

3DMM-RF: Convolutional Radiance Fields for 3D Face Modeling

Stathis Galanakis^{1,2}, Baris Gecer², Alexandros Lattas^{1,2}, and Stefanos Zafeiriou^{1,2}

¹Imperial College of London

²Huawei

Abstract

Facial 3D Morphable Models are a main computer vision subject with countless applications and have been highly optimized in the last two decades. The tremendous improvements of deep generative networks have created various possibilities for improving such models and have attracted wide interest. Moreover, the recent advances in neural radiance fields, are revolutionising novel-view synthesis of known scenes. In this work, we present a facial 3D Morphable Model, which exploits both of the above, and can accurately model a subject’s identity, pose and expression and render it in arbitrary illumination. This is achieved by utilizing a powerful deep style-based generator to overcome two main weaknesses of neural radiance fields, their rigidity and rendering speed. We introduce a style-based generative network that synthesizes in one pass all and only the required rendering samples of a neural radiance field. We create a vast labelled synthetic dataset of facial renders, and train the network on these data, so that it can accurately model and generalize on facial identity, pose and appearance. Finally, we show that this model can accurately be fit to “in-the-wild” facial images of arbitrary pose and illumination, extract the facial characteristics, and be used to re-render the face in controllable conditions.

1. Introduction

Photorealistic 3D face modeling and reconstruction from a single image is a widely researched field in computer vision due to its numerous applications, such as avatar creation, virtual makeup and speech-driven face animation. Since the introduction of 3D Morphable Models (3DMM) in 1999 by the seminal work of Blanz and Vetter [3] to model faces by statistical linear models, the majority of the research stroll around linear representation of faces. Many follow-up works have studied the integration of 3DMMs with Deep Neural Networks [72, 78], Mesh Convolutions [52], and Generative Adversarial Net-

works (GANs) [22, 23, 25] in order to improve its representational strength in high-frequency details and photorealism. In this study, we explore the potential of recently emerged non-linear representation approach called Neural Radiance Fields (NeRFs) [51] to evolve 3DMMs into its strong neural volumetric representation.

NeRFs [51] have recently demonstrated an immense progress in novel-view synthesis [51], relighting [8, 71], and reconstruction [14]. They consist of fully connected neural networks that learn to implicitly represent a scene and its appearance parameters. Such networks can be optimized by using only a few dozens camera views of the scene and can be queried to generate novel views of that scene. Despite their photorealistic renderings and high 3D consistency, most initial NeRF-based methods focus on modeling and overfitting on a single subject, at a single pose.

On the contrary, methods like the 3D Morphable Model(3DMM) [3], create a statistical model of human faces by hundreds or thousands of 3D scans, and can be used as a prior to reconstructing a 3D face from a single image, with controllable identity, pose and appearance. Despite their flexibility and the tremendous improvements in their ability to generate photorealistic faces [22], they are difficult to re-create authentic facial renderings and their rendering relies on expensive skin shading models [43].

Currently, there has been tremendous progress in high-resolution 2D image generation, using generative adversarial networks [38, 40]. Such methods have been shown to successfully generate photorealistic human faces [38, 40]. Extensions have been proposed, which combined with a 3DMM, can even generate facial images with controllable facial, camera and environment attributes [73]. However, the domain of such methods lies in the 2D space and attempting to freely interpolate camera or illumination parameters creates unwanted artifacts.

In this work, we create an implicit 3D Morphable Model, by leveraging a powerful style-based deep generator network, with a flexible radiance field representation and volume rendering module. Since NeRFs heavily overfit to a

single scene, they are not constrained by the capacity of the neural component. On the other hand, generic models, such GANs and 3DMM, compress information fed into the model from hundreds of thousands of data samples. In order to build a generic NeRF representation, we propose an architecture that can predict the whole volumetric representation in a single inference. The architecture utilizes deconvolutional layers as in a traditional generator network, thus provides spatial consistency, computational and memory efficiency so that it can model a large-scale dataset. In order to unlock the full potential of our approach, we generate a synthetic dataset including 10,000 photorealistically rendered 3D faces with identity, expression, pose and illumination variations. Therefore, the proposed volumetric representation model can disentangle attributes such as identity, expression, illumination and pose, exposing an advantage over single scene/object [51] and large-scale unsupervised [13] approaches. In summary, our main contributions are:

- We present 3DMM-RF, a controllable parametric face model inspired by Neural Radiance Fields.
- We show that the proposed 3DMM-RF model can be used for 3D reconstruction of “in-the-wild” face images.
- Our convolutional generator architecture with depth-based sampling strategy can generate the samples required for rendering a volumetric radiance field, in a single pass.
- We introduce a large-scale synthetic dataset that disentangles identity, expression, camera and illumination.

2. Related Work

2.1. 3D Face Modeling

Ever since the original 3D Morphable Model introduced by [3], there have been many studies in that direction such as extension to facial expressions [1, 9, 10, 12, 45, 46, 76, 87], large-scale dataset releases [7, 16, 59, 70], reconstruction by deep regression networks [25, 63, 75, 77], and reconstruction by analysis-by-synthesis with more advanced features [5, 22, 23]. Due to its linear nature, the original 3DMMs under-represent the high-frequency information and often result in overly-smoothed geometry and texture models. In terms of preserving photorealism and high-frequency signals, non-linear generative models have been shown to be very successful in 2D image synthesis [28, 37, 41, 81], thus, non-linear 3D face modeling has been widely studied in the context of deep generative networks [72, 74, 79], GANs [22, 23, 42], and VAEs [2, 49, 64, 85]. We refer the reader to [69, 96] for a more detailed presentation of the 3D face modeling approaches.

2.2. Neural Radiance Fields

Original NeRF paper [51] showed promising results for becoming an improvement to the conventional scene representations e.g. point-clouds, meshes. This and its extensions [50, 91] can implicitly represent scenes by using a differentiable neural rendering algorithm, without the need of 3D supervision. One of the drawbacks of NeRF approaches is the computationally heavy rendering process and its long training time. Many approaches have been introduced focusing on reducing training and inference time [21, 32, 66, 90]. Other approaches focus on 3D shape reconstruction [54, 82, 88], human bodies registration [60, 61, 86], dealing with non-static scenes [55, 62], scene editing [35, 48] and scene relighting [8, 71]. NeRF-based networks’ applications have been extended on representing human faces based on a video [19, 31, 55, 56, 83] or a single image [20, 33, 95]. The recent work i3DMM [89] is the first to represent a 3DMM using an implicit function, using signed distance functions (SDF). Our method can be categorized as the latter type of networks and its application is quite similar to [95] and [33]. Our main distinction is the fact that both of them perform a low-dimension rendering and then improve the image quality through upsampling layers [33] or by a Refine network [95], whereas our method doesn’t require any optimization after the rendering step.

