RigNeRF: Fully Controllable Neural 3D Portraits

ShahRukh Athar*
Stony Brook University
sathar@cs.stonybrook.edu

Zexiang Xu
Adobe Research
zexu@adobe.com

Kalyan Sunkavalli
Adobe Research
sunkaval@adobe.com

Eli Shechtman
Adobe Research
elishe@adobe.com

Zhixin Shu
Adobe Research
zshu@adobe.com

(a) Pose Control (b) Expression Control (c) Novel View Synthesis (d) Video based reanimation with novel view synthesis

Figure 1. Our method, RigNeRF, enables full control of head-pose and facial expression of a human portrait using a 3DMM-guided deformable neural radiance field. With a short portrait video for training, RigNeRF can reanimate the subject with arbitrary (a) head-pose and (b) facial expressions. It also allows controllable viewpoints of the 3D scene (c). Naturally RigNeRF can be used to transfer facial animation from a driving video sequence to the 3D portraits faithfully (d).

Abstract

Volumetric neural rendering methods, such as neural radiance fields (NeRFs), have enabled photo-realistic novel view synthesis. However, in their standard form, NeRFs do not support the editing of objects, such as a human head, within a scene. In this work, we propose RigNeRF, a system that goes beyond just novel view synthesis and enables full control of head pose and facial expressions learned from a single portrait video. We model changes in head pose and facial expressions using a deformation field that is guided by a 3D morphable face model (3DMM). The 3DMM effectively acts as a prior for RigNeRF that learns to predict only residuals to the 3DMM deformations and allows us to render novel (rigid) poses and (non-rigid) expressions that were not present in the input sequence. Using only a smartphone-captured short video of a subject for training, we demonstrate the effectiveness of our method on free view synthesis of a portrait scene with explicit head pose and expression controls.

1. Introduction

Photo-realistic editing of human portraits is a long-standing topic in the computer graphics and computer vision community. It is desirable to be able to control certain attributes of a portrait, such as 3D viewpoint, lighting, head pose, and even facial expression, after capturing. It also has great potential in AR/VR applications where a 3D immersive experience is valuable. However, it is a challenging task: modeling and rendering a realistic human portrait with complete control over 3D viewpoint, facial expressions, and head pose in natural scenes remains elusive, despite the longtime interest and recently increased research.

3D Morphable Face Models (3DMMs) [4] were among the earliest attempts towards a fully controllable 3D human

*Work done while interning at Adobe Research.

head model. 3DMMs use a PCA-based linear subspace to control face shape, facial expressions, and appearance independently. A face model of desired properties can be rendered in any view using standard graphics-based rendering techniques such as rasterization or ray-tracing. However, directly rendering 3DMMs [4], which only models face region, is not ideal for photo-realistic applications as its lacks essential elements of the human head such as hair, skin details, and accessories such as glasses. Therefore, it is better employed as an intermediate 3D representation [21, 47, 48] due to its natural disentanglement of face attributes such as shape, texture, and expression, which makes 3DMMs an appealing representation for gaining control of face synthesis.

On the other hand, recent advances on neural rendering and novel view synthesis [3, 6, 13, 14, 28, 29, 32, 33, 35, 39, 50, 52, 53] have demonstrated impressive image-based rendering of complex scenes and objects. Despite that, existing works are unable to simultaneously generate high quality novel views of a given natural scene and control the objects within it, including that of the human face and its various attributes. In this work, we would like to introduce a system to model a fully controllable portrait scene: with camera view control, head pose control, as well as facial expression control.

Control of head-pose and facial expressions can be enabled in NeRFs via a deformation module as done in [35, 36, 39]. However, since those deformations are learnt in a latent space, they cannot be explicitly controlled. A natural way add control to head-pose and facial expressions, via deformations, is by parameterizing the deformation field using the 3DMM head-pose and facial expression space. However, as shown in Fig 7, such naive implementation of a deformation field leads to artefacts during the reanimation due to the loss of rigidity and incorrect modelling of facial expressions. To address these issue, we introduce RigNeRF, a method that leverages a 3DMM to generate a coarse deformation field which is then refined by corrective residual predicted by an MLP to account for the non-rigid dynamics, hair and accessories. Beyond giving us a controllable deformation field, the 3DMM acts as an inductive bias allowing our network to generalize to *novel* head poses and expressions that were not observed in the input video.

Our model is designed to be trained on a short video captured using a mobile device. Once trained, RigNeRF allows for explicit control of head pose, facial expression and camera viewpoint. Our results capture rich details of the scene along with details of the human head such as the hair, beard, teeth and accessories. Videos reanimated using our method maintain high fidelity to both the driving morphable model in terms of facial expression and head-pose and the original captured scene and human head.

In summary, our contributions in this paper are as follows: 1) We propose a neural radiance field capable of full control of the human head along with simultaneously modelling the full 3D scene it is in. 2) We experimentally demonstrate the loss of rigidity when dynamic neural radiance fields are reanimated. 3) We introduce a deformation prior that ensures rigidity of the human head during reanimation thus significantly improves its quality.

2. Related works

RigNeRF is a method for full control of head pose, facial expressions, and novel view synthesis of 3D portrait scene. It is closely related to recent work on neural rendering, novel view synthesis, 3D face modeling, and controllable face generation.

