MoRF: Morphable Radiance Fields for Multiview Neural Head Modeling

Daoye Wang
Prashanth Chandran
daowang@ethz.ch
prashanth.chandran@disneyresearch.com
ETH Zurich
Zurich, Switzerland
DisneyResearch|Studios
Zurich, Switzerland

Gaspard Zoss
Derek Bradley
Paulo Gotardo
gaspard.zoss@disneyresearch.com
derek.bradley@disneyresearch.com
paulo.gotardo@disneyresearch.com
DisneyResearch|Studios
Zurich, Switzerland

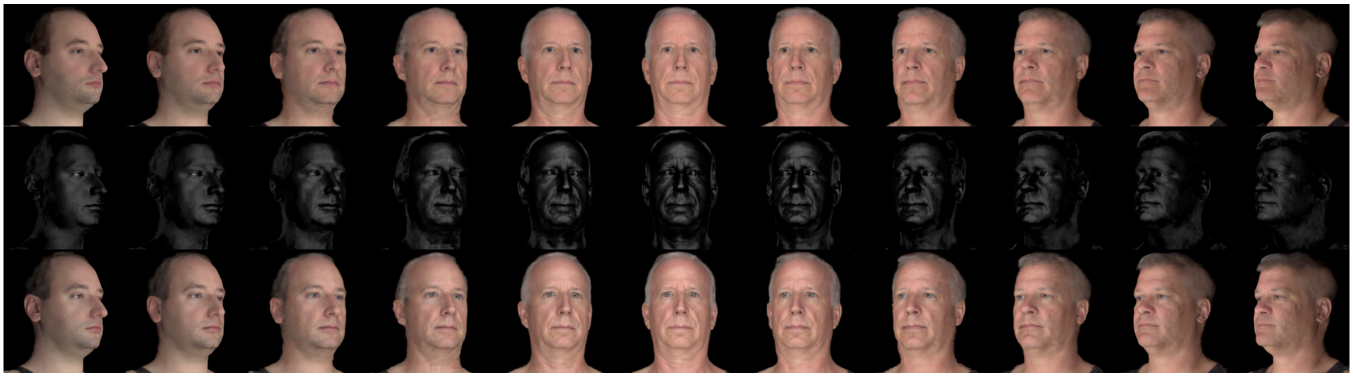

Figure 1: MoRF leverages a high-quality image database with polarization-based separation of diffuse and specular reflection to learn a generative model that can synthesize novel human head volumes for photorealistic, multiview-consistent rendering. Left to right: the rendering viewpoint changes while morphing between three different subjects (left, center, right). Top to bottom: the modeled diffuse, specular, and full RGB color.

ABSTRACT

Recent research work has developed powerful generative models (*e.g.*, StyleGAN2) that can synthesize complete human head images with impressive photorealism, enabling applications such as photorealistically editing real photographs. While these models can be trained on large collections of unposed images, their lack of explicit 3D knowledge makes it difficult to achieve even basic control over 3D viewpoint without unintentionally altering identity. On the other hand, recent Neural Radiance Field (NeRF) methods have already achieved multiview-consistent, photorealistic renderings but they are so far limited to a single facial identity. In this paper, we propose a new Morphable Radiance Field (MoRF) method that extends a NeRF into a generative neural model that can realistically synthesize multiview-consistent images of complete human heads, with variable and controllable identity. MoRF allows for morphing between particular identities, synthesizing arbitrary new identities, or quickly generating a NeRF from few images of a new subject, all while providing realistic and consistent rendering under novel viewpoints. We train MoRF in a supervised fashion by leveraging a high-quality database of multiview portrait images of several people, captured in studio with polarization-based separation of diffuse and specular reflection. Here, we demonstrate how MoRF is a strong new step forwards towards generative NeRFs for 3D neural head modeling.

CCS CONCEPTS

• **Computing methodologies** → **Rendering**; **Neural networks**; **Volumetric models**.

KEYWORDS

neural radiance fields, novel view synthesis, generative models, neural rendering, photoreal human synthesis.

ACM Reference Format:
Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Gotardo. 2022. MoRF: Morphable Radiance Fields for Multiview Neural Head Modeling. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings (SIGGRAPH '22 Conference Proceedings), August 7–11, 2022, Vancouver, BC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3528233.3530753

SIGGRAPH '22 Conference Proceedings, August 7–11, 2022, Vancouver, BC, Canada
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9337-9/22/08…$15.00
https://doi.org/10.1145/3528233.3530753

1 INTRODUCTION

For over two decades, parametric 3D face models have been employed in computer vision and graphics applications. These models are typically built from datasets of 3D facial scans and characterize the space of facial shapes, often through linear blends of the input data. Face models offer compact representations that can be used for compression, to provide deformation priors for applications like image-based reconstruction, and they can even be sampled to generate synthetic new faces. A main drawback of traditional face models, however, is that they only represent the 3D shape (and sometimes rudimentary appearance) of facial skin areas only; they cannot represent the more complex components like facial hair, scalp hair, eyes and so on, including the complex reflectance properties of all face components. Thus, current face models are unsuited for generating photorealistic full-head renditions of people.

Data-driven photorealistic face modeling has been a topic of recent research, which has led to very powerful generative models like the StyleGAN variants [Karras et al. 2019]. These deep neural networks can generate full head portraits with a fidelity that matches natural images. While it is possible to separate some semantic components of these models for edits [Abdal et al. 2020; Härkönen et al. 2020; Shen et al. 2020], the results are typically limited to nearly frontal portraits and lack precise consistency in 3D geometry and appearance when rendering from multiple different viewpoints [Chandran et al. 2021].

To achieve multiview-consistent face modeling, an attractive recent development is the Neural Radiance Field (NeRF). NeRFs represent 3D scenes using fully connected neural networks that predict the density and radiance at any given point in a continuous volume, for any view direction. Once trained on images of a scene, they allow for synthesizing photorealistic novel views with accurate view-consistency due to their volumetric nature. Traditional NeRFs are designed to represent only a single scene, but have been shown to extend also to dynamic scenes. In the context of faces, NeRFs can represent high-quality, complete 3D heads and even render novel photoreal views of a talking performance. What is currently lacking is a morphable radiance field model with the power to synthesize arbitrary full-head volumes for multiview-consistent rendering.

In this work we aim to fill this gap by presenting *MoRF*, a framework for morphable radiance fields, that allows multiview consistent photorealistic neural head modeling. MoRF is trained in a supervised way from a dataset of 3D head scans and corresponding sparse multiview images that were used for 3D reconstruction. We extend the NeRF framework by training a single volumetric model that can represent all identities in the dataset, where each identity is represented by a latent code. Following recent work on dynamic NeRFs, we learn subject-specific deformation fields that semantically align the identities to a common canonical NeRF. We show that high-quality results can be achieved by partially supervising the deformation field on skin areas, for which we have correspondences across identities, and automatically learn how to best deform non-skin regions like hair, where correspondences are not available. We further leverage polarization-based capture to separate the output radiance into diffuse and specular reflection; the view-dependent specular component is then constrained using low-frequency spherical harmonics. As a result, MoRF enables new applications such as morphing between full-head models, synthesizing new identities, and generating a NeRF from just one image of new test subjects, all while offering photorealistic novel view rendering as in Fig. 1. We believe MoRF is a strong step forward towards effective generative NeRFs for high-quality full-head modeling beyond facial skin.