2.3. Deep Generative Models

The impressive photorealistic results of the GAN paper [27] resulted in its widespread application [36, 38–40, 69]. Currently, there is a large effort of combining NeRF approaches with GANs starting with pi-GAN [15] and GRAF [67]. Most recent approaches like GIRAFFE [53], StyleNerf [29], CIPS-3D [93] and EG-3D [14] have shown remarkable image quality results. GIRAFFE [53] and StyleNerf [29] are two-stage networks, having a Multi-Layer Perceptron (MLP) at low resolution and then up-sampling, CIPS-3D [94] synthesizes each pixel independently whereas EG-3D [14] introduces a 3D aware generative model base on the tri-plane representation. A distinction of our method is that these approaches use an MLP which predicts feature vectors while ours renders RGB α output directly. Also, some of these approaches require a two-step sampling procedure, while 3DMM-RF uses a depth-based sampling strategy, which needs only a single pass.

3. 3D Face Model by Neural Radiance Fields

In this work, we describe an implicit parametric facial model, with disentangled identity, pose and appearance. The NeRF representation is unfit for such a task, as a) it represents a rigid scene, b) optimizing a scene with a

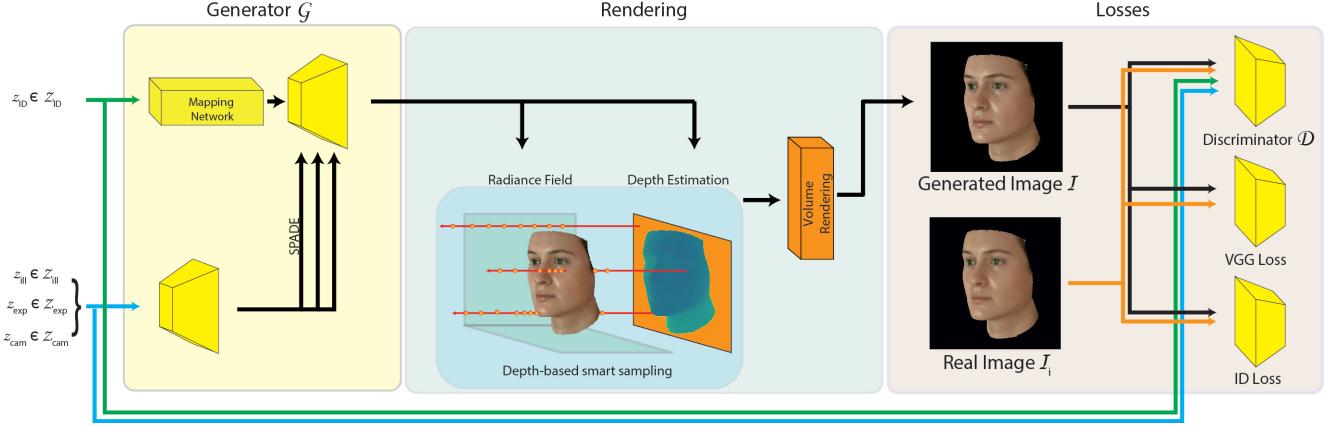


Figure 1. Overview of 3DMM-RF, a generative network which creates the sampling space of a neural radiance field depicting a human face, in one pass, based on the depths acquired from the depth prediction and then renders an image directly using the volume rendering algorithm [51]. The Generator \mathcal{G} is conditioned on a subject’s identity vector \mathbf{z}_{ID} , as well as the expression parameters \mathbf{z}_{exp} , the scene illumination parameters \mathbf{z}_{ill} and the camera parameters \mathbf{z}_{cam} . \mathbf{z}_{ID} passes through a mapping network, while the rest are transformed by SPADE layers before being given as input to \mathcal{G} . Based on a ray-face intersection depth estimation, the generator directly generates the samples needed for volume rendering which speeds up rendering. Finally, the network is trained on a synthetic dataset, using adversarial, perceptual (VGG) and identity losses.

large number of identities, poses and appearance requires an intractable optimization. In this manner, we introduce 3DMM-RF, a model that can represent and render a controllable non-rigid facial scene, using a style-based generator [40], that generates an instance of an implicit neural radiance field. Moreover, 3DMM-RF learns to approximate the area of dense samples for each view, so that a rendering can be achieved with a single query of the network.

Considering a 3D space as a neural rendering space, 3DMM-RF is a neural morphable model \mathcal{S} that renders a facial image $\mathbf{I} \in \mathbb{R}^{512 \times 512 \times 3}$ as follows:

$$\mathbf{I} = \mathcal{S}(\mathbf{z}_{ID}, \mathbf{z}_{exp}, \mathbf{z}_{cam}, \mathbf{z}_{ill}) \quad (1)$$

where $\mathbf{z}_{ID} \in \mathbb{R}^{512}$ describes an identity latent code, $\mathbf{z}_{exp} \in \mathbb{R}^{20}$ the expression 3DMM blendshapes, $\mathbf{z}_{cam} \in \mathbb{R}^3$ the camera position, whereas $\mathbf{z}_{ill} \in \mathbb{R}^8$ the illumination parameters. The 3DMM-RF model \mathcal{S} consists of style-based generator \mathcal{G} that generates a volumetric radiance field, a volume rendering module, and a discriminator \mathcal{D} . An overview of the method is shown in Fig. 1, the network architecture is presented in Sec. 3.1, the training using the synthetic dataset in Sec. 3.2, and the fitting process in Sec. 3.3.

3.1. The Architecture

3.1.1 Scene Representation

Consider a neural radiance field (NeRF) [51] with an implicit function $F_\Theta(x, y, z, \theta, \phi) \rightarrow (r, g, b, \sigma)$, that maps a world coordinate (x, y, z) and a viewing direction (θ, ϕ) , to a colour (r, g, b) with density σ . A rigid scene can be im-

plicitly learned by the radiance field, by fitting F_Θ on different views of the scene. To render a novel view, F_Θ is queried using hierarchical sampling, for a number of points on each ray intersecting the camera, and volume rendering [51] is used to produce the final colour.

Contrary to the typical NeRF approach [51] which implicitly represents a scene with a trained MLP, we train a convolutional generator network \mathcal{G} (Sec. 3.1.3) that concurrently generates all necessary K samples for each ray that passes through each pixel of the rendered image I , in a single pass. Each sample contains colour r, g, b and density σ values, like the sample vectors in NeRF [51].