Neural Scene Representations and Novel View Synthesis. RigNeRF is related to recent work in neural rendering and novel view synthesis [3, 6, 13, 14, 25, 28, 29, 31–36, 39, 40, 45, 49–53]. Neural Radiance Fields (NeRF) use a Multi-Layer Perceptron (MLP), F, to learn a volumetric representation of a scene. For every 3D point and the direction from which the point is being viewed, F predicts its color and volume density. For any given camera pose, F is first evaluated densely enough throughout the scene using hierarchical volume sampling [33], then volume rendering is used to render the final image. F is trained by minimizing the error between the predicted color of a pixel and its ground truth value. While NeRFs are able to generate photo-realistic images for novel view synthesis, it is only designed for a static scene and is unable to represent scene dynamics. Specifically designed for dynamic portrait video synthesis, our approach not only models the dynamics of human faces, but also allows specific controls on the facial animation.

Dynamic Neural Scene Representations. Although NeRF [33] is designed for a static scene, several works have attempted to extend it to model dynamic objects or scene. There is a line of work [27, 28, 39, 50] that extend NeRF to dynamic scenes by providing as input a time component and along with it imposing temporal constraints either by using scene flow [28, 50] or by using a canonical frame [39]. Similarly, Nerfies [35] too work with dynamic scenes by mapping to a canonical frame, however it assumes that the movement is small. In [36], authors build upon [35] and use an ambient dimension in order to model topological changes in the deformation field. The deformation fields in these approaches are conditioned on learnt latent codes without specific physical or semantic meaning, and therefore not controllable in an intuitive manner. RigNeRF, similar to [35, 36], models the portrait video by mapping to a canonical frame but in addition also enables full parameterized control of head pose and facial expression.

Controllable Face Generation. Recent breakthroughs in Generative Adversarial Networks(GANs) [15, 17–20, 55]

have enabled high-quality image generation and manipulation. They also inspired a large collection of work [2, 7–9, 24, 38, 42, 43, 46, 47] focusing on face image manipulation and editing. However, majority of these work are intrinsically image-based and lack explicit 3D representation. Therefore it is challenging to enable high-quality view synthesis and 3D controls of the portraits such as large pose changes or extreme facial expressions. Another line of work [1, 10, 22, 23] made use of 3D Morphable Model as intermediate 3D face representation to reanimate face images/videos. While being able to model head poses with great detail, thanks to the disentangled representation in 3DMM, they often unable to perform novel view synthesis as they focus on face region but neglect the geometry or appearance of the scene. Similarly, NerFACE [13] uses neural radiance fields to model a 4D face avatar and allows pose/expression control on the head. However, they assume a static background and fixed camera, thus cannot perform view synthesis of the person or the scene. In contrast, our method RigNeRF provides full control over the head pose and facial expressions of the person captured in the portrait video while simultaneously being able to synthesize novel views of the 3D portrait scene.

Hybrid Representations. The photorealism of volumetric and implicit representations have encouraged works that combine them with classical representations in order improve reconstruction [5] or lend control over foreground [12, 30, 37]. In [30], authors learn a deformation field along with a texture mapping to reanimate human bodies. Similarly, [37] learns a 3D skinning field to accurately deform points according to the target pose. Neither [30] nor [37] model the full 3D scene. In contrast, RigNeRF models the whole 3D scene with full control of head-pose, facial expressions and viewing directions.

3. RigNeRF

In this section, we describe our method, RigNeRF, that enables novel view synthesis of 3D portrait scenes and arbitrary control of head pose and facial expressions. A Neural Radiance Field (NeRF) [33] with a per-point deformation is used to control the head pose and facial expressions of the subject. The deformation field deforms the rays of each frame to a canonical space, defined by a the 3DMM in the frontal head-pose and neutral expression, where the colors are sampled. In order to model deformations due to both head-pose (a rigid deformation) and facial expressions (a non-rigid deformation) and to correctly deform facial details such as hair and glasses, the deformation field is defined as the sum of the 3DMM deformation field and a residual deformation predicted by a deformation MLP.

3.1. Deformable Neural Radiance Fields

A neural radiance field (NeRF) is defined as a continuous function $F : (\boldsymbol{\gamma}_m(\mathbf{x}), \boldsymbol{\gamma}_n(\mathbf{d})) \rightarrow (\mathbf{c}(\mathbf{x}, \mathbf{d}), \sigma(\mathbf{x}))$, that, given the position of a point in the scene $\mathbf{x}$ and the direction it is being viewed in, $\mathbf{d}$, outputs the color $\mathbf{c} = (r, g, b)$ and the density σ. F is usually represented as a multi-layer perceptron (MLP) and $\boldsymbol{\gamma}_m : \mathbb{R}^3 \rightarrow \mathbb{R}^{3+6m}$ is the positional encoding [33] defined as $\boldsymbol{\gamma}_m(\mathbf{x}) = (\mathbf{x}, ..., \sin(2^k\mathbf{x}), \cos(2^k\mathbf{x}), ...)$ where m is the total number of frequency bands and $k \in \{0, ..., m-1\}$. The expected color of the pixel through which a camera ray passes is calculated via volume rendering. The parameters of F are trained to minimize the L2 distance between the expected color and the ground-truth.

