Learning Personalized High Quality Volumetric Head Avatars from Monocular RGB Videos

Ziqian Bai[1,2*] Feitong Tan[1] Zeng Huang[1] Kripasindhu Sarkar[1] Danhang Tang[1]
Di Qiu[1] Abhimitra Meka[1] Ruofei Du[1] Mingsong Dou[1] Sergio Orts-Escolano[1]
Rohit Pandey[1] Ping Tan[2] Thabo Beeler[1] Sean Fanello[1] Yinda Zhang[1]
[1] Google [2] Simon Fraser University

Figure 1. Our technique builds a 3D avatar representation of a person using just a single short monocular RGB video (*e.g.*, 1-2 minutes). We leverage a 3DMM to track the user's expressions. By anchoring a neural radiance field to the 3DMM geometry, we generate a volumetric photorealistic 3D avatar that can be rendered with user-defined expression and viewpoint. Note that our method works well on challenging materials, *e.g.*, hair and dramatic expressions. Please see our webpage augmentedperception.github.io/monoavatar for more results.

Abstract

We propose a method to learn a high-quality implicit 3D head avatar from a monocular RGB video captured in the wild. The learnt avatar is driven by a parametric face model to achieve user-controlled facial expressions and head poses. Our hybrid pipeline combines the geometry prior and dynamic tracking of a 3DMM with a neural radiance field to achieve fine-grained control and photorealism. To reduce over-smoothing and improve out-of-model expressions synthesis, we propose to predict local features anchored on the 3DMM geometry. These learnt features are driven by 3DMM deformation and interpolated in 3D space to yield the volumetric radiance at a designated query point. We further show that using a Convolutional Neural Network in the UV space is critical in incorporating spatial context and producing representative local features. Extensive experiments show that we are able to reconstruct high-quality avatars, with more accurate expression-dependent details, good generalization to out-of-training expressions, and quantitatively superior renderings compared to other state-of-the-art approaches.

1. Introduction

Creating a controllable human avatar is a fundamental piece of technology for many downstream applications, such as AR/VR communication [20,31], virtual try-on [37], virtual tourism [13], games [42], and visual effects for movies [12, 18]. Prior art in high-quality avatar generation typically requires extensive hardware configurations (*i.e.*, camera arrays [6, 12, 31], light stages [18, 29], dedicated depth sensors [8]), or laborious manual intervention [1]. Alternatively, reconstructing avatars **from monocular RGB videos** significantly relaxes the dependency on equipment setup and broadens the application scenarios. However, monocular head avatar creation is highly ill-posed due to the dual problems of reconstructing and tracking highly articulated and deformable facial geometry, while model-

*Work was conducted while Ziqian Bai was an intern at Google.

ing sophisticated facial appearance. Traditionally, 3D Morphable Models (3DMM) [7, 24] have been used to model facial geometry and appearance for various applications including avatar generation [9, 16, 22]. However, 3DMMs do not fully capture subject-specific static details and dynamic variations, such as hair, glasses, and expression-dependent high frequency details such as wrinkles, due to the limited capacity of the underlying linear model.

Recent works [3, 15] have incorporated neural radiance fields in combination with 3DMMs for head avatar generation to achieve photorealistic renderings, especially improving challenging areas, such as hair, and adding view-dependent effects, such as reflections on glasses. The pioneering work of NerFACE [15] uses a neural radiance field parameterized by an MLP that is conditioned on 3DMM expression parameters and learnt per-frame latent codes. While they achieve photorealistic renderings, the reliance on an MLP to directly decode from 3DMM parameter space leads to the loss of fine-grain control over geometry and articulation. Alternatively, RigNeRF [3] learns the radiance field in a canonical space by warping the target head geometry using a 3DMM fit, which is further corrected by a learnt dense deformation field parameterized by another MLP. While they demonstrate in-the-wild head pose and expression control, the use of two global MLPs to model canonical appearance and deformations for the full spatial-temporal space leads to a loss of high frequency details, and an overall uncanny appearance of the avatar. Both of these works introduce new capabilities but suffer from lack of detail in both appearance and motion because they attempt to model the avatar's global appearance and deformation with an MLP network.

In this paper, we propose a method to learn a neural head avatar from a monocular RGB video. The avatar can be controlled by an underlying 3DMM model and deliver high-quality rendering of arbitrary facial expressions, head poses, and viewpoints, which retain fine-grained details and accurate articulations. We achieve this by learning to predict expression-dependent spatially local features on the surface of the 3DMM mesh. A radiance field for any given 3D point in the volume is then obtained by interpolating the features from K-nearest neighbor vertices on the deformed 3DMM mesh in target expression, and passing them through a local MLP to infer density and color. The local features and local MLP are trained jointly by supervising the radiance field through standard volumetric rendering on the training sequence [30]. Note that our networks rely on the local features to model appearance and deformation details, and leverages the 3DMM to model only the global geometry.