3.1.2 Depth-based Smart Volume Rendering

NeRF rendering is plagued by its computational load since, typically per pixel, firstly a uniform sampling step is required to find the densest area, and another to extract a dense sampling area for volume rendering. For our approach, a given ray r will terminate as soon as it intersects with the facial surface. This means that samples, which are a bit far from the surface, will not contribute to the final RGB values. Based on that idea, we significantly speed-up rendering, by predicting the ray-face intersection depth D_r , for the ray r and sample only around this area. The predicted depth D_r is modelled as a Gaussian distribution with mean D_{μ_r} and standard variation D_{std_r} . Overall, for a ray r containing K samples, our model generates N channels, where $N = 4K + 2$, which the first $4K$ channels represent each sample across the ray r , and two additional channels

for the depth prediction. Throughout the layers of the generator, sample values ($4K$) evolve together with depth estimation, meaning that they are highly correlated thanks to the ground truth provided by the synthetic data. Thus, generated sampled values are aligned with the estimated depth values in the final radiance field.

Therefore, by generating only all the required samples along the ray that are close to the facial surface, we can directly employ the volume rendering, bypassing the multiple MLP queries required by NeRF and its importance sampling strategy [51]. For the single ray r , we separate the depth prediction channels and reshape the rest into $K \times 4$ samples along the ray. Each sample $k \in 1 \dots K$, includes the predicted $\mathbf{c}_k = (r_k, g_k, b_k)$ colour values and their corresponding density σ_k . Using the predicted depth mean D_{μ_r} and standard variation D_{std_r} to sample the depth values, the final colour $\hat{\mathbf{C}}_r$ is measured using standard volume rendering [51]:

$$\begin{aligned} \hat{\mathbf{C}}_r &= \sum_{k=1}^K w_k \mathbf{c}_k, \\ w_k &= T_k (1 - \exp(-\sigma_k (t_{k+1} - t_k))), \\ T_n &= \exp \left(- \sum_{k=1}^{K-1} \sigma_k (t_{k+1} - t_k) \right), \\ t_i &\sim \mathcal{N}(D_{\mu_r}, D_{std_r}) \text{ and } t_{k+1} \geq t_k \end{aligned} \quad (2)$$

w_k is each sample's contribution and t_k is its depth.

3.1.3 Convolutional Radiance Field Generator

We extend the above single-ray prediction approach, to the prediction of all the rays required to render an image $\mathbf{I} \in \mathbb{R}^{512 \times 512 \times 3}$. In this manner, we introduce a generator network \mathcal{G} that generates all the required samples of neural radiance field in a single pass, as it is shown in Fig. 1. The capacity of the deep convolutional generator enables it to generate the radiance field of arbitrary scenes, conditioned on certain attributes. The generator consists of a fully-convolutional mapping network [40], which translates a subject's identity vector \mathbf{z}_{ID} to an extended latent space identity vector w_{ID} , and a synthesis network that generates all the ray samples of interest.

In addition to the extended latent space identity vector w_{ID} , we condition the generator with properties relating to the reconstructed scene, namely the blendshape expression vector \mathbf{z}_{exp} , the camera position \mathbf{z}_{cam} and the scene illumination parameters \mathbf{z}_{ill} , which include the light source direction, the diffuse, specular and ambient intensity. We also apply positional encoding [51] to the camera position and the light source direction vectors. We observed that the generator G performs better when it receives the scene parameters as a 2-dimensional vector instead of 1-dimensional. This is achieved by firstly passing the parameters $\mathbf{z}_{exp}, \mathbf{z}_{cam}, \mathbf{z}_{ill}$

through two modulated convolutional layers [40] which are then fed to the network through SPADE layers [57].

Following the typical adversarial training strategy [28, 40], a discriminator \mathcal{D} is needed to achieve photorealistic results. We follow the architecture of the StyleGAN2 [40] discriminator using the conditional strategy introduced by [38]. Moreover, the discriminator is conditioned by both identity \mathbf{z}_{ID} and the rest parameters $\mathbf{z}_{exp}, \mathbf{z}_{cam}, \mathbf{z}_{ill}$, which are fed to it as a conditional label.

3.2. Training by Synthetic Face Dataset

Given the lack of accurately labelled facial images with paired depth maps and labelled variations in illumination, expression and pose, we train the proposed architecture using a synthetically generated dataset. This vast training dataset consists of arbitrary photorealistically rendered facial images, which are based on a 3D Morphable Model. They are also paired with metadata concerning their characteristics and rendering, information difficult to acquire for “in-the-wild” images. Despite using synthetic data, our generator’s finetuning step (Sec. 3.3) accurately captures real facial images.

We generate a synthetic dataset using a morphable model (LSFM [6]), a facial texture generator (TGAN [22]) and a facial material network (AvatarMe++ [43]). Specifically, we draw 10,000 facial shape samples from LSFM, and another 10,000 facial texture samples from TBGAN based on an identity vector. These samples are purposefully uncorrelated to increase subject diversity. However, the textures lack photorealistically rendering material properties, and thus we pass them through an image-to-image translation network (AvatarMe++), which translates each facial texture to spatially varying (a) diffuse albedo, (b) specular albedo, (c) diffuse normals and (d) specular normals. As shown in [43], these can be used for photorealistically rendering facial images under arbitrary environment conditions using shaders such as the Blinn-Phong [4], which we implement in PyTorch3D [43, 65]. Moreover, we sample various real expression blendshapes \mathbf{Z}_{exp} from 4DFAB [12], which correspond to our 3DMM. Finally, we define a space of the environment illumination \mathbf{Z}_{ill} with plausible RGB values and random direction for n_l light sources, and plausible RGB values for ambient illumination, and a space of the frontal-hemisphere rendering camera parameters \mathbf{Z}_{cam} (up to 30° on each axis). The models we use [6, 22, 43] are trained on diverse datasets (see each paper for details), and their trained models or training data are in the public domain.