NeRFs, as defined above, are designed for static scenes and offer no control over the objects within the scene. In order to model a dynamic scene, NeRFs are extended by additionally learning a deformation field to map each 3D point of the scene to a canonical space, where the volumetric rendering takes place [35, 36, 39]. The deformation field is also represented by an MLP $D_i : \mathbf{x} \rightarrow \mathbf{x}_{\text{can}}$ where D_i is defined as $D(\mathbf{x}, \omega_i) = \mathbf{x}_{\text{can}}$ and ω_i is a per-frame latent deformation code. In addition to a deformation code, ω_i, a per-frame appearance code is also used [35, 36, 39], ϕ_i, thus the final radiance field for the i-th frame is as follows:

$$(\mathbf{c}(\mathbf{x}, \mathbf{d}), \sigma(\mathbf{x})) = F\left(\boldsymbol{\gamma}(D(\mathbf{x}, \omega_i)), \boldsymbol{\gamma}(\mathbf{d}), \phi_i\right) \quad (1)$$

In addition to the parameters of F, each ω_i and ϕ_i are also optimized through stochastic gradient descent. While the aforementioned modifications are able generate novel views [35, 39] of dynamic videos and handle small movement of objects in the scene [35], the deformations they estimate are conditioned on learnt deformation codes that can be arbitrary. Instead, we seek intuitive deformation controls that explicitly disentangles and controls facial appearance based on camera viewpoint, head pose and expression.

3.2. A 3DMM-guided deformation field

RigNeRF enables novel view synthesis of dynamic portrait scene and arbitrary control of head pose and facial expressions. For each frame i, we first extract its head-pose and expression parameters $\{\beta_{i,\text{exp}}, \beta_{i,\text{pose}}\}$ using DECA [11] and landmark fitting [16]. Next, we shoot rays through each pixel, p, of the frame and deform each point on the ray, $\mathbf{x}$, to a position in the canonical space, $\mathbf{x}_{\text{can}} = (x', y', z')$, where its color is computed. A natural way to parameterize this canonical space, and any deviations from it, is using 3DMMs [4, 26]. Thus, RigNeRF's canonical space is defined as the one where the head has zero head-pose and a neutral facial expression.

Unfortunately, a 3DMM is only defined accurately for a subset of points on the head—3DMM fitting is often not perfect and they do not model hair, glasses, etc.—and is

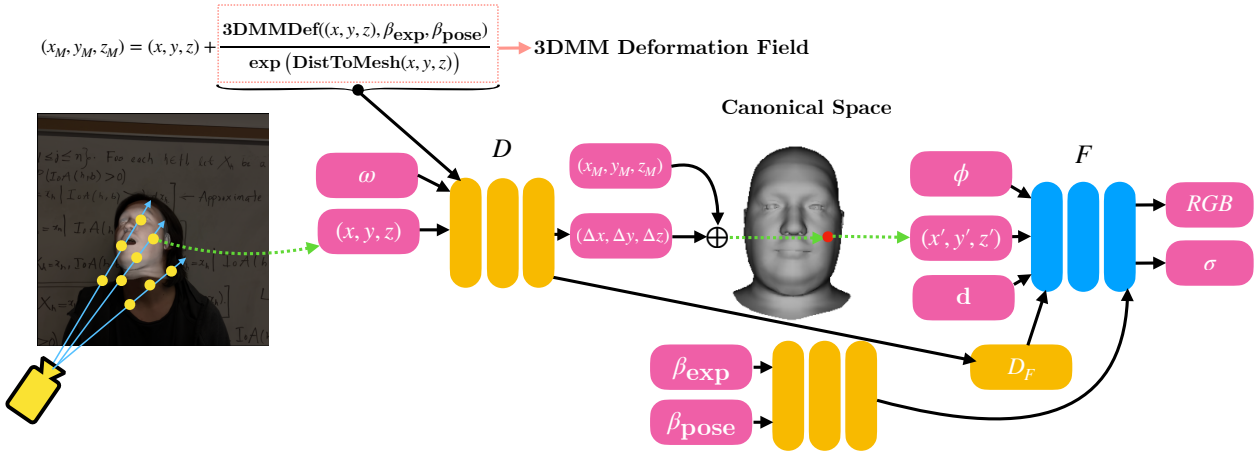

Figure 2. **Overview of RigNeRF.** RigNeRF is a deformable NeRF architecture that consists of two learnable MLPs: a deformation MLP D and a color MLP F. Given an image, we shoot rays through each of its pixels. For every ray, we deform each point on it according to a 3DMM-guided deformation field. This deformation field is the sum of the 3DMM deformation field (see Sect 3.2) and the residual predicted by the deformation MLP, D. Next, the deformed point is given as input to the color MLP, F, which additionally takes as input the pose and expression parameters $\{\beta_{\text{exp}}, \beta_{\text{pose}}\}$, the viewing direction $\mathbf{d}$ and an appearance embedding ϕ to predict the color and density. The final color of the pixel is calculated via volume rendering.

undefined for point in the rest of 3D space. Hence, a deformation MLP $D_i : \mathbf{x} \rightarrow \mathbf{x}_{\text{can}}$ is still necessary to perform the transformation to the canonical space. However, as detailed in Sect 4.3, we find that directly predicting the deformation to the canonical space gives rise to artefacts during reanimation. The artefacts arise due to the inability of D to 1) maintain the rigidity of the head and 2) to model facial expressions correctly.