2 RELATED WORK

This section reviews the closely related work on traditional methods for face modeling and rendering, before discussing the recent explosion of neural rendering methods, with a focus on NeRFs.

Linear 3D Morphable Models (3DMMs) were originally proposed by Blanz and Vetter [Blanz and Vetter 1999; Egger et al. 2020] to model 3D meshes and textures of faces and were later improved into multi-linear models [Vlasic et al. 2005] with disentangled control geometry (identity, expression), and appearance. More recently, non-linear face models have also been proposed as deep neural networks with fully-connected or mesh-convolutional layers [Abrevaya et al. 2019; Chandran et al. 2020; Gong et al. 2019; Ranjan et al. 2018]. These and other recent mesh-based models [Li et al. 2017] typically focus on facial skin areas and fail to reproduce the complex geometry and appearance of all the different components of a complete human head, such as hair, eyes, and inner mouth. The automated construction of complete 3D morphable head models has been investigated in [Ploumpis et al. 2020], but without hair. Instead of a mesh, Yenamandra et al. [2021] fit an implicit surface model to full-head 3D scans to represent geometry and per-point colors, without modeling view-dependent colors due to specular reflectance. Their model generates smooth surfaces (particularly on hair areas) and lacks a method for rendering photoreal images. Ramon et al. [2021] learn a model of full-head geometry only, for 3D reconstruction from few views, without a generative appearance model. Thus, while techniques do exist for modeling all head components, generating and rendering photoreal digital humans still requires a lot of manual work by skilled artists.

The recent advent of neural rendering is changing this scenario very quickly, as these techniques now generate images of complete human heads (and full bodies) with impressive photorealism. StyleGAN and its derivatives [Karras et al. 2021, 2019, 2020] are popular and powerful full-head models that generate synthetic images with a large variety of facial identities and photorealistic appearance. These 2D image models are trained in an unsupervised way over very large face image datasets, without requiring posed images and proxy 3D geometry. However, the lack of an explicit 3D model makes it challenging to have basic control of 3D viewpoint without unintentionally altering the identity and appearance of the modeled object. Instead of fully operating on the 2D image grid, more recent generators [Gu et al. 2022; Nguyen-Phuoc et al. 2019; Niemeyer and Geiger 2021] model an intermediate 3D volume of latent features and explicitly apply a rigid transformation to it, before projecting features onto the 2D image plane. But 2D kernels are still used for upsampling, to manage rendering speed. This combined 3D-2D neural rendering only mitigates multiview inconsistency without removing it; 3D pose and appearance are not fully disentangled.

Another class of volumetric neural models is already capable of novel view synthesis with impressive photorealism and 3D consistency. So far, such models are typically limited to the geometry and appearance of the single scene on which it was trained on (*e.g.*, a human head with a particular identity [Lombardi et al. 2019; Ma et al. 2021]). Here, we focus on the more closely related methods based on Neural Radiance Fields (NeRFs) [Mildenhall et al. 2020], which model continuous volumes implicitly, using multilayer perceptron (MLP) networks, and whose resolution is not limited by explicit volume discretization. NeRFs are trained to learn properties of 3D points within a semi-translucent neural volume that is rendered via traditional ray-marching techniques.

In particular, we focus on NeRF models that represent properties of a non-static volume, to model human heads. Most of these models represent a dynamic scene whose geometry and appearance change slightly over time [Athar et al. 2021; Gafni et al. 2021; Li et al. 2021; Martin-Brualla et al. 2021; Park et al. 2021a,b; Pumarola et al. 2020; Tretschk et al. 2021; Wang et al. 2020; Xian et al. 2021; Xie et al. 2021]. These works mostly address the challenging scenario of learning deformable head (or body) models from monocular video; they have so far been applied mainly to scenes with small shape deformations or small appearance changes. In contrast, our work is not restricted to monocular capture and builds a generative model that encodes larger variability in both geometry and appearance across complete human heads of different identities (*e.g.*, large changes in hair styles). Gao et al. [2020] use metalearning on multiple subjects to derive a prior for fine tuning a subject-specific NeRF from a single (frontal) portrait image. Raj et al. [2021] propose a multi-identity NeRF conditioned on pixel-aligned local features encoded from input images, without inducing a latent space for identity. These NeRFs cannot be easily sampled to generate novel, fully synthetic subjects.

Other recent work has already attempted, with some success, to learn NeRFs that can synthesize multiple object identities. GRAF [Schwarz et al. 2020] learns a generative model for a 3D radiance field that is conditioned on shape and appearance latent codes. And π-GAN [Chan et al. 2021] also learns a multi-identity NeRF model using network layers that are modulated by a noise vector, in StyleGAN fashion [Karras et al. 2019]. CIPS-3D [Zhou et al. 2021] uses a similar NeRF architecture, but it is shallow for 3D geometry and includes an additional deep 2D network for appearance, which unfortunately adds multi-view inconsistencies (as with [Gu et al. 2022; Niemeyer and Geiger 2021]). Training these NeRF-based GANs requires large datasets of unposed images. They can be difficult to train and generalize well since they are all conditioned on latent codes (global or localized [DeVries et al. 2021; Wang et al. 2020]) without explicit correspondence of facial parts across identities. In contrast, as noted by [Park et al. 2021b], dynamic NeRFs that model head deformation over time incorporate a 3D warping field that explicitly brings the data into semantic correspondence within a canonical 3D space, where each position of the canonical radiance field is better constrained when training.

This paper describes our novel generative NeRF model that improves on the warp field component and better constrains the canonical NeRF network for modeling heads of various identities (geometry and appearance). Differently from the NeRF GANs above, we do not employ adversarial training nor large datasets of unposed images with mostly frontal heads. Instead, we rely on a multiview image database of head portraits (captured in studio with polarization-based separation of specular and diffuse reflection) and show how a high-quality morphable head model can be trained in a simple, fully supervised way. Our comprehensive set of constraints (see Section 3.2) allows us to train our high-quality generative model that truly disentangles 3D pose and identity.