During training, we use the above to generate vast amounts of training data. The complete dataset rendering function is defined as $\mathcal{R}(\hat{\mathbf{z}}_{ID}, \mathbf{z}_{exp}, \mathbf{z}_{ill}, \mathbf{z}_{cam}) \rightarrow \mathbb{R}^{w,h,4}$, where $\hat{\mathbf{z}}_{ID}$ is an identity vector, \mathbf{z}_{exp} is an expression vector, \mathbf{z}_{ill} is an illumination vector, \mathbf{z}_{cam} is a camera vector and w, h is the image shape. The output’s 3 channel are the

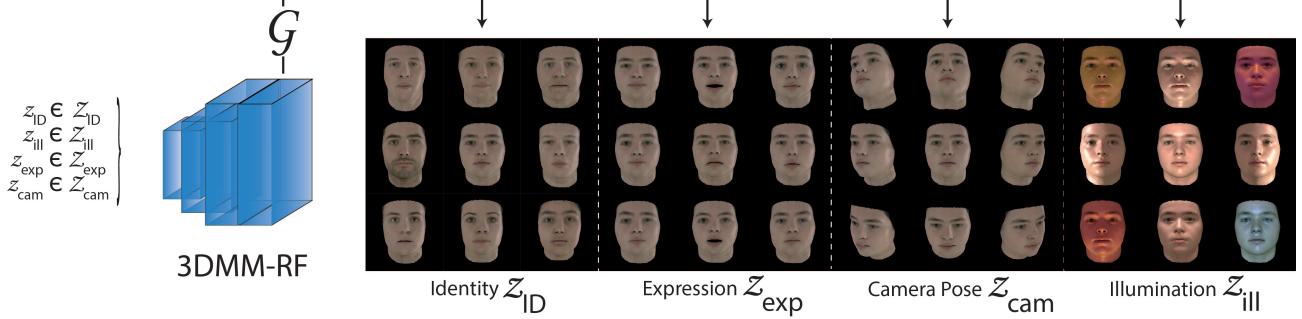


Figure 2. Once trained, 3DMM-RF has learned disentangled latent representations for facial attributes and scene parameters, and can easily be queried to render facial images. From left to right, we show samples of a) subject identity $\mathbf{z}_{ID} \in \mathcal{Z}_{ID}$, b) subject expression $\mathbf{z}_{exp} \in \mathcal{Z}_{exp}$, c) camera pose $\mathbf{z}_{cam} \in \mathcal{Z}_{cam}$, and d) scene illumination $\mathbf{z}_{ill} \in \mathcal{Z}_{ill}$.

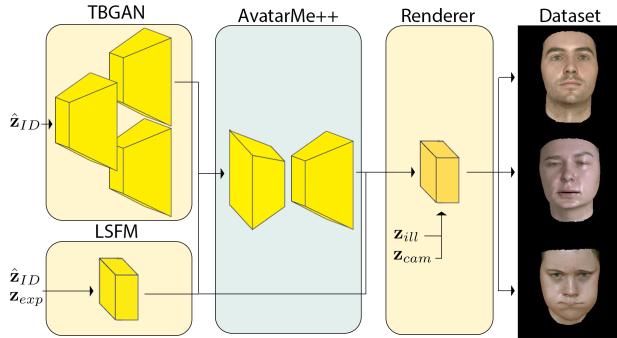


Figure 3. Overview of our dataset generation pipeline: For each training image, its identity vector $\hat{\mathbf{z}}_{ID}$ and expression vector \mathbf{z}_{exp} are used to generate diverse facial textures and shapes with TB-GAN [22] and LSFN [6]. Then AvatarMe++ [43] is used to acquire reflectance textures, which are photorealistically rendered with PyTorch3D [65], given then illumination \mathbf{z}_{ill} and camera \mathbf{z}_{cam} vectors.

RGB rendering facial image, and the last channel is the rasterized camera-space depth, using SoftRasterizer [47]. For each training iteration i , we generate each labelled image and its depth, $\mathcal{R}(\hat{\mathbf{z}}_{ID_i}, \mathbf{z}_{exp_i}, \mathbf{z}_{ill_i}, \mathbf{z}_{cam_i}) = \{\mathbf{I}_i, \mathbf{D}_i\}$, by generating a random identity vector $\hat{\mathbf{z}}_{ID_i}$, and drawing arbitrary expression blendshapes $\mathbf{z}_{exp_i} \in \mathcal{Z}_{exp}$, camera params $\mathbf{z}_{cam_i} \in \mathcal{Z}_{cam}$ and illumination params $\mathbf{z}_{ill_i} \in \mathcal{Z}_{ill}$ from their sets. Finally, to increase compatibility with “in-the-wild” images, we use a state-of-the-art face recognition network [17] on the rendered image \mathbf{I}_i , to acquire the latent identity vector \mathbf{z}_{ID_i} which we give to our model instead of $\hat{\mathbf{z}}_{ID_i}$, to facilitate the use of an identity loss.

3.3. Fitting for 3D Face Reconstruction

We can utilize our 3DMM-RF model to 3D reconstruct any single “in-the-wild” image. Given any face image, our goal is to get the optimal latent vectors for identity \mathbf{z}_{ID} , expression \mathbf{z}_{exp} , camera position \mathbf{z}_{cam} , and illumination \mathbf{z}_{ill}

that can reconstruct the target face. Firstly, we align the target face by using a face detector network [30] and also acquire its identity latent vector $\hat{\mathbf{z}}_{ID}$ by a face recognition network [18]. Then by using a facial landmark detection network [11], we extract the 2D face landmarks and get a face mask based on these landmarks. We initialize the input identity latent vector \mathbf{z}_{ID} as the identity latent vector $\hat{\mathbf{z}}_{ID}$, and set the expression parameters \mathbf{z}_{exp} to a zero vector, camera position \mathbf{z}_{cam} to a frontal view and illumination \mathbf{z}_{ill} to an average illumination setting. During the optimization, we first freeze our network and optimize the \mathbf{z} parameters where $\mathbf{z} = \{\mathbf{z}_{ID}, \mathbf{z}_{exp}, \mathbf{z}_{cam}, \mathbf{z}_{ill}\}$, by comparing the final rendered image and the target image by a combination of loss functions as follows:

$$L_{fitting} = L_{photometry} + L_{vgg} + L_{ID} + L_{landmarks} \quad (3)$$

where $L_{photometry}$ is the MSE loss (Photometry) between the rendered and the final image, L_{vgg} is the perceptual loss as introduced in [92], L_{ID} is the loss between the identity feature maps based on [18] and $L_{landmarks}$ is the facial landmarks loss. We define $L_{landmarks}$ as the L_2 -distance between the activation maps of both images which are fed into the facial landmark network [11]. To further optimize the reconstruction of the face, we finetune the generator network \mathcal{G} parameters in addition to the input parameters \mathbf{z} by the same loss function with 200-fold smaller learning rate. This approach helps to recover identity more precisely and helps to bridge the domain gap between the synthetic training data and “in-the-wild“ images.