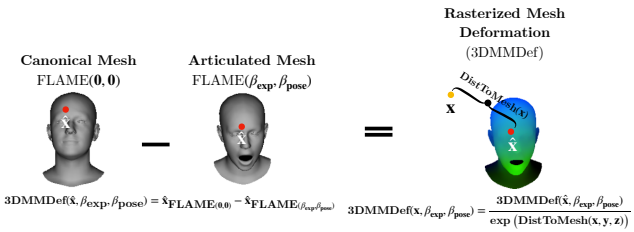

Figure 3. **The 3DMM deformation field.** The 3DMM deformation field at any point in space, $\mathbf{x}$, is equal to the deformation of its closest neighbor on the mesh, $\hat{\mathbf{x}}$, weighted by the inverse of the exponential of the distance between $\mathbf{x}$ and $\hat{\mathbf{x}}$.

In order to fix this and ensure RigNeRF is able to handle both rigid deformations due to head-pose changes and non-rigid deformations due to changes in facial expressions, we use a deformation field prior derived using the 3DMM. For expression and head-pose parameters, $\{\beta_{\text{exp}}, \beta_{\text{pose}}\}$, the value of the 3DMM deformation field at any point $\mathbf{x} = (x, y, z)$ is:

$$\text{3DMMDef}(\mathbf{x}, \beta_{\text{exp}}, \beta_{\text{pose}}) = \frac{\text{3DMMDef}(\hat{\mathbf{x}}, \beta_{\text{exp}}, \beta_{\text{pose}})}{\exp(\text{DistToMesh}(\mathbf{x}))} \quad (2)$$

where, $\text{3DMMDef}(\mathbf{x})$ is the value of the 3DMM deformation field, $\hat{\mathbf{x}} = (\hat{x}, \hat{y}, \hat{z})$ is the closest point to (x, y, z) on the mesh and $\text{DistToMesh} = ||\mathbf{x} - \hat{\mathbf{x}}||$ is the distance between $\mathbf{x}$ and $\hat{\mathbf{x}}$. The 3DMM deformation of any point on the mesh, $\hat{\mathbf{x}}$, is given by the difference between its position in the canonical space (i.e when the mesh had a zero head-pose and neutral facial expression) and its current articulation, as follows:

$$\text{3DMMDef}(\hat{\mathbf{x}}, \beta_{\text{exp}}, \beta_{\text{pose}}) = \hat{\mathbf{x}}_{\text{FLAME}(\mathbf{0},\mathbf{0})} - \hat{\mathbf{x}}_{\text{FLAME}(\beta_{\text{exp}},\beta_{\text{pose}})} \quad (3)$$

where, $\hat{\mathbf{x}}_{\text{FLAME}(\mathbf{0},\mathbf{0})}$ is the position of $\mathbf{x}$ in the canonical space and $\hat{\mathbf{x}}_{\text{FLAME}(\beta_{\text{exp}},\beta_{\text{pose}})}$ is its position with head pose and facial expression parameters $\{\beta_{\text{exp}}, \beta_{\text{pose}}\}$.

The RigNeRF deformation field can now be defined as the sum of the 3DMM deformation field and the residual predicted by D, as follows

$$\begin{aligned} \hat{D}(\mathbf{x}) &= \text{3DMMDef}(\mathbf{x}, \beta_{\text{i,exp}}, \beta_{\text{i,pose}}) \\ &\quad + D(\boldsymbol{\gamma}_a(\mathbf{x}), \boldsymbol{\gamma}_b(\text{3DMMDef}(\mathbf{x}, \beta_{\text{i,exp}}, \beta_{\text{i,pose}})), \omega_i) \\ \mathbf{x}_{\text{can}} &= \mathbf{x} + \hat{D}(\mathbf{x}) \end{aligned} \quad (4)$$

where, $\hat{D}(\mathbf{x})$ is the value of the RigNeRF deformation field at $\mathbf{x}$, $\{\boldsymbol{\gamma}_a, \boldsymbol{\gamma}_b\}$ is the positional embedding on $\mathbf{x}$ and $\text{3DMMDef}(\mathbf{x}, ...)$ respectively and ω_i is the deformation embedding for current frame. We use ω_i to model deformations that cannot be accounted for by head-pose and expression changes. Experimentally, we find that conditioning D directly on the expression and pose parameters, $\{\beta_{\text{i,exp}}, \beta_{\text{i,pose}}\}$, leads to overfitting and poor generalization. This is likely due to the high dimensionality of the code (59), that makes it prone to overfitting. Instead, we condition D on the 3DMM

deformation of the point $\mathbf{x}$, 3DMMDef$(\mathbf{x}, \beta_{\text{i,exp}}, \beta_{\text{i,pose}})$. Since 3DMMDef$(\mathbf{x}, \beta_{\text{i,exp}}, \beta_{\text{i,pose}}) \in \mathbb{R}^3$, it is itself relatively low dimensional, and it can be pushed into higher dimensions by adjusting the number of frequencies of its positional embedding, $\boldsymbol{\gamma}_b$. We find that using $b = 2$ frequencies in $\boldsymbol{\gamma}_b$ for the 3DMM deformation, 3DMMDef$(\mathbf{x}, \beta_{\text{i,exp}}, \beta_{\text{i,pose}})$, works the best. In Fig 4, we show renders of both the output of D and the RigNeRF deformation field, $\hat{D}$, as described in Eq. (4). In Fig 4(c), we see that D generates the accurate deformation around the glasses, which the 3DMM deformation cannot do, for both head-poses. In Fig 4(d), we see that the $\hat{D}$ is only concentrated on the head as it should be.