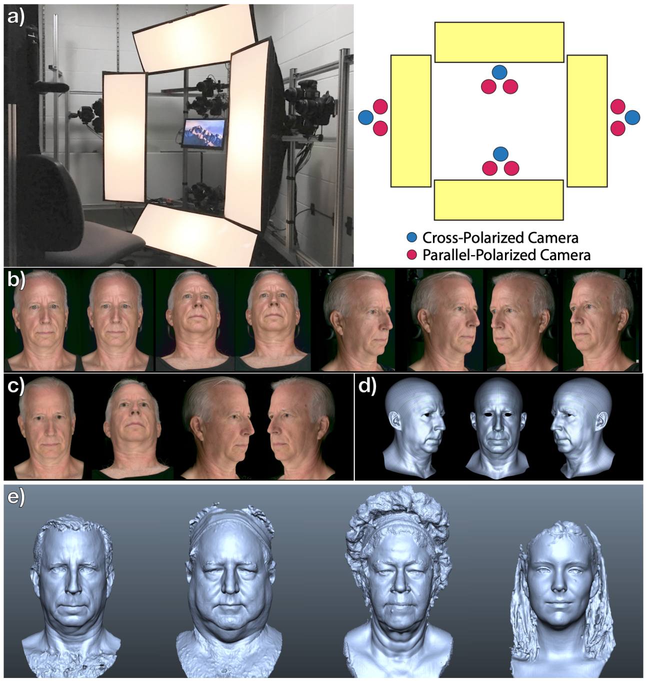

Figure 2: MoRF is trained in a supervised way using a high-quality multiview image database with full head portraits, captured in studio with 12 cameras and polarized light: 4 camera triplets (a), each including a parallel-polarized stereo pair (b) and a cross-polarized camera (c) that only captures diffuse reflection; (d) a canonical mesh is pre-aligned to the skin areas of all 3D head scans and used to supervise the deformation field in MoRF; (e) full-head multiview stereo is also used to supervise the densities of MoRF.

3 MORF: MORPHABLE RADIANCE FIELDS

We now present our solution for learning 3D neural models that can generate photorealistic, multiview consistent images of complete human heads, with variable and controllable identity. To achieve this goal, we leverage the recent advances brought by research on NeRFs, which we extend to a new generative model of neural volumes dubbed Morphable Radiance Fields (MoRF).

3.1 Modeling Scenario

Instead of adopting adversarial training from a large collection of unposed images, we train MoRF in a simpler, fully supervised fashion that leverages a high-quality, multiview image database with full head portraits of several people captured in studio conditions. This allows us to go beyond modeling mostly frontal faces, and capture a better variety of viewpoints for each subject. In addition, polarization-based capture allows us to explicitly and separately model view-invariant diffuse reflection (color) and view-dependent specular reflection. Fig. 2 shows our capture setup with 12 DSLR

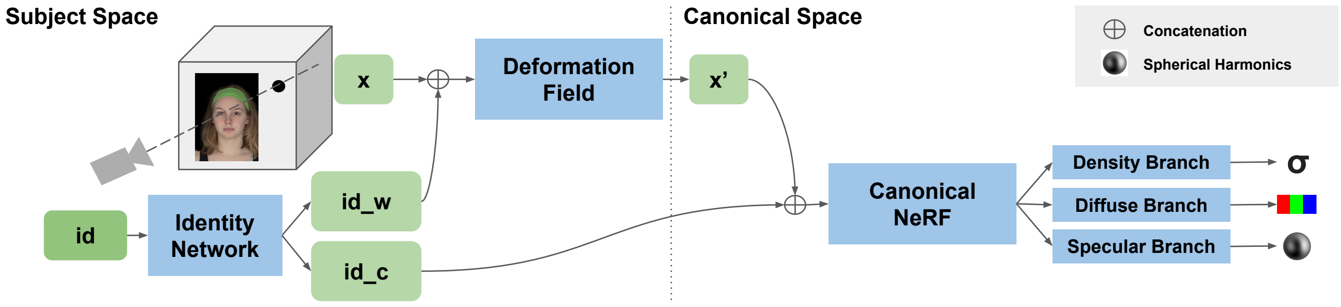

Figure 3: The architecture of MoRF includes three main MLPs: (*i*) the *Identity Network* first translates a subject-specific id code into a geometric deformation ($\mathrm{id_w}$) and canonical identity ($\mathrm{id_c}$) codes, used to condition the other two MLPs; then (*ii*) the *Deformation Field* MLP warps the world space into a canonical one, where (*iii*) the *Canonical NeRF* models a semantically-aligned radiance field comprising density, view-invariant diffuse color, and omnidirectional specular color using spherical harmonics.

cameras organized into 4 triplets; each triplet includes a narrow-baseline stereo camera pair that is parallel polarized with respect to the lighting, plus an additional cross-polarized camera that only captures diffuse reflection. Lighting is constant and mostly frontal, coming from four flash boxes with horizontal polarization filters. Lighting (environment map) and camera matrices are fully calibrated. The data includes full-head 3D reconstructions from multi-view stereo, with a canonical mesh topology (Fig. 2 (d)) registered to each 3D scan to semantically align skin areas across subject identities. Rigid head pose normalization is also applied across all subjects. Appearance maps (albedo, specular intensity and roughness) are also available for generating traditional skin renders from arbitrary views. More details are described in [Riviere et al. 2020].

3.2 Generative NeRF Model

To generate multiview-consistent images of its morphable neural volume, MoRF extends the rendering approach used by NeRF and its variants. The network architecture of MoRF includes three main MLP components: (1) an *Identity Network*, (2) a *Deformation Field*, and (3) a novel *Canonical NeRF* model (Fig. 3). These MLPs are summarized next and detailed in the subsections below.

The identity network takes as input a subject-specific latent **id** code and maps it to a pair of new codes,

$$[\ \mathbf{id}_\mathrm{w}, \mathbf{id}_\mathrm{c}\ ] = \mathrm{ID}(\mathbf{id}), \tag{1}$$

that separately encode pure geometric deformation and canonical appearance within a shape-normalized space. Then, the resulting code $\mathbf{id}_\mathrm{w}$ conditions the non-rigid deformation field MLP that warps the 3D world space $\mathbf{x} = (x, y, z)$ into a canonical 3D space,

$$\mathbf{x}' = \mathrm{DF}(\mathbf{x}; \mathbf{id}_\mathrm{w}). \tag{2}$$

Thus, we explicitly align (spatially and semantically) corresponding head regions within the canonical space to facilitate modeling of the observed variability across different subjects. Appearance in canonical space is modeled by the third MLP, the Canonical NeRF,

$$[\ \sigma, \mathbf{c}_d, \mathbf{c}_s\ ] = \mathrm{CN}(\mathbf{x}'; \mathbf{id}_\mathrm{c}). \tag{3}$$

This MLP has three output branches, instead of the usual two, since we leverage polarization-based image capture to explicitly model diffuse and specular radiances separately. Thus, besides the per-point density σ and viewpoint-invariant diffuse RGB color $\mathbf{c}_d$, this network also outputs a separate specular code $\mathbf{c}_s \in \mathbb{R}^K$, representing a local distribution of specular radiance over the full 3D sphere of allowed viewing (ray) directions $\mathbf{r}$. We model this distribution within a low-frequency spherical harmonics (SH) subspace of order K, a natural design choice also used in Yu et al. [2021]. Here, with $K = 3$, this constraint allows us to enforce *smoothness across viewpoints* and train with as few as K^2 viewpoints (see supplementary file). Also, this MLP becomes an *omnidirectional* NeRF that can be evaluated once at any point $\mathbf{x}'$ to provide a single output for all possible viewing directions $\mathbf{r}$. The full color observed from a particular $\mathbf{r}$ is $c(\mathbf{r}) = \mathbf{c}_d + \mathbf{c}_s^T \mathrm{SH}(\mathbf{r})$, where $\mathrm{SH}(\mathbf{r})$ is the SH basis along $\mathbf{r}$. Since our (color calibrated) images are captured under white light, we add a (scalar) neutral specular color to the diffuse RGB color $\mathbf{c}_d$.