4. Experiments

4.1. Disentanglement control

One of the goals of the training is the network to be able to efficiently disentangle its latent space. This means that each dimension of the resulting subspace affects a different variation factor. As shown in Fig. 4, while modifying one

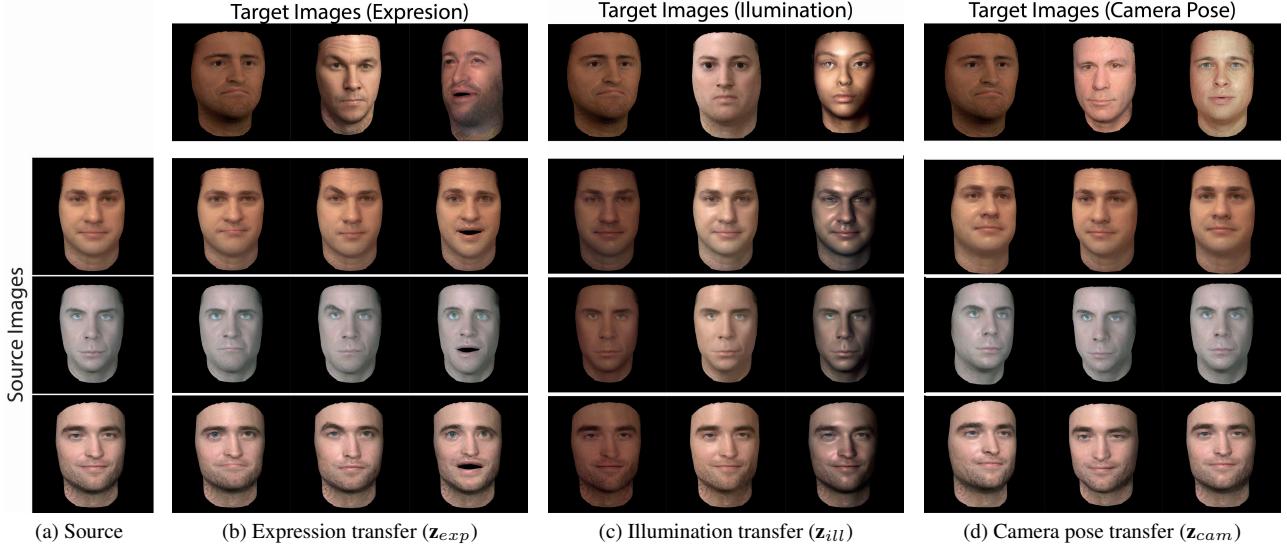


Figure 4. We showcase the ability of our network to disentangle facial and scene properties, by reconstructing the source images on the left, and replacing some of their attributes, with those acquired from target images.

of the input parameters, our synthesis network is capable of preserving the rest unchanged. Given the input images of Fig. 4a, the renderings in Fig. 4b demonstrate that 3DMM-RF achieves expression transfer from the desired target images to the input ones, while it preserves the same identities, lighting conditions and camera poses in each case.

4.2. Comparison with other models

4.2.1 Qualitative comparison

We compare the fitting ability of our method with MoFaNeRF [95] and HeadNeRF [33], both NeRF-based face models. MoFaNeRF is a parametric model which maps face images into a vector space including identity, expression and appearance features. HeadNeRF is another NeRF-based parametric head model achieving photorealistic results. Fig. 5 shows the fitting results for the same input images, which are presented in Fig. 5a. Even though MoFaNeRF appears to perform well (Fig. 5b), in some cases it completely misses to recreate a human face. Moreover, although HeadNeRF achieves to reproduce more photorealistic human faces than MoFaNeRF does (Fig. 5c), it doesn't always capture the right identity. It is clearly shown that, except rows a&e, the final rendered identity is not the same as the one appearing in the input image. As shown in Fig. 5d, our approach achieves both high-quality output while it preserves the right identity. Even for rows a&e where HeadNeRF manages to reconstruct the right identity, our approach produces finer details.

Moreover, we compare our implicit 3DMM modeling and rendering, with two state-of-the-art explicit 3DMM fit-

Method	L1 \downarrow	LPIPS \downarrow	SSIM \uparrow
MoFaNeRF [95]	0.37	0.098	0.9155
HeadNeRF [33]	0.241	0.05	0.9506
3DMM-RF (Ours)	0.216	0.035	0.9563

Table 1. Quantitative comparison between 3DMM-RF(Ours), MoFaNeRF [95] and HeadNeRF [33] using image comparison metric

ting methods, GANFIT [24] and AvatarMe++ [43]. Fig. 6 shows our results on two subjects, from a central and side pose, compared to the above methods. Even though above methods [24, 43] use explicit rendering based on a mesh, we showcase that our reconstruction is on-par with such methods, and our neural facial rotation accurately maintains the facial and environment characteristics.

4.2.2 Quantitative comparison

We measure the ability of our network of preserving the identity features by reproducing the steps followed in [23] and [26]. Firstly, we reconstruct every image of the Labeled Faces in the Wild Dataset (LFW) [34] using our method and then we feed them both to a face recognition network [58]. Then, we compare the cosine similarity distributions of the activations in the network's embedding layer between pairs of 1) the original and the reconstructed-rendered image in Fig. 7a 2) images containing the same/different identities in Fig. 7b. We show that the reconstructed images have more than 0.75 cosine similarity to the original images in average, and same/different pairs of LFW dataset are still distinguishable after our reconstruction. Both results show