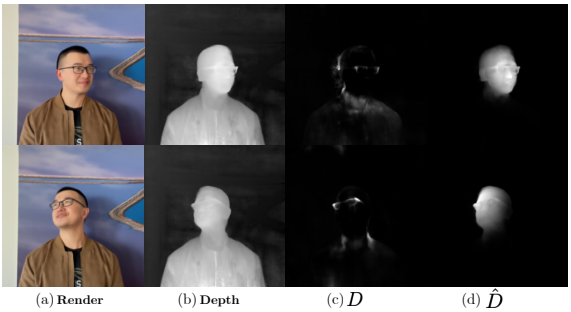

Figure 4. **Visualize learnt depth and deformation in RigNeRF.** Here we show the depth, the magnitude of the output of the Deformation MLP D, and magnitude of the $\hat{D}$ i.e the sum of the 3DMM Deformation and D. In (b), we can see that despite large changes in head-pose, the depth remains consistent. Next, in (c), we see D generates a deformation around the glasses for both poses so that it can be accurately deformed along with the head. Finally, in the last column we see how $\hat{D}$ is only concentrated on the head.

3.3. 3DMM-conditioned Appearance

In order to accurately model expression and head-pose based textures, such as teeth, we condition F on both expression and head-pose parameters and on features extracted from the penultimate layer of the deformation MLP $D(\boldsymbol{\gamma}_a(\mathbf{x}), ...)$. We find that using these features as input improves the overall quality of the render, please check the supplementary for details. Thus, once a point $\mathbf{x}$ has been deformed to its location in the canonical space, $\mathbf{x}_{\text{can}}$, using Eq. (4), its color is calculated as follows:

$$\mathbf{c}(\mathbf{x}, \mathbf{d}), \sigma(\mathbf{x}) = F(\boldsymbol{\gamma}_c(\mathbf{x}_{\text{can}}), \boldsymbol{\gamma}_d(\mathbf{d}), \phi_i, D_{F,i}(\mathbf{x}_{\text{can}}), \beta_{\text{i,exp}}, \beta_{\text{i,pose}}) \tag{5}$$

where, $\mathbf{d}$ is the viewing direction, $\boldsymbol{\gamma}_c, \boldsymbol{\gamma}_d$ is the positional embedding on $\mathbf{x}_{\text{can}}$ and $\mathbf{d}$, and $D_{F,i}(\mathbf{x}_{\text{can}})$ are features from the penultimate layer of the deformation MLP $D(\boldsymbol{\gamma}_a(\mathbf{x}), ...)$. The pixel color for p is then calculated using volume rendering and parameters of RigNeRF are minimized w.r.t to the ground truth color of p. The full architecture is shown in Fig 2.

4. Results

In this section, we show results of head-pose control, facial expression control, and novel view synthesis using RigNeRF. For each scene, the model is trained on a short portrait video captured using a consumer smartphone.

Baseline approaches. To the best of our knowledge, RigNeRF is the first method that enables dynamic control of head-pose, facial expressions along with the ability to synthesize novel views of full portrait scenes. Thus, there is no existing work for an apple-to-apple comparison. We qualitatively and quantitatively compare our method to three other methods that perform closely related tasks: (1) HyperNeRF [36]: a state-of-the-art method using NeRFs for novel view synthesis of dynamic portrait scenes, *without* any control (2) NerFACE [12], a state-of-the-art method using NeRF for face dynamics control *without* modeling camera viewpoint and the entire scene, and (3) First Order Motion Model (FOMM) [44], a general-purpose image reanimation pipeline.

When generating renanimation videos, RigNeRF, HyperNeRF [36] and NerFACE [12] require an appearance code for rendering; we use the appearance code of the first frame here. Similarly, RigNeRF and Nerfies [35] require a deformation code and we use the deformation code from the first frame. Full videos of the reanimation can be found in the supplementary material. We strongly urge the readers to refer to the videos to evaluate the quality of the results.

Training Data Capture and Training details. The training and validation data was captured using an iPhone XR or iPhone 12 for all the experiments in the paper. In the first half of the capture, we ask the subject to enact a wide range of expressions and speech while trying to keep their head still as the camera is panned around them. In the next half, the camera is fixed at head-level and the subject is asked to rotate their head as they enact a wide range of expressions. Camera parameters are calculated using COLMAP [41]. We calculate the expression and shape parameters of each frame in the videos using DECA [11] and further optimize them using via standard landmark fitting using the landmarks predicted by [16] and camera parameters given by COLMAP [41]. All training videos are between 40-70 seconds long ($\sim$ 1200-2100 frames). Due to compute restrictions, the video is down-sampled and the models are trained at 256x256 resolution. We use coarse-to-fine and vertex deformation regularization [35] to train the deformation network $D(\mathbf{x}, \omega_i)$. Please find full details of each experiment in the supplementary.