3.2.1 ID Network Constraints. The design of the ID network defines our latent space model, Eq. 1, which is an essential part of MoRF. This network takes as input a single **id** $\in \mathbb{R}^D$ code and learns to predict codes $\mathbf{id}_\mathrm{w}$ and $\mathbf{id}_\mathrm{c}$ as to best condition the other two MLPs. An **id** code is optimized per subject during training. Although we could directly optimize for per-subject codes $\mathbf{id}_\mathrm{w}$ and $\mathbf{id}_\mathrm{c}$, this approach becomes poorly constrained as the number of training subjects and the dimensionality of these codes increase. Instead, by learning to predict $\mathbf{id}_\mathrm{w}$ and $\mathbf{id}_\mathrm{c}$, we obtain smoother and highly disentangled embeddings for these codes (see Section 4). In our experiments, we trained with $S = 15$ subjects and a small $D = 4$. The compact **id** codes are initialized from Gaussian white noise, and optimized with the weights of the 3 MLPs, with an ID loss

$$\mathcal{L}_{ID} = \lambda_{ID} \sum_s \left\|\mathbf{id}^s\right\|_2^2. \tag{4}$$

The weight λ_{ID} controls the **id** space regularization near the origin.

3.2.2 Deformation Field Constraints. Since there can be large variations in head shape across different identities, MoRF adopts the common approach of shape alignment for appearance modeling [Blanz and Vetter 1999; Cootes et al. 2001]. MoRF learns a 3D deformation field that maps all heads onto a unique canonical space that better constrains the training of each position of the multi-subject canonical NeRF. Effectively, the observed geometric variability is modeled as smooth deformations of a canonical 3D shape [Newcombe et al. 2015]. The design of our deformation field, Eq. 2, leverages these lessons and also recent adaptations of these ideas for use with NeRFs. Our deformation field MLP is largely similar to the one proposed by [Park et al. 2021a], and outputs a spatially-varying 6-dimensional vector that encodes a 3D rotation $\mathbf{R_x}$ and 3D translation $\mathbf{t_x}$. We then obtain $\mathbf{x}' = \mathbf{R_x}(\mathbf{x} + \mathbf{t_x})$. Empirically, we obtained better deformation fields without the elastic energy regularizer of Part et al. [2021a]. Instead, we control the smoothness of the warp based on the number of frequency bands used to position-encode the input $\mathbf{x}$. Additionally, a major benefit of our design is that we can partially supervise the deformation field for each identity by leveraging our dataset with dense semantic alignment of the skin surface across all subjects, as provided by the registered common 3D mesh topology shown in Fig. 2 (c). We compute the average 3D mesh across all our training subjects and use it to define the canonical surface for facial skin. We consider nearly 600K vertices $\mathbf{v}$ on our subdivided meshes. Then, our deformation loss is

$$\mathcal{L}_{DF} = \lambda_{DF} \sum_{s,\mathbf{v}} \left\| (\mathbf{v}' + \xi \mathbf{n}') - \mathrm{DF}(\mathbf{v} + \xi \mathbf{n}; \mathbf{id}^s_{\mathrm{w}}) \right\|_2^2, \qquad (5)$$

computed in the canonical space. Above, λ_{DF} is a weight and $\mathbf{v}$ is a vertex (with normal $\mathbf{n}$) on the s-th subject's mesh that corresponds to vertex $\mathbf{v}'$ (with normal $\mathbf{n}'$) in our canonical 3D mesh; $\xi \sim \mathcal{N}(0, \sigma)$ is a normal random variable with small standard deviation $\sigma = 0.05$ mm. Thus, we supervise the warp field within a thin volume near the surface of the mesh (face and neck), where reliable correspondences across subjects are available. In other areas, semantic correspondence is lacking due to the large variability in both geometry and appearance due to different hair styles and accessories. In those areas, the model learns appropriate correspondences as guided by the above smoothness constraint (on the whole volume) and the other losses. We also allow MoRF to model large changes in hair style using the canonical NeRF's $\mathbf{id}_{\mathrm{c}}$ code, as described next.

3.2.3 Canonical NeRF Constraints. Our multi-subject NeRF consists of a main MLP trunk and 3 output MLP branches, Fig. 3. The input to the trunk, $[\ \mathbf{x}', \mathbf{id}_{\mathrm{c}}\ ]$, is also supplied to its mid layer via skip links. Thus, the task of the initial layers can be understood as transforming the input into a localized canonical identity code, with the back-end then conditioned by both global and local codes. The MLP architectures are detailed in the supplementary file.

Although the deformation field above helps constrain the canonical NeRF, the warp supervision lacks semantic correspondences for the complex variation in hair styles and optional accessories. Thus, the canonical NeRF is also required to generate new density in some large areas that would otherwise encode empty space (*e.g.*, long hair). This fact motivates conditioning all of the canonical NeRF outputs on $\mathbf{id}_{\mathrm{c}}$, rather than conditioning only the color outputs as in previous work (see ablation in supplementary file). In a way, $\mathbf{id}_{\mathrm{c}}$ can be understood as playing a similar role as the hyper-coordinates used in [Park et al. 2021b] to model topology changes due to changing facial expressions. Here, we use a similar mechanism to model variation in hair styles. We find that, in some underconstrained cases, the canonical NeRF can learn to rely strongly on $\mathbf{id}_{\mathrm{c}}$, leading to an undesirable attenuation of the deformation field. Our deformation supervision via Eq. 5 greatly helps us mitigate this problem.

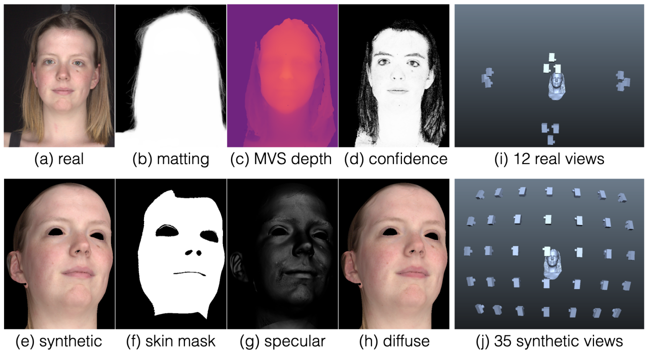

Figure 4: MoRF is supervised using real (a) and synthetic (e) images, with rendering losses on foreground pixels (b),(f). Density losses are applied on the background (b) and in front of the surface using multiview stereo (MVS) depth (c) and confidence maps (d). The 12 real views (i) are augmented with 35 synthetic views (j) for training.

