TEM, ISM, and landmark-based sampling all have positive effects on model representation ability, and synthesize more faithful results in the visual comparison.

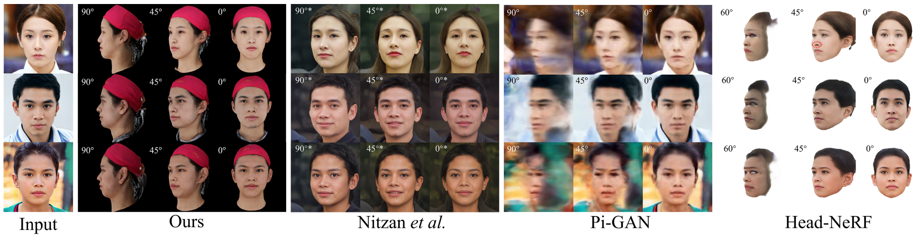

Fig. 8. The fitting and facial rotating results compared with previous methods.

4.4 Application Results

Image-based fitting. The fitted result to the testing set and in-the-wild images are shown in Figure 7. More results, comparison with single-view reconstruction methods, and failure cases can be found in the supplementary material.

Facial rotating. We fit our model to a single image and rotate the fitted face by rendering it from a side view, as shown in Figure 8. The facial rotating results is compared with Nitzan *et al.* [32], Pi-GAN[4] and HeadNeRF[16]. We can see that our method synthesizes a plausible result even at a large angle (close to ±90°) while the facial appearance and shape are maintained. Nitzan *et al.*and Pi-GAN are GAN-based networks, while HeadNeRF is a parametric NeRF trained with the help of traditional 3DMM. The results of all these three methods contain obvious artifacts when the face is rotated at a large angle.

Face rigging and editing. As shown in Figure 1 and Figure 3, after the model is fitted or generated, we can rig the face by driving the expression code ϵ, and edit the face by changing shape code β and appearance code α. Please watch our results in the video and supplementary materials.

5 Conclusion

In this paper, we propose MoFaNeRF that is the first facial parametric model based on neural radiance field. Different to the previous NeRF variants that focuses on a single or a small collection of objects, our model disentangles the shape, appearance, and expression of human faces to make the face morphable in a large-scale solution space. MoFaNeRF can be used in multiple applications and achieves competitive performance comparing to SOTA methods.

Limitation. Our model doesn't explicitly generate 3D shapes and focuses on free-view rendering performance. This prevents our model from being directly used in traditional blendshapes-based driving and rendering pipelines. Besides, the single-view fitting of MoFaNeRF only works well for relatively diffused lighting, while the performance will degrade for extreme lighting conditions. In the future work, we believe that introducing illumination model with MoFaNeRF will improve the generalization and further boost the performance.

Acknowledgement This work was supported by the NSFC grant 62025108, 62001213, and Tencent Rhino-Bird Joint Research Program. We thank Dr. Yao Yao for his valuable suggestions and Dr. Yuanxun Lu for proofreading the paper.

References

1. Bagautdinov, T., Wu, C., Saragih, J., Fua, P., Sheikh, Y.: Modeling facial geometry using compositional vaes. In: CVPR. pp. 3877–3886 (2018)
2. Blanz, V., Vetter, T., et al.: A morphable model for the synthesis of 3d faces. In: SIGGRAPH. vol. 99, pp. 187–194 (1999)
3. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. TVCG **20**(3), 413–425 (2013)
4. Chan, E., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In: CVPR (2021)
5. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In: CVPR. pp. 5799–5809 (2021)
6. Chen, J., Zhang, Y., Kang, D., Zhe, X., Bao, L., Jia, X., Lu, H.: Animatable neural radiance fields from monocular rgb videos. arXiv preprint arXiv:2106.13629 (2021)
7. Cheng, S., Bronstein, M., Zhou, Y., Kotsia, I., Pantic, M., Zafeiriou, S.: Meshgan: Non-linear 3d morphable models of faces. arXiv preprint arXiv:1903.10384 (2019)
8. Dai, H., Pears, N., Smith, W., Duncan, C.: Statistical modeling of craniofacial shape and texture. IJCV **128**(2), 547–571 (2019)
9. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. NIPS **29**, 658–666 (2016)
10. Egger, B., Smith, W.A., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., et al.: 3d morphable face models—past, present, and future. ToG **39**(5), 1–38 (2020)
11. Gafni, G., Thies, J., Zollhofer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: CVPR. pp. 8649–8658 (2021)
12. Gao, C., Shih, Y., Lai, W.S., Liang, C.K., Huang, J.B.: Portrait neural radiance fields from a single image. https://arxiv.org/abs/2012.05903 (2020)
13. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR. pp. 2414–2423 (2016)
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. NIPS **27** (2014)
15. Gu, J., Liu, L., Wang, P., Theobalt, C.: Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis (2021)
16. Hong, Y., Peng, B., Xiao, H., Liu, L., Zhang, J.: Headnerf: A real-time nerf-based parametric head model. In: CVPR. pp. 20374–20384 (2022)
17. Horé, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: ICPR. pp. 2366–2369 (2010)
18. Jiang, Z.H., Wu, Q., Chen, K., Zhang, J.: Disentangled representation learning for 3d face shape. In: CVPR. pp. 11957–11966 (2019)
19. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV. pp. 694–711. Springer (2016)
20. Jourabloo, A., Liu, X.: Large-pose face alignment via cnn-based dense 3d model fitting. In: CVPR. pp. 4188–4196 (2016)
21. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR. pp. 4401–4410 (2019)
22. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: CVPR. pp. 8110–8119 (2020)

23. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1867–1874 (2014)
24. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. arXiv preprint arXiv:1907.11922 (2019)
25. Li, H., Weise, T., Pauly, M.: Example-based facial rigging. ToG **29**, 32 (2010)
26. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ToG **36**(6), 194 (2017)
27. Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Lv, Z.: Neural 3d video synthesis. arXiv preprint arXiv:2103.02597 (2021)
28. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. https://arxiv.org/abs/2011.13084 (2020)
29. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: Neural free-view synthesis of human actors with pose control. arXiv preprint arXiv:2106.02019 (2021)
30. Luo, L., Xue, D., Feng, X.: Ehanet: An effective hierarchical aggregation network for face parsing. Applied Sciences **10**(9), 3135 (2020)
31. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV. pp. 405–421 (2020)
32. Nitzan, Y., Bermano, A., Li, Y., Cohen-Or, D.: Face identity disentanglement via latent space mapping. ToG **39**, 1 – 14 (2020)
33. Noguchi, A., Sun, X., Lin, S., Harada, T.: Neural articulated radiance field. arXiv preprint arXiv:2104.03110 (2021)
34. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. arXiv preprint arXiv:2011.12948 (2020)
35. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. arXiv preprint arXiv:2106.13228 (2021)
36. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Bao, H., Zhou, X.: Animatable neural radiance fields for human body modeling. arXiv preprint arXiv:2105.02872 (2021)
37. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: CVPR (2021)
38. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-NeRF: Neural radiance fields for dynamic scenes. https://arxiv.org/abs/2011.13961 (2020)
39. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: CVPR. pp. 10318–10327 (2021)
40. Raj, A., Zollhofer, M., Simon, T., Saragih, J., Saito, S., Hays, J., Lombardi, S.: Pixel-aligned volumetric avatars. In: CVPR. pp. 11733–11742 (2021)
41. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR. pp. 815–823 (2015)
42. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: Graf: Generative radiance fields for 3d-aware image synthesis. In: CVPR (2021)
43. Tewari, A., Zollhöfer, M., Garrido, P., Bernard, F., Kim, H., Pérez, P., Theobalt, C.: Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In: CVPR. pp. 2549–2559 (2018)

44. Tran, L., Liu, F., Liu, X.: Towards high-fidelity nonlinear 3d face morphable model. In: CVPR. pp. 1126–1135 (2019)
45. Tran, L., Liu, X.: Nonlinear 3d face morphable model. In: CVPR. pp. 7346–7355 (2018)
46. Tran, L., Liu, X.: On learning 3d face morphable model from in-the-wild images. PAMI (2019)
47. Tretschk, E., Tewari, A., Golyanik, V., Zollhöfer, M., Lassner, C., Theobalt, C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a deforming scene from monocular video. https://arxiv.org/abs/2012.12247 (2020)
48. Vlasic, D., Brand, M., Pfister, H., Popović, J.: Face transfer with multilinear models. ToG **24**(3), 426–433 (2005)
49. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: CVPR. pp. 4690–4699 (2021)
50. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: CVPR. pp. 8798–8807 (2018)
51. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP **13**(4), 600–612 (2004)
52. Wang, Z., Bagautdinov, T., Lombardi, S., Simon, T., Saragih, J., Hodgins, J., Zollhofer, M.: Learning compositional radiance fields of dynamic human heads. In: CVPR. pp. 5704–5713 (2021)
53. Xian, W., Huang, J.B., Kopf, J., Kim, C.: Space-time neural irradiance fields for free-viewpoint video. In: CVPR. pp. 9421–9431 (2021)
54. Xiao, Y., Zhu, H., Yang, H., Diao, Z., Lu, X., Cao, X.: Detailed facial geometry recovery from multi-view images by learning an implicit function. In: AAAI (2022)
55. Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., Cao, X.: Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In: CVPR (2020)
56. Yenamandra, T., Tewari, A., Bernard, F., Seidel, H.P., Elgharib, M., Cremers, D., Theobalt, C.: i3dmm: Deep implicit 3d morphable model of human heads. In: CVPR. pp. 12803–12813 (2021)
57. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: CVPR. pp. 4578–4587 (2021)
58. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)
59. Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: CVPR. pp. 3661–3670 (2021)
60. Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.W.: Deep single-image portrait relighting. In: CVPR. pp. 7194–7202 (2019)
61. Zhu, H., Yang, H., Guo, L., Zhang, Y., Wang, Y., Huang, M., Shen, Q., Yang, R., Cao, X.: Facescape: 3d facial dataset and benchmark for single-view 3d face reconstruction. arXiv preprint arXiv:2111.01082 (2021)