4.1. Evaluation on Test Data

We evaluate RigNeRF, HyperNeRF [36], NerFACE [12] and FOMM [44] on held out images on the captured video

(a) Fixed view with varying expression and pose

(b) Varying expression, pose and view

Figure 5. **Qualitative comparison by reanimation with novel facial expression, head-pose, and camera view parameters.** Here we reanimate RigNeRF, HyperNeRF [36] and NerFACE [12] using facial expression and head-pose derived from source images (top-row). We observe that while HyperNeRF [36] is able to generate realistic looking images of a portrait, it is unable to control head pose or facial expression in the result. On the other hand, NerFACE [12] attempts to render the correct pose and expression, but is unable to generate plausible face regions. Since, NerFACE [12] lacks an explicit deformation module it is unable to model deformation due to head-pose and facial expression changes. In contrast, our approach RigNeRF can effectively control the head pose, facial expression, and camera view, generating high quality facial appearance.

	Subject 1			*Subject 2*			*Subject 3*			*Subject 4*		
Models	PSNR ↑	LPIPS ↓	FaceMSE ↓	PSNR ↑	LPIPS ↓	FaceMSE ↓	PSNR ↑	LPIPS ↓	FaceMSE ↓	PSNR ↑	LPIPS ↓	FaceMSE ↓
RigNeRF (Ours)	29.55	0.136	9.6e-5	29.36	0.102	1e-4	28.39	0.109	8e-5	27.0	0.092	2.3e-4
HyperNeRF [36]	24.58	0.22	8.14e-4	22.55	0.1546	9.48e-4	19.29	0.26	2.74e-3	21.19	0.182	1.58e-3
NerFACE [13]	24.2	0.217	7.84e-4	24.57	0.174	6.7e-4	28.00	0.1292	1.2e-4	28.47	0.134	2.7e-4
FOMM [44]	11.45	0.432	7.65e-3	12.7	0.582	6.31e-3	10.17	0.601	1.7e-2	11.17	0.529	6.8e-3

Table 1. Quantitative results of Subject 1,2,3 and 4 on test data. Our results are better than HyperNeRF [36], NerFACE [12] and FOMM [44] on most metrics across all subjects.

sequences. We use the camera view, pose and expression parameters of these images. Since RigNeRF and HyperNeRF [36] use a per-frame deformation ω_i, we can't use the first frame (which is what we use as default for reanimation) to perform a direct comparison with the ground truth image as it may have a different deformation to the canonical space than the first frame. Therefore, we first optimize for the deformation code, ω_v, of a given validation image by minimizing rendering error wrt that frame as follows:

$$\omega_v = \min_{\omega} ||C_p(\omega; \mathbf{x}, \mathbf{d}, \theta, \phi_0, \beta_{\text{i,exp}}, \beta_{\text{i,pose}}) - C_p^{GT}|| \quad (6)$$

where, $C_p(\omega; \mathbf{x}, \mathbf{d}, \theta, \phi_0)$ is the predicted color at pixel p generated using Eq. (5) and volume-rendering , ϕ_0 is the appearance code of the first frame, θ are the parameters of F as defined in Eq. (5) and C_p^{GT} is the ground-truth pixel value. Note, we *only* optimize ω, all other parameters of the radiance field are kept fixed. We optimize Eq. (6) for 200 epochs which we observe to be more than enough to find the loss plateau. Once the optimization finishes, we report the final MSE, PSNR, LPIPS and Face MSE i.e the MSE only over the face region. We use no such optimization with NerFACE [12] since it does not have a deformation module and on FOMM [44] since it is a image-based method.

As can be seen in Table 1, our method outperforms HyperNeRF [36], NerFACE [12] and FOMM [44] on held-out test images. RigNeRF, HyperNeRF [36] and NerFACE [12] are trained on dynamic portrait videos with changing head-pose and facial expressions, HyperNeRF [36], lacking any head-pose and facial expression control, is unable to generate the head-poses and facial expressions seen in the held-out

(a) **Image-driven Reanimation**

(b) **Novel-View Synthesis**

(c) **3DMM-based control of 3D Portraits**

Figure 6. Applications of **RigNeRF**. RigNeRF allows for full control head-pose, facial expressions and viewing direction of 3D portrait scenes. This enables application like (a) Image-driven Reanimation, (b) Novel View Synthesis and (c) 3DMM-based control of 3D Portraits. In (a)-top row, we show images of "driving sequences" from which we extract pose and expression parameters; the results from 4 subjects are shown in (a)-bottom where we synthesize realistic portrait frames that closely matches the driving pose and expression. We show in (b) a set of view synthesis results where we fix the head pose and facial expression, rendering results with varying camera positions, in which we show high-quality results with dramatic view changes. In (c), we demonstrate the application of controlling the portrait appearance with explicit 3DMM parameters. Within each row, (c)-column 1 and (c)-column 2 have the same pose but different expressions; (c)-column 2 and (c)-column 3 have the different poses but share the same expression. The inset shows the input 3DMM pose and expression, both of which are faithfully rendered in the corresponding results. Please find more results in the supplemental document and video.

test set. Even as we optimized the deformation code of HyperNeRF [36], we found it hard to fit the unseen test images. NerFACE [12], on the other hand, lacking a deformation module, therefore is unable to model the dynamism of head-pose changes solely by concatenating pose and expression parameters as input the NeRF MLP. As a result, NerFACE [12] generates significant artefacts on the face regions (see the third row of Fig 5(a) and Fig 5(b)). FOMM [44], being an image based method, is unable to model novel views. Qualitative results of FOMM [44] can be found in the supplementary. In contrast to other methods, RigNeRF, thanks to the use of a 3DMM-guided deformation module, is able to model head-pose, facial expressions and the full 3D portrait scene with high fidelity, thus giving better reconstructions with sharp details.