To supervise MoRF's output density σ, we define a new loss that promotes sparsity along the ray cast to render each pixel p. This loss comprises two terms that, for each camera view c, make use of: (*i*) a set of foreground (head) pixels in a *matting mask* $\mathcal{M}_c$, Fig. 4 (b); and (*ii*) a multiview stereo (MVS) *depth map* $\mathcal{Z}_c$, Fig. 4 (c). Let $\mathbf{x}_{pi}$ denote the i-th 3D point sampled along the ray of p, and σ_{pi} be the density and z_{pi} the depth at $\mathbf{x}_{pi}$. We define the density loss as

$$\mathcal{L}_{\sigma} = \sum_{s,c,p} \left(\lambda_m \sum_i |\sigma_{pi}| \mathbf{1}_{[p \notin \mathcal{M}_c]} + \lambda_z \lambda_{cp} \sum_i |\sigma_{pi}| \mathbf{1}_{[z_{pi} < \mathcal{Z}_{cp} - \delta_z]} \right), \qquad (6)$$

where the indicator $\mathbf{1}_{[\chi]}$ is 0 when the condition χ is false. While the first term discourages non-zero density outside the matting mask, the second term encourages zero density at points in front of the surface, but not closer than $\delta_z = 5$ mm ("line-of-sight" prior). The weight λ_z of the depth term is modulated by the per-point confidence derived from the MVS reconstruction, which typically depicts skin areas with high confidence $\lambda_{cp} \approx 1.0$, Fig. 4 (d). The NeRF is otherwise free to improve full-head geometry during training.

The per-point diffuse and specular colors $\mathbf{c}_d$ and $\mathbf{c}_s$ output by MoRF (Eq. 3) are supervised using training images $\mathcal{I}_c$ from parallel- and cross-polarized cameras, Fig. 2 (b)-(c). The rendering loss is:

$$\mathcal{L}_{RGB} = \sum_{s,c,p} \left\| \mathcal{I}_{cp} - \sum_i \omega_{pi} \left(\mathbf{c}_{dpi} + \mathbf{c}_{spi}^T \mathrm{SH}(\mathbf{r}) \mathbf{1}_{[c \notin C_{\mathcal{XP}}]} \right) \right\|_2^2, \qquad (7)$$

where the weights ω_{pi} are defined by the σ_{pi}, for integrating along the ray [Mildenhall et al. 2020], and $C_{\mathcal{XP}}$ is the set of cross-polarized cameras. Besides the 12 real images for each subject, we augment our training data with synthetic images of the face that are rendered in high quality at new viewpoints using traditional ray-tracing and the captured appearance parameters [Riviere et al. 2020], Fig. 4 (e)-(h). In total, we add 35 synthetic viewpoints per subject, Fig. 4 (j). These synthetic images contain only skin pixels and provide no matting nor RGB supervision for non-skin areas (e.g. hair, eyes). Nevertheless, we found that this type of augmentation helps guide

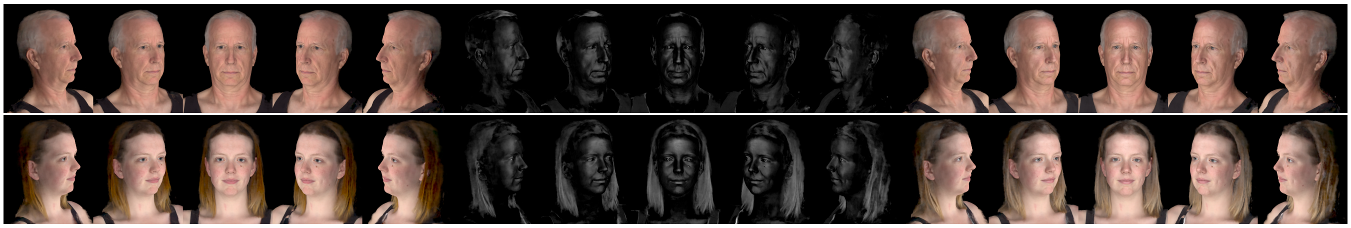

Figure 5: Example of synthesized "turntable" views of a single MoRF model evaluated for two (out of 15) training subjects: rendered diffuse color (left), specular color (mid), and full color (right); see other views and subjects in supplementary video.

and improve convergence during training, while also constraining facial silhouettes and the per-point specular output of MoRF.

3.2.4 Training Details. We train MoRF by optimizing the MLP weights and per-subject **id** codes as to minimize the sum of the losses above, with $\lambda_{ID} = 0.1$, $\lambda_{DF} = 10$, $\lambda_m = 0.1$, and $\lambda_z = 0.5$. Following established practice, we position-encode the input points $\mathbf{x}$ and $\mathbf{x}'$ of our MLPs using 8 frequency bands, with a coarse-to-fine training scheme [Park et al. 2021a] that masks higher frequency bands (it reaches full-band capacity after 50% of training iterations). In early training stages, we apply only the rendering loss $\mathcal{L}_{RGB}$ to let the canonical NeRF more quickly learn sufficient head geometry and appearance. We initially set a large $\delta_z = 100$ mm in the depth loss to constrain only points far from the surface, then gradually decrease this distance. All loss terms are enabled after 20% of training epochs. For the results presented next, we trained MoRF with an Adam optimizer for 400K iterations, which takes about 80 hours (15 subjects) on a single NVidia GTX3090 GPU.

4 RESULTS

We now show results of MoRF in different tasks, including image reconstruction, novel view synthesis, the generation of novel synthetic heads by mixing codes between pairs of training subjects, interpolating novel identities within the learned latent subspaces, and fitting a pre-trained MoRF to images of novel subjects left out of training. Unless stated otherwise, all results were obtained from a single MoRF model trained on a 15-subject dataset that includes images as illustrated in Fig. 2 and Fig. 4. Additional experiments in the supplementary file show: a comparison against stereo depth; the contribution of spherical harmonics constraints; the removal of warp supervision; the PSNR of novel view generation; PSNR of MoRF vs single-identity NeRF; a comparisons to variants of MoRF that closely resemble other baseline methods in the literature; and proof-of-concept results on modeling facial expressions with MoRF.

Modeling Quality. We trained MoRF on a total of 180 real images and 1050 synthetic images from across the 15 different subjects. Examples of novel "turntable" views synthesized by MoRF for two training subjects are shown in Fig. 5. The figure shows separate renderings of diffuse and specular colors, as well as the final full color rendering, to better illustrate the contribution of separating these components explicitly in the model (see other views and subjects in the supplementary material). As with NeRFs, images rendered with MoRF show high photorealism and multiview consistency and closely match the full head identity depicted in the real images used for training (Fig. 2). While geometry and diffuse appearance remain consistent, view-dependent specular highlights realistically change across the rendered viewpoints. The resulting renders even capture the reflection of the light sources on the subject's eyes. As expected, modeling error is predominantly found on the hair areas due to the complex geometry and also the small number of real training images used to supervise hair areas. After training, the modeling fidelity of this MoRF model, as measure by PSNR, was 35.98 on the real training images and 36.62 on a set of synthetic validation views (skin areas only, not included in training). Note that MoRF achieves this performance with a single, multi-subject radiance field model. A PSNR comparison against a one-subject NeRF and variants of MoRF's architecture is given in the supplementary file.