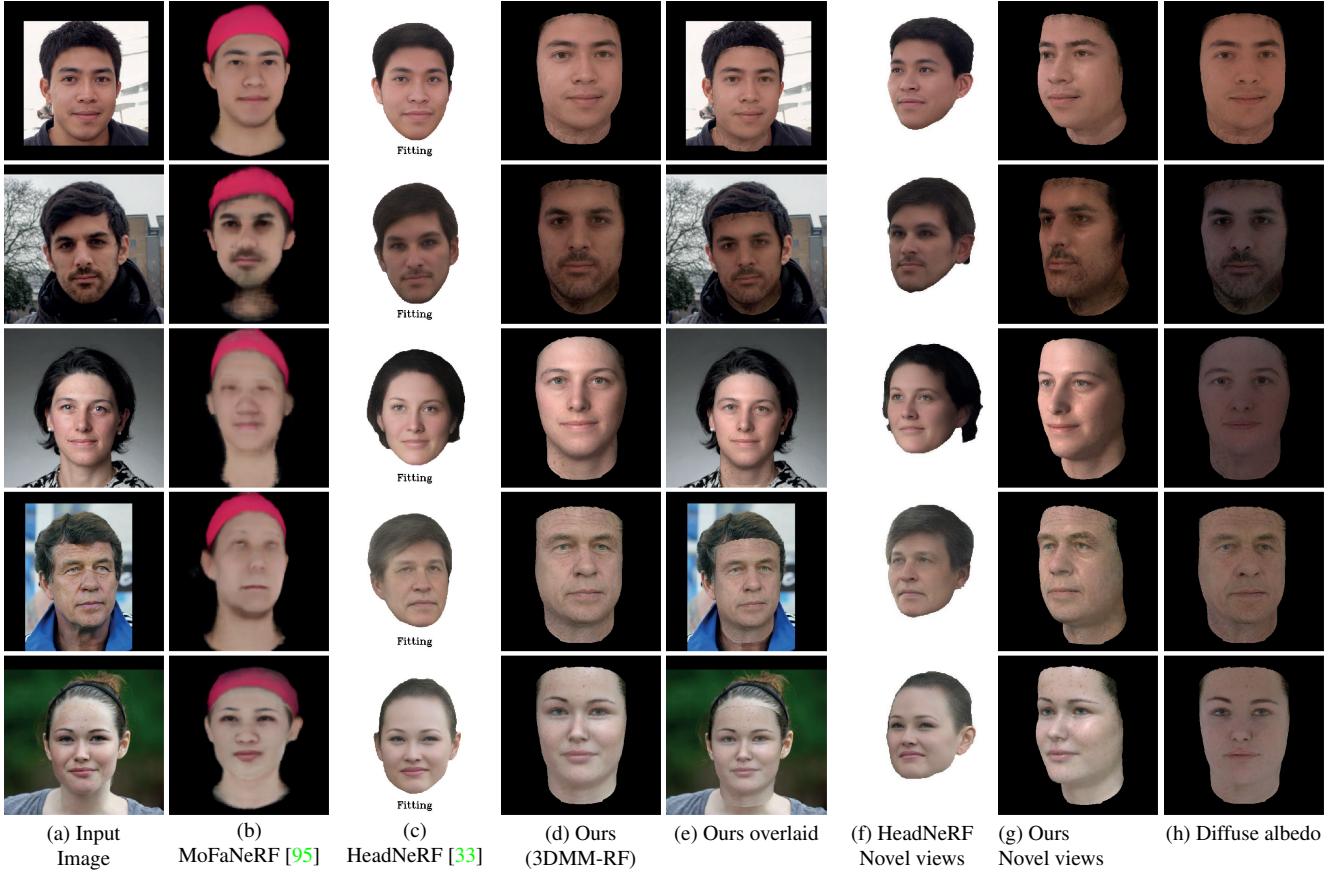


Figure 5. This figure contains a qualitative comparison between our method (3DMM-RF), MoFaNeRF [95] and HeadNeRF [33]. From left to right, the columns include the input image, MoFaNeRF’s fitted prediction, HeadNeRF’s prediction, 3DMM-RF’s prediction, a novel view rendered by HeadNeRF and 3DMM-RF and ours under diffuse albedo.

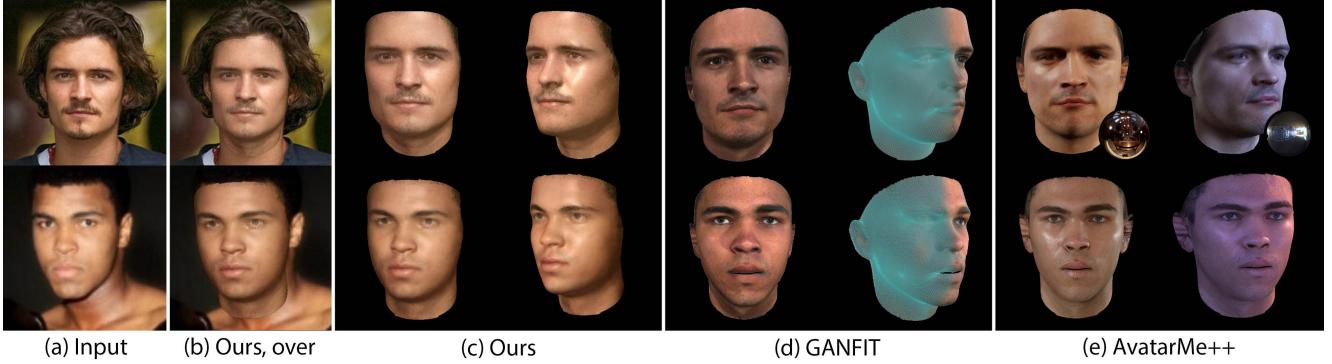


Figure 6. Comparison between our results and explicit 3DMM fitting methods, GANFIT [24] and AvatarMe++ [43]. We showcase that our reconstruction and rendering are on-par with these state-of-the-art methods, despite being implicit.

that identities are well-preserved. Our performance closely matches with the other state-of-the-art methods that are specializing in identity preservation such as [23, 25, 26, 80].

Additionally, we randomly select 550 images from the CelebAMask-HQ [44] following the method introduced by HeadNeRF [33]. We do fit our model on those images and

compare the fitted images with the state-of-the-art models MoFaNeRF [95] and HeadNeRF [33]. In contrast with the approach proposed by the authors of HeadNeRF, we do not train any of the networks using the rest images contained in this dataset. As comparison metrics, we use L_1 -distance, SSIM [84] and LPIPS [92]. The results are shown in Tab. 1

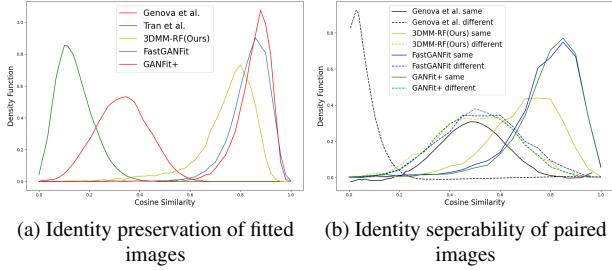


Figure 7. 7a: The distribution of cosine similarity between the original images and the reconstructed-rendered images of LFW dataset, which indicates strong identity preservation of the reconstruction of our model. 7b: The distributions between same (solid line) and different (dashed line) pairs of LFW dataset after they have been reconstructed and rendered by our method. The separability of the two distributions shows that the identity is being preserved.

which clearly shows that our method outperforms.