4.2. Reanimation with pose and expression control

In this section we show results of reanimating a portrait video using both RigNeRF, HyperNeRF [36] and NerFACE [12] using expression and head-pose parameters as the driving parameters. Per-frame expression and head-pose parameters from the driving video are extracted using DECA [11]+Landmark fitting [16] and are given as input to RigNeRF in Eq. (4) and Eq. (5). Since HyperNeRF [36] does not take as input head-pose or expression parameters, it's forward pass remains unchanged.

First, in Fig 5(a) we show the results of changing head-pose and expression using a driving video while keeping view constant. As one can see, RigNeRF captures the driving head-pose and facial expressions with high fidelity without compromising the reconstruction of the entire 3D scene. In contrast, we see that HyperNeRF [36] is unable to change facial expressions or head-pose due to the lack of controls, while NerFACE [12] generates significant artefacts on the face, especially when head-pose is changed. In Fig 5(b), we show the results of *novel view synthesis along with changing head-pose and facial expressions*. Again, we see that RigNeRF reanimates the subject with the accurate head-pose and

Figure 7. **A qualitative comparison between RigNeRF and HyperNeRF+E/P.** Here we show a qualitative comparison between RigNeRF and HyperNeRF+E/P when reanimated using source images. We see that HyperNeRF+E/P generates many artefacts during reanimation due to its inability to model deformations correctly. In constrast, RigNeRF generates realistic reanimations with high fidelity to both the head-pose and facial expressions.

facial expressions *and is able to do so regardless of viewing direction* and without compromising the reconstruction of the background 3D scene. Again, HyperNeRF is unable to change the head-pose and facial expressions due to the aforementioned reason while NerFACE [12] generates significant artefacts during reanimation.

In Fig 6, we show more qualitative results of RigNeRF. We use three different applications of RigNeRF to demonstrate its flexibility and full controllability of a portrait scene. In Fig 6-(a), we show additional results of image-driven animation. The result frames (Fig 6-(a)-bottom) closely reproduce the head pose and facial expression shown in the driving frames (Fig 6-(a)-top). In Fig 6-(b), we show results of varying camera views while fixing (an arbitrary) facial expression and head pose. We demonstrate robust view synthesis performance with dramatic view changes. In Fig 6-(c), we show that RigNeRF can take user-specified 3DMM parameters as input to generate high-quality portrait images: each frame faithfully reproduces the face and expression provided by a set of 3DMM parameters shown in the inset.

4.3. Comparison with HyperNeRF+E/P

In this section we compare against HyperNeRF [36] with added pose and expression control, which we named HyperNeRF+E/P. The forward pass of this model is as follows

$$\begin{aligned} \mathbf{x}_{\text{can}} &= D(\boldsymbol{\gamma}_a(\mathbf{x}), \beta_{\text{i,exp}}, \beta_{\text{i,pose}}, \omega_i) \\ \mathbf{w} &= H(\boldsymbol{\gamma}_l(\mathbf{x}), \omega_i) \\ \mathbf{c}(\mathbf{x}, \mathbf{d}), \sigma(\mathbf{x}) &= F(\boldsymbol{\gamma}_c(\mathbf{x}_{\text{can}}), \boldsymbol{\gamma}_d(\mathbf{d}), \phi_i, \mathbf{w}, \beta_{\text{i,exp}}, \beta_{\text{i,pose}}) \end{aligned} \tag{7}$$

where, H is the ambient MLP and $\mathbf{w}$ are the ambient coordinates [36]. In Table 2, we show a quantitative comparison between RigNeRF and HyperNeRF+E/P. We see that RigNeRF is able to generate better reconstructions of the face and generates images that are perceptually closer to the ground truth, as measured by LPIPS [54], than those generated by HyperNeRF+E/P. In Fig 7 we show a qualitative comparison when both methods are reanimated using source images, where the head-pose and expression significantly differ from the training set. As can be seen, while HyperNeRF+E/P is able to turn the head, it is unable to do so rigidly (see row 3, column 2 and row 3, column 4 of Fig 7). Further, it is also unable to model facial expressions accurately and generates artefacts on the face region. This further demonstrates the benefit of our 3DMM-guided deformation learning.