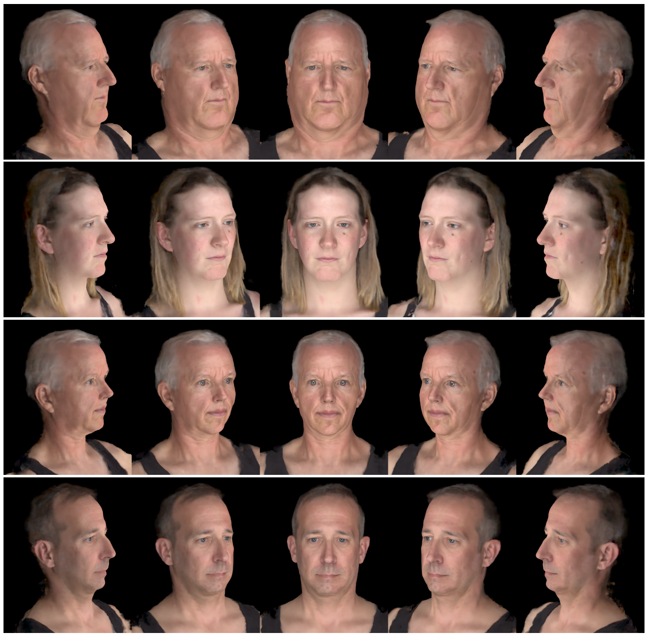

Figure 6: Example "turntable" views for three novel synthetic subjects obtained by mixing deformation and canonical identity codes ($\mathbf{id_w}$ and $\mathbf{id_c}$). Additional views and subjects are shown in the supplementary document and video.

Generating Photoreal Subjects. A key benefit of learning a generative model is the ability to synthesize novel photoreal subjects that are different from those seen during training. This task is greatly

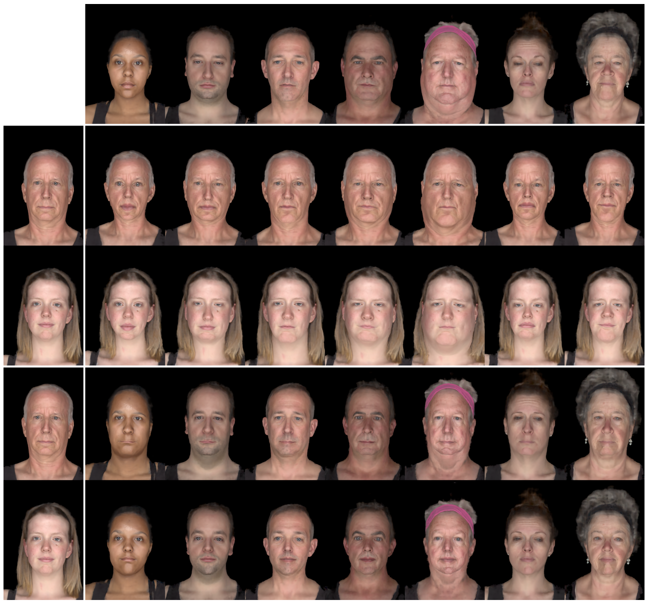

Figure 7: Novel synthetic subjects obtained by mixing deformation and canonical identity codes ($\mathbf{id}_\mathrm{w}$ and $\mathbf{id}_\mathrm{c}$) for two training subjects: (rows 2-3) the deformation code is taken from the subject at the top; and (rows 4-5) the canonical identity code is taken from the subject at the top.

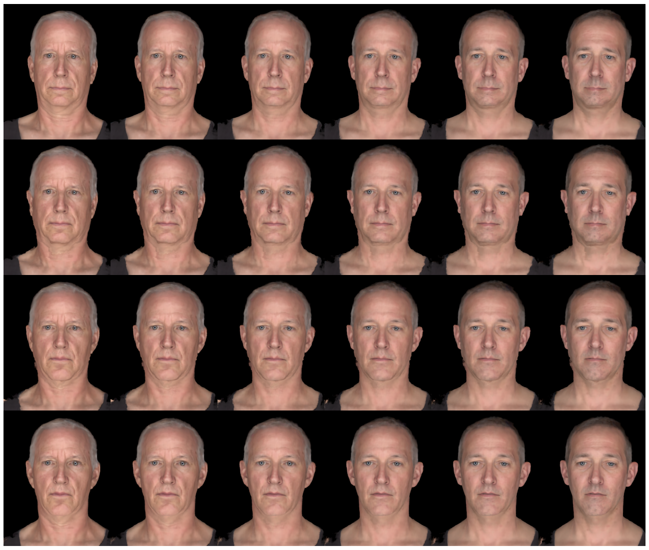

Figure 8: Novel synthetic heads interpolated between two real subjects (top left, bottom right) along the deformation code ($\mathbf{id}_\mathrm{w}$, vertical axis) and canonical identity code ($\mathbf{id}_\mathrm{c}$, horizontal axis); see also the supplementary video.

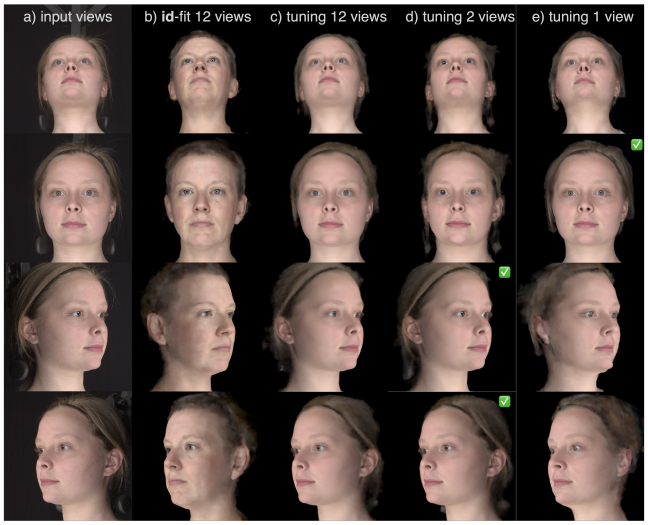

Figure 9: Fitting the pre-trained MoRF to images and 3D mesh of a new subject: (a) 4 of the 12 input images; (b) result from fitting id, then $\mathbf{id}_\mathrm{w}$ and $\mathbf{id}_\mathrm{c}$; and (c)-(e) results of code fitting, followed by tuning of the model's MLPs to 12, 2, and 1 input images (as marked), quickly yielding a new-subject NeRF.

facilitated by having a shared latent space across identities, from which new codes can be sampled randomly, while existing codes for training subjects can be easily blended to generate photorealistic renditions of people that do not exist in reality. Recall that MoRF has two intermediary latent codes, one for 3D geometry ($\mathbf{id}_\mathrm{w}$) and one for appearance ($\mathbf{id}_\mathrm{c}$) in the canonical space. Here, we take several pairs of real training subjects and combine the $\mathbf{id}_\mathrm{w}$ code of the first, with the $\mathbf{id}_\mathrm{c}$ code of the second subject in each pair (and *vice versa*), before feeding the new codes into MoRF. Novel turntable views of some of these synthetic humans are shown in Fig. 6, demonstrating the desired multiview consistency that is part of our main goal. Other synthesized subjects are tabulated in Fig. 7 and show how the appearance of a subject can be mapped onto the geometry of several other subjects, and *vice versa*. Once again, the results are photorealistic renditions of novel identities, showing a good disentanglement of the two code components that are predicted by MoRF's ID MLP. Another simple way of generating new subjects is to continuously interpolate across pairs of $\mathbf{id}_\mathrm{w}$ and/or $\mathbf{id}_\mathrm{c}$ codes, as show in Fig. 8, where interpolation is done between the two real training subjects on the top-left and bottom-right of the figure. The results are highly realistic novel identities with multiview-consistent renditions.