4.3. Ablation Study

One of the key parts of our approach is the use of SPADE [57] layers for feeding the scene parameters to the network. We examine the importance of these, by implementing another approach without any SPADE layers, in which all the parameters are given together as input to the network. Fitting the image of Fig. 8a to both networks, we present our network’s reconstructed image in Fig. 8b and a novel view of it in Fig. 8c, whilst Fig. 8d,8e include the fitted results performed by the network without SPADE layers. Comparing those figures, it is easily shown that the w/o SPADE layers network performs worse. We also notice that the resulting pose isn’t the desired one, meaning that SPADE layers are a key factor to the disentanglement of the scene parameters.

Another ablation study focuses on the fitting pipeline. We examine the importance of each loss function and the finetuning final step by a leave-one-out experiment. As shown in Fig. 8, the results in Fig. 8b are the best in comparison with the others, meaning that the current used loss functions, the SPADE layers and the finetuning step play a key role to the performance of the approach. Qualitatively, however, it’s very difficult to distinguish the importance of the L_{ID} loss. To measure it, we compare the cosine identity distance between the input image of Fig. 8a with the reconstructed images of Fig. 8b and Fig. 8f using the identity detector provided by deepFace [68]. We found that Fig. 8a is more similar to Fig. 8b than Fig. 8f, meaning that the L_{ID} loss plays an important role in the fitting pipeline.

4.4. Limitations and Future Work

Even though our network can achieve photo-realistic results, it still has some limitations. One of them is the fact that it cannot render other parts of the human head such

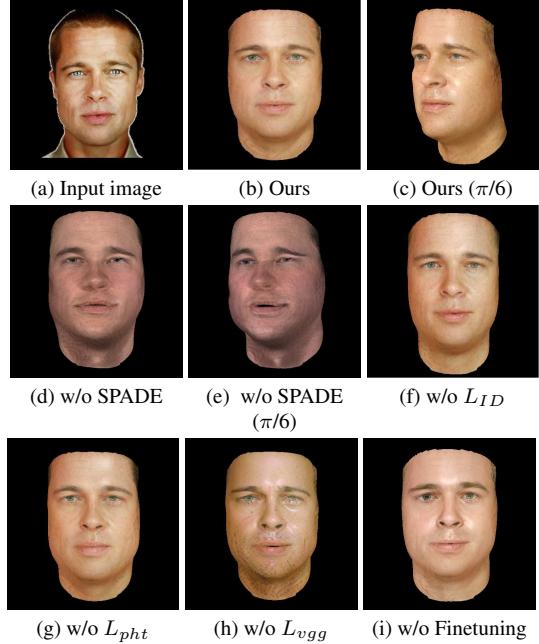


Figure 8. Leave-one-out ablation study to visualize the contribution of various components of our fitting pipeline.

as ears and hair. This is because the synthetic images we used for training don’t include any of these parts. On top of that, in some situations where some hair occludes parts of the forehead, our network gets baffled. In future work, we think of using methods that are capable of producing human faces including all the parts. Finally, the training data suffer from flattened eyes, as they are represented by a single facial mesh, which affects the final rendering.

Another limitation is the fact that our training images don’t include any wearable like glasses. As a result, in some cases, our network didn’t perform well in getting the right face texture because of the losses we are using during the fitting procedure. We consider finetuning our network in datasets including this type of data.

5. Conclusion

This work presents a facial deep 3D Morphable Model, which can accurately model a subject’s identity, pose and expression and render it under arbitrary illumination. This is achieved by utilizing a powerful deep style-based generator to bypass two main weaknesses of neural radiance fields, rigidity and rendering speed, by generating at one pass all the samples required for the volume rendering algorithm. We have shown that this model can accurately be fit to “in-the-wild” facial images of arbitrary pose and illumination conditions, extract the facial characteristics, and be used to re-render the face under controllable conditions.

References

- [1] Brian Amberg, Reinhard Knothe, and Thomas Vetter. Expression invariant 3D face recognition with a morphable model. In *2008 8th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2008*, pages 1–6. IEEE, 2008. [2](#)
- [2] Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. Modeling facial geometry using compositional vaes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3877–3886, 2018. [2](#)
- [3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH ’99*, 1999. [1, 2](#)
- [4] James F Blinn. Models of light reflection for computer synthesized pictures. In *Proceedings of the 4th annual conference on Computer graphics and interactive techniques*, pages 192–198, 1977. [4](#)
- [5] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3D face morphable models “In-the-Wild”. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-January, pages 5464–5473, 2017. [2](#)
- [6] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2):233–254, 2018. [4, 5](#)
- [7] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3D morphable model learnt from 10,000 faces. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-December, pages 5543–5552, 2016. [2](#)
- [8] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. Nerd: Neural reflectance decomposition from image collections. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12664–12674, 2021. [1, 2](#)
- [9] Sofien Bouaziz, Yangang Wang, and Mark Pauly. Online modeling for realtime facial animation. *ACM Transactions on Graphics*, 32(4):40, 2013. [2](#)
- [10] Martin Breidt, Heinrich H. Biilthoff, and Cristobal Curio. Robust semantic analysis by synthesis of 3D facial motion. In *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*, pages 713–719. IEEE, 2011. [2](#)
- [11] 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 *International Conference on Computer Vision*, 2017. [5](#)
- [12] Chen Cao, Yanlin Weng, Shun Zhou, Yiyi Tong, and Kun Zhou. FaceWarehouse: A 3D facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014. [2, 4](#)
- [13] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient Geometry-aware 3D Generative Adversarial Networks. page 28. [2](#)
- [14] 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 *arXiv*, 2021. [1, 2](#)
- [15] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5795–5805, 2021. [2](#)
- [16] Hang Dai, Nick Pears, William Smith, and Christian Duncan. A 3D morphable model of craniofacial shape and texture variation. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-October, pages 3104–3112, 2017. [2](#)
- [17] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. [5](#)
- [18] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:4685–4694, 2019. [5](#)
- [19] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8645–8654, 2021. [2](#)
- [20] Chen Gao, Yichang Shih, Wei-Sheng Lai, Chia-Kai Liang, and Jia-Bin Huang. Portrait neural radiance fields from a single image. *arXiv preprint arXiv:2012.05903*, 2020. [2](#)
- [21] Stephan J. Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14326–14335, 2021. [2](#)
- [22] Baris Gecer, Alexandros Lattas, Stylianos Ploumpis, Jiankang Deng, Athanasios Papaioannou, Stylianos Moschoglou, and Stefanos Zafeiriou. Synthesizing coupled 3d face modalities by trunk-branch generative adversarial networks. In *European conference on computer vision*, pages 415–433. Springer, 2020. [1, 2, 4, 5](#)
- [23] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. GANFIT: Generative Adversarial Network Fitting for High Fidelity 3D Face Reconstruction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1155–1164, Long Beach, CA, USA, June 2019. IEEE. [1, 2, 6, 7](#)
- [24] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. [6, 7](#)