	Subject 1			*Subject 2*		
Models	PSNR ↑	LPIPS ↓	FaceMSE ↓	PSNR ↑	LPIPS ↓	FaceMSE ↓
RigNeRF (Ours)	29.55	0.136	9.6e-5	29.36	0.102	1e-4
HyperNeRF+Exp	31.3	0.161	1.3e-4	30	0.116	1.9e-4

Table 2. Quantitative comparison between RigNeRF and HyperNeRF+E/P. We see that RigNeRF generates better reconstruction of the face with lower perceptual distance to the ground-truth as compared to HyperNeRF+E/P.

5. Limitations and Conclusion

Our method has certain limitations. First, it is subject specific and trains an individual model for each scene. Due to the need to capture sufficient expression and head-pose variations for training, out method currently requires a training sequences ranging from 40-70 seconds. Additionally, like all other NeRF-based methods, the quality of camera-registration affects the quality of the results. Being a method that allows photorealistic facial reanimation, RigNeRF may have potentially negative societal impact if misused. We discuss this further in the supplementary.

In conclusion, we present RigNeRF, a volumetric neural rendering model for fully controllable human portraits. Once trained, it allows full control of head-pose, facial expression, and viewing direction. In order to ensure generalization to novel head-pose and facial expression we use 3DMM-guided deformation field. This deformation field allows us to effectively model and control both the rigid deformations caused by head-pose change and non-rigid deformations of changes in facial expressions. Training with a short portrait video, RigNeRF enables applications includes image-based face reanimation, portrait novel-view synthesis, and 3DMM-based control of 3D portraits.

6. Acknowledgements

We would like to thank the anonymous CVPR reviewers for taking the time to review and suggest improvements to the paper. ShahRukh Athar is supported by a gift from Adobe, Partner University Fund 4DVision Project, and the SUNY2020 Infrastructure Transportation Security Center.

References

[1] ShahRukh Athar, Albert Pumarola, Francesc Moreno-Noguer, and Dimitris Samaras. Facedet3d: Facial expressions with 3d geometric detail prediction. *arXiv preprint arXiv:2012.07999*, 2020. 3

[2] S Athar, Z Shu, and D Samaras. Self-supervised deformation modeling for facial expression editing. 2020. 3

[3] Mojtaba Bemana, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. X-fields: Implicit neural view-, light-and time-image interpolation. 2020. 2

[4] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. 1999. 1, 2, 3

[5] Aggelina Chatziagapi, ShahRukh Athar, Francesc Moreno-Noguer, and Dimitris Samaras. Sider: Single-image neural optimization for facial geometric detail recovery. *arXiv preprint arXiv:2108.05465*, 2021. 3

[6] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *CVPR*, 2021. 2

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 3

[8] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 3

[9] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5154–5163, 2020. 3

[10] M. Doukas, Mohammad Rami Koujan, V. Sharmanska, A. Roussos, and S. Zafeiriou. Head2head++: Deep facial attributes re-targeting. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3:31–43, 2021. 3

[11] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. volume 40, 2021. 3, 5, 7

[12] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *CVPR*, June 2021. 3, 5, 6, 7, 8

[13] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction, 2020. 2, 3, 6

[14] 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

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[16] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3, 5, 7

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2

[21] H. Kim, P. Garrido, A. Tewari, Weipeng Xu, Justus Thies, M. Nießner, Patrick Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37:1 – 14, 2018. 2

[22] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM TOG*, 2018. 3

[23] M. Koujan, M. Doukas, A. Roussos, and S. Zafeiriou. Head2head: Video-based neural head synthesis. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*, pages 319–326, Los Alamitos, CA, USA, may 2020. IEEE Computer Society. 3

[24] Marek Kowalski, Stephan J Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. Config: Controllable neural face image generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 299–315. Springer, 2020. 3

[25] Christoph Lassner and Michael ZollhÃ¶fer. Pulsar: Efficient sphere-based neural rendering. In *CVPR*, 2021. 2

[26] 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, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 3

[27] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, and Zhaoyang Lv. Neural 3d video synthesis, 2021. 2

[28] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, 2021. 2

[29] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. 2020. 2

[30] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM TOG*, 2021. 3

[31] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering, 2021. 2

[32] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. *arXiv:2008.02268*, 2020. 2

[33] 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. 2020. 2, 3

[34] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. *arXiv preprint arXiv:2104.10078*, 2021. 2

[35] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 2, 3, 5

[36] 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. *arXiv preprint arXiv:2106.13228*, 2021. 2, 3, 5, 6, 7, 8

[37] 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. 3

[38] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: One-shot anatomically consistent facial animation. *International Journal of Computer Vision*, 128(3):698–713, 2020. 3

[39] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *CVPR*, 2021. 2, 3

[40] Gernot Riegler and Vladlen Koltun. Stable view synthesis. In *CVPR*, 2021. 2

[41] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 5

[42] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *ECCV*, 2018. 3

[43] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, 2017. 3

[44] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. 2019. 5, 6, 7

[45] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. 2019. 2

[46] Ayush Tewari, Mohamed Elgharib, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 3

[47] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images, cvpr 2020. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, june 2020. 2, 3

[48] Justus Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering. *ACM Transactions on Graphics (TOG)*, 2019. 2

[49] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *CVPR*, 2021. 2

[50] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *CVPR*, 2021. 2

[51] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *NIPS*, 33, 2020. 2

[52] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv:2010.07492*, 2020. 2

[53] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv:2010.07492*, 2020. 2

[54] 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. 8

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2