Fitting to New Subjects. MoRF and its ID spaces encode global identity information that can serve as prior in different tasks. Here, we show results of fitting the pre-trained MoRF to images of novel subjects not seen at training, effectively producing a NeRF from as few as one image and an optional 3D mesh. We adopt hierarchical fitting similar to that used with StyleGAN [Abdal et al. 2019]: we first fit an **id** code to initialize $\mathbf{id}_\mathrm{w}$ and $\mathbf{id}_\mathrm{c}$, which are then optimized further in their own subspaces. We found that MoRF trained on only 15 subjects is not expressive enough to faithfully reproduce a novel arbitrary identity, even when fitting to 12 images with a 3D mesh, Fig. 9 (b). In fact, even powerful StyleGAN models that are trained on thousands of identities cannot represent identity features that are unique to an arbitrary person not seen during training. This

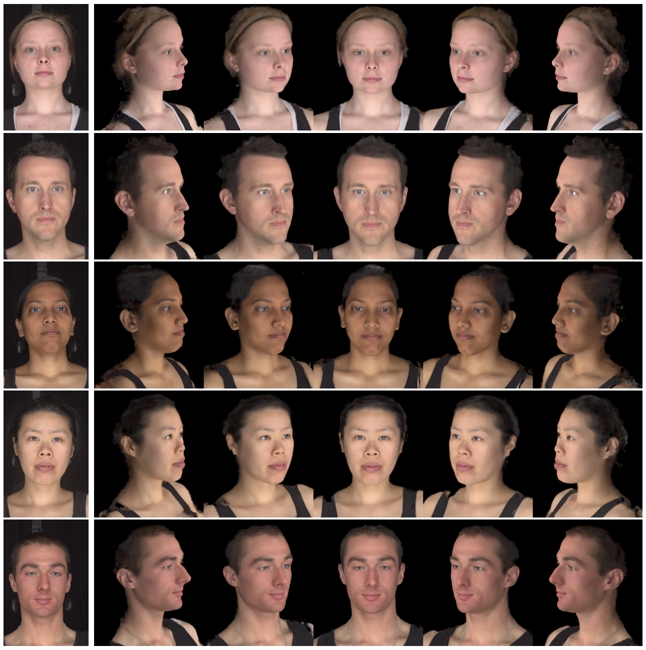

Figure 10: Tuning the pre-trained MoRF to *11 input views*, including the 3D mesh: (left) real image *left out for validation*; (right) 5 synthesized views of the 5 out-of-sample subjects.

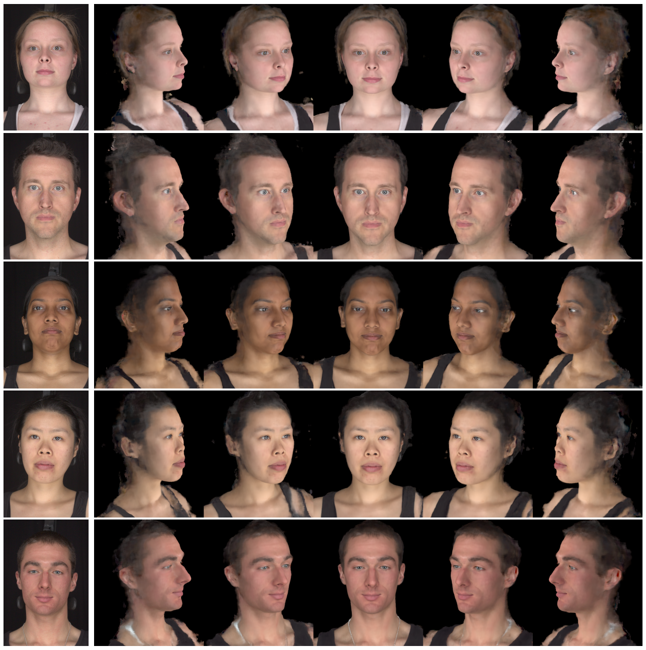

Figure 11: Tuning the pre-trained MoRF to a *single input view* (left), including the 3D mesh; (right) 5 synthesized "turntable" views of the 5 out-of-sample subjects.

fact has led to recent fitting methods that also tune the generator's weights (MLPs), for more faithful fits [Roich et al. 2021]. With this tuning approach, we can faithfully fit MoRF to an arbitrary new identity, Fig. 9 (c). Compared to training a standard NeRF for this subject, tuning MoRF takes about 1% of the number of iterations (2K vs 200K) and 5% of the optimization time (0.5h vs 20h). And MoRF can be tuned to as few as 1 input image, Fig. 9 (d)-(e); for these results, rendering PSNR on the remaining views left out of the optimization was, respectively, 28.33 and 23.72 (the single frontal view leads to worse renderings on the sides of the head). Finally, we tuned MoRF on a total of 5 test subjects, individually, using 11 real images and holding a frontal view out for PSNR evaluation, Fig. 10. On average, tuning with and without the 3D mesh provided similar PSNR (33.34 vs 33.36), indicating good constraining by the 11 images alone. As before, rendering errors are predominantly found on hair. When tuning to only one real view, Fig. 11, PSNR over the other 11 held-out images was 26.4 (25.8 without the 3D mesh) over the 5 subjects. Fig. 10 and Fig. 11 also show novel rendered views of these 5 out-of-sample subjects. A detailed description of the fitting and tuning procedures and additional results on all subjects are presented in the supplementary material.

5 CONCLUSION

We present MoRF, a morphable radiance field model that extends NeRFs into generative models that synthesize novel volumes and photorealistic images with full disentanglement of identity and 3D pose. Analogous to other morphable models for faces, MoRF offers parametric control over the resulting identity, for applications like synthesizing photorealistic new people or quickly fitting a NeRF to few images of a novel subject. The results show high-quality, multiview-consistent rendering from a wide range of views and complete with all head components such as eyes and hair.