- [25] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Fast-GANFIT: Generative Adversarial Network for High Fidelity 3D Face Reconstruction. *Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 1, 2, 7
- [26] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3D morphable model regression. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 8377–8386, 2018. 6, 7
- [27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 2
- [28] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 3, pages 2672–2680, 2014. 2, 4
- [29] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *International Conference on Learning Representations*, 2022. 2
- [30] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection, 2021. 5
- [31] 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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2
- [32] Peter Hedman, Pratul P. Srinivasan, Ben Mildenhall, Jonathan T. Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5855–5864, 2021. 2
- [33] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. 2022. 2, 6, 7
- [34] Gary B. Huang, Vedit Jain, and Erik Learned-Miller. Unsupervised joint alignment of complex images. In *ICCV*, 2007. 6
- [35] Zhang Jiakai, Liu Xinhang, Ye Xinyi, Zhao Fuqiang, Zhang Yanshun, Wu Minye, Zhang Yingliang, Xu Lan, and Yu Jingyi. Editable free-viewpoint video using a layered neural representation. In *ACM SIGGRAPH*, 2021. 2
- [36] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. 2
- [37] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018. 2
- [38] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020. 1, 2, 4
- [39] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018. 2
- [40] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 1, 2, 3, 4
- [41] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014. 2
- [42] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Trianafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. AvatarMe: Realistically Renderable 3D Facial Reconstruction "in-the-wild". In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 760–769, 2020. 2
- [43] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos P Zafeiriou. Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–1, 2021. 1, 4, 5, 6, 7
- [44] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7
- [45] Hao Li, Thibaut Weise, and Mark Pauly. Example-based facial rigging. *ACM SIGGRAPH 2010 Papers, SIGGRAPH 2010*, 29(4):32, 2010. 2
- [46] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, 36(6):194, 2017. 2
- [47] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft Rasterizer: A Differentiable Renderer for Image-Based 3D Reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2019. 5
- [48] Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. Editing conditional radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5753–5763, 2021. 2
- [49] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics*, 37(4):68, 2018. 2
- [50] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7206–7215, 2021. 2
- [51] 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 *European conference on computer vision*, pages 405–421. Springer, 2020. 1, 2, 3, 4
- [52] Stylianos Moschoglou, Stylianos Ploumpis, Mihalis Nicolaou, Athanasios Papaioannou, and Stefanos Zafeiriou. 3DFaceGAN: Adversarial nets for 3D face representation, generation, and translation. *arXiv preprint arXiv:1905.00307*, 2019. 1
- [53] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [54] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5569–5579, 2021. 2
- [55] 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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5845–5854, 2021. 2
- [56] 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. *ACM Trans. Graph.*, 40(6), dec 2021. 2
- [57] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 4, 8
- [58] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In Mark W. Jones Xianghua Xie and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015. 6
- [59] 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, Sept. 2009. 2
- [60] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021. 2
- [61] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 2
- [62] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10313–10322, 2021. 2
- [63] Mallikarjun B. R., Ayush Tewari, Tae-Hyun Oh, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, and Christian Theobalt. Monocular Reconstruction of Neural Face Reflectance Fields. *arXiv:2008.10247 [cs]*, Aug. 2020. 2
- [64] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3D faces using convolutional mesh autoencoders. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11207 LNCS:725–741, 2018. 2
- [65] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 4, 5
- [66] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14315–14325, 2021. 2
- [67] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: generative radiance fields for 3d-aware image synthesis. *CoRR*, abs/2007.02442, 2020. 2
- [68] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE, 2020. 8
- [69] Sahil Sharma and Vijay Chahar. 3d face reconstruction in deep learning era: A survey. *Archives of Computational Methods in Engineering*, 01 2022. 2
- [70] William A. P. Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua Tenenbaum, and Bernhard Egger. A Morphable Face Albedo Model. *arXiv:2004.02711 [cs]*, Apr. 2020. 2
- [71] Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7491–7500, 2021. 1, 2
- [72] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgarib, Hans-Peter Seidel, Patrick Perez, Michael Zollhofer, and Christian Theobalt. FML: Face Model Learning From Videos. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10804–10814, Long Beach, CA, USA, June 2019. IEEE. 1, 2
- [73] Ayush Tewari, Mohamed Elgarib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020. 1
- [74] Ayush Tewari, Michael Zollhofer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Perez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2549–2559, 2018. 2
- [75] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Pérez, and Christian

- Theobalt. MoFA: Model-Based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, volume 2018-January, pages 1274–1283, 2017. 2
- [76] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Transactions on Graphics*, 34(6):181–183, 2015. 2
- [77] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-January, pages 1493–1502, 2017. 2
- [78] Luan Tran and Xiaoming Liu. On learning 3D face morphable model from in-the-wild images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. 1
- [79] Luan Tran and Xiaoming Liu. On Learning 3D Face Morphable Model from In-the-wild Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. 2
- [80] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-January, pages 1283–1292, 2017. 7
- [81] Arash Vahdat and Jan Kautz. NVAE: A Deep Hierarchical Variational Autoencoder. *arXiv:2007.03898 [cs, stat]*, July 2020. 2
- [82] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 34:27171–27183, 2021. 2
- [83] Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. Learning compositional radiance fields of dynamic human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5704–5713, June 2021. 2
- [84] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 7
- [85] Shih En Wei, Jason Saragih, Tomas Simon, Adam W. Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. VR facial animation via multiview image translation. *ACM Transactions on Graphics*, 38(4):67, 2019. 2
- [86] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. *CVPR*, 2022. 2
- [87] Fei Yang, Dimitri Metaxas, Jue Wang, Eli Shechtman, and Lubomir Bourdev. Expression flow for 3D-Aware face componant transfer. *ACM Transactions on Graphics*, 30(4):1–10, 2011. 2
- [88] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 2
- [89] T Yenamandra, A Tewari, F Bernard, HP Seidel, M Elgarib, D Cremers, and C Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE / CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2
- [90] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5732–5741, 2021. 2
- [91] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv:2010.07492*, 2020. 2
- [92] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5, 7
- [93] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. CIPS-3D: A 3D-Aware Generator of GANs Based on Conditionally-Independent Pixel Synthesis. 2021. 2
- [94] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788*, 2021. 2
- [95] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. *arXiv preprint arXiv:2112.02308*, 2021. 2, 6, 7
- [96] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt. State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications. *Computer Graphics Forum*, 37(2):523–550, May 2018. 2