As a main limitation, MoRF was so far trained on only 15 subjects; more training subjects will naturally constrain the latent space better, improving sampling of random new subjects, and fitting to new subjects with limited input. A larger training dataset should also allow for semantic control to be "discovered" by traversing the latent space using facial attribute classifiers, as done for StyleGAN [Härkönen et al. 2020; Shen et al. 2020]. Still, even with 15 subjects, we can already synthesize over 200 (nearly 15^2) novel, high-quality human heads via simple code mixing. Another natural extension would be to include training data with additional, non-neutral facial expressions (see proof of concept in the supplementary file). We also plan on adding more camera views to better model complex hair styles. Currently, interpolating between a subject with short hair and a subject with long hair tends to produce floating hair near the midway points (see supplementary video). While polarization-based capture already provides a more comprehensive and constrained model, future work will also extend MoRF to model diffuse and specular albedo separate from lighting. Despite the current limitations, we believe that MoRF already presents an impactful step forwards towards high-quality generative NeRFs for full-head modeling beyond facial skin.

REFERENCES

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2StyleGAN: How to embed images into the StyleGAN latent space?. In *Proc. ICCV*. IEEE, 4432–4441.

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020. Image2StyleGAN++: How to edit the embedded images?. In *Proc. CVPR*. IEEE, 8296–8305.

Victoria Fernandez Abrevaya, Adnane Boukhayma, Stefanie Wuhrer, and Edmond Boyer. 2019. A Decoupled 3D Facial Shape Model by Adversarial Training. In *Proc. ICCV*.

ShahRukh Athar, Zhixin Shu, and Dimitris Samaras. 2021. FLAME-in-NeRF : Neural control of Radiance Fields for Free View Face Animation. arXiv:2108.04913 [cs.CV]

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Siggraph*, Vol. 99. 187–194.

Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis. In *Proc. CVPR*.

Prashanth Chandran, Derek Bradley, Markus Gross, and Thabo Beeler. 2020. Semantic Deep Face Models. In *International Conference on 3D Vision*. 345–354.

Prashanth Chandran, Sebastian Winberg, Gaspard Zoss, Jérémy Riviere, Markus Gross, Paulo Gotardo, and Derek Bradley. 2021. Rendering with Style: Combining Traditional and Neural Approaches for High-Quality Face Rendering. *ACM Trans. Graph.* 40, 6 (dec 2021).

Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. 2001. Active Appearance Models. *IEEE Trans. PAMI* 23, 6 (jun 2001), 681–685.

Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. 2021. Unconstrained Scene Generation with Locally Conditioned Radiance Fields. In *Proc. ICCV*.

Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 2020. 3D Morphable Face Models - Past, Present and Future. *ACM Trans. Graph.* 39, 5 (2020).

Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2021. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. In *Proc. CVPR*. 8649–8658.

Chen Gao, Yichang Shih, Wei-Sheng Lai, Chia-Kai Liang, and Jia-Bin Huang. 2020. Portrait Neural Radiance Fields from a Single Image. *arXiv preprint arXiv:2012.05903* (2020).

S. Gong, L. Chen, M. Bronstein, and S. Zafeiriou. 2019. SpiralNet++: A Fast and Highly Efficient Mesh Convolution Operator. In *Proc. ICCV Workshops*. 4141–4148.

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2022. StyleNeRF: A Style-based 3D Aware Generator for High-resolution Image Synthesis. In *International Conference on Learning Representations*.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. GANSpace: Discovering Interpretable GAN Controls. In *Proc. NeuIPS*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 9841–9850.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-Free Generative Adversarial Networks. In *Proc. NeurIPS*.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proc. CVPR*. 4401–4410.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*. IEEE, 8110–8119.

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a Model of Facial Shape and Expression from 4D Scans. *ACM Trans. Graph.* 36, 6 (nov 2017), 17 pages.

Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. 2021. Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes. In *Proc. CVPR*.

Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural Volumes: Learning Dynamic Renderable Volumes from Images. *ACM Trans. Graphics (Proc. SIGGRAPH)* 38, 4, Article 65 (July 2019), 14 pages.

Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. 2021. Pixel Codec Avatars.

Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. 2021. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *Proc. CVPR*.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

Richard A Newcombe, Dieter Fox, and Steven M Seitz. 2015. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proc. CVPR*. 343–352.

Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. 2019. HoloGAN: Unsupervised Learning of 3D Representations From Natural Images. In *Proc. ICCV*.

Michael Niemeyer and Andreas Geiger. 2021. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. In *Proc. CVPR*.

Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. 2021a. Nerfies: Deformable Neural Radiance Fields. *Proc. ICCV* (2021).

Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. 2021b. HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields. *ACM Trans. Graph.* 40, 6, Article 238 (dec 2021).

Stylianos Ploumpis, Evangelos Ververas, Eimear O'Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William Smith, Baris Gecer, and Stefanos P Zafeiriou. 2020. Towards a complete 3D morphable model of the human head. *IEEE Trans. PAMI* (2020).

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2020. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Proc. CVPR*.

Amit Raj, Michael Zollhoefer, Tomas Simon, Jason Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. 2021. PVA: Pixel-aligned Volumetric Avatars. In *Proc. CVPR*.

Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. 2021. H3D-Net: Few-Shot High-Fidelity 3D Head Reconstruction. In *Proc. ICCV*. 5620–5629.

Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. 2018. Generating 3D faces using Convolutional Mesh Autoencoders. In *Proc. ECCV*.

Jérémy Riviere, Paulo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2020. Single-Shot High-Quality Facial Geometry and Skin Appearance Capture. *ACM Trans. Graphics (Proc. SIGGRAPH)* 39, 4, Article 81 (2020), 12 pages.

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. 2021. Pivotal Tuning for Latent-based Editing of Real Images. *arXiv preprint arXiv:2106.05744* (2021).

Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. 2020. GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. In *Proc. NeurIPS*.

Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. 2020. InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs. *IEEE Trans. PAMI* PP (Oct. 2020).

Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. 2021. Non-Rigid Neural Radiance Fields: Reconstruction and Novel View Synthesis of a Dynamic Scene From Monocular Video. In *Proc. ICCV*. IEEE.

Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. 2005. Face Transfer with Multilinear Models. *ACM Trans. Graph.* 24, 3 (2005), 426–433.

Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhöfer. 2020. Learning Compositional Radiance Fields of Dynamic Human Heads.

Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. 2021. Space-time Neural Irradiance Fields for Free-Viewpoint Video. In *Proc. CVPR*. 9421–9431.

Christopher Xie, Keunhong Park, Ricardo Martin-Brualla, and Matthew Brown. 2021. FiG-NeRF: Figure-Ground Neural Radiance Fields for 3D Object Category Modelling. In *International Conference on 3D Vision (3DV)*.

Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. 2021. i3DMM: Deep Implicit 3D Morphable Model of Human Heads. In *Proc. CVPR*. 12803–12813.

Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. 2021. PlenOctrees for Real-time Rendering of Neural Radiance Fields. In *Proc. ICCV*.

Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. 2021. CIPS-3D: A 3D-Aware Generator of GANs Based on Conditionally-Independent Pixel Synthesis. *arXiv preprint arXiv:2110.09788* (2021). arXiv:2110.09788 [cs, eess]