Learning local features is critical in achieving a high-quality head avatar. To this end, we train an image-to-image translation U-Net that transforms the 3DMM deformations in the UV space to such local features. These UV-space features are then attached to the corresponding vertices of the 3DMM mesh geometry. We show that learning features from such explicit per-vertex local displacement of the 3DMM geometry makes the model retain high-frequency expression-dependent details and also generalizes better to out-of-training expressions, presumably because of the spatial context between nearby vertices incorporated by the convolutional neural network (CNN). An alternative approach is to feed the 3DMM parameters directly into a CNN decoder running on the UV space. However, we found this produces severe artifacts on out-of-training expressions, particularly given a limited amount of training data, *e.g.* for a lightweight, 1-minute data collection procedure during the avatar generation process.

In summary, our contributions are as follows: we propose a neural head avatar representation based on a 3DMM-anchored neural radiance field, which can model complex expression-dependent variations, but requires only monocular RGB videos for training. We show that a convolutional neural network running on per-vertex displacement in UV space is effective in learning local expression-dependent features, and delivers favorable training stability and generalization to out-of-training expressions. Experiments on real-world datasets show that our model provides competitive controllability and generates sharper and detail enriched rendering compared to state-of-the-art approaches.

2. Related Works

Building photorealistic representations of humans has been widely researched in the past few decades. Here, we mainly discuss prior art in head avatar and refer readers to the state-of-the-art surveys [14, 55] for a comprehensive literature review.

Monocular Explicit Surface Head (Face) Avatars. Traditionally, a typical approach to create head (or face) avatars from monocular RGB videos is using a 3D Morphable Models (3DMM) as the foundation and adding personalized representations, such as corrected blendshapes [16, 22], detail texture maps [16, 22], image-based representations [9], and secondary components [21]. Early works use various optimizations to obtain the personalized representations from monocular data, including analysis-by-synthesis [14, 16, 21, 55] as well as shape-from-shading [16, 22]. Recent approaches replace optimizations by regressions with Deep Neural Networks (DNNs) [10, 40, 48], or integrate optimizations with deep learning components [4, 5]. More recent methods leverage neural textures [17] to generate photorealistic appearances.

The main drawback of these methods is that they rely on explicit meshes with a fixed topology, making it hard to handle out-of-model details such as hair and accessories like glasses and apparels. In contrast, our hybrid method combines geometric priors with implicit representations, lead-

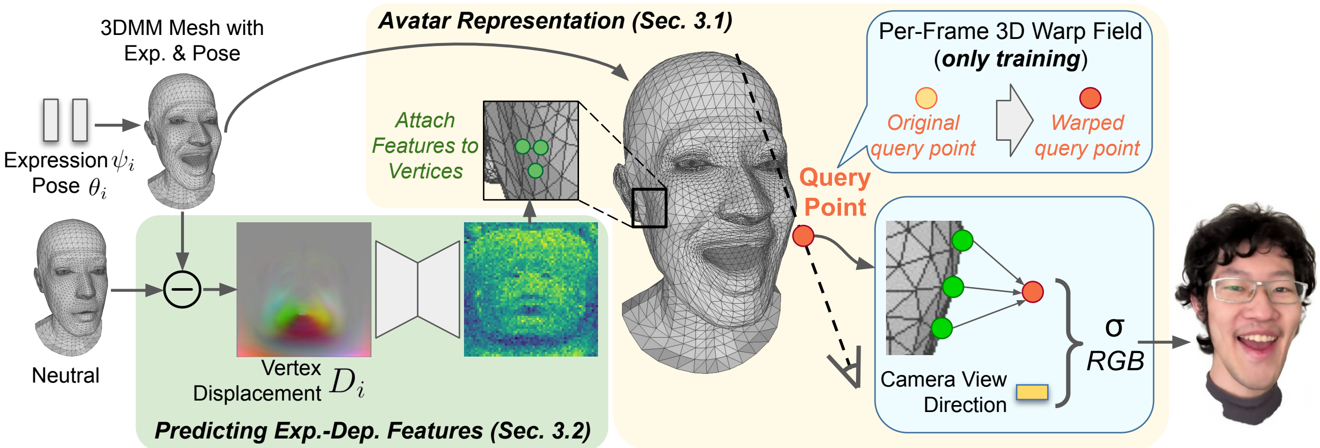

Figure 2. Overview of our pipeline. The core of our method is the Avatar Representation (Sec. 3.1. Shown as the yellow area) based on a 3DMM-anchored neural radiance field (NeRF), which are decoded from local features attached on the 3DMM vertices. Then, we use volumetric rendering to compute the output image. To predict the vertex-attached features (Sec. 3.2. Shown as the green area), we first compute the vertex displacements from the 3DMM expression and pose, then process the displacements in UV space with Convolutional Neural Networks (CNNs), and sample the obtained features back to mesh vertices.

ing to a significantly larger representation capacity.

Monocular Implicit Head Avatars. Recent work proposes to extend 3DMM with implicit 3D representations. NerFACE [15] introduces a dynamic neural radiance field (NeRF) conditioned on 3DMM expression codes which can render a view-consistent avatar with volumetric rendering. Since NerFACE directly inputs the 3DMM expression codes into MLPs without using any shape or spatial information from 3DMM, their model is quite under-constrained for monocular reconstruction, and suffers from severe artifacts for data with challenging expressions. RigNeRF [3] uses 3DMM derived warping field to deform the camera space into a canonical space, and defines a canonical NeRF conditioned on 3DMM codes. However, their model uses a dense MLP-based architecture to memorize the appearance and deformation for the full head, leading to oversmooth results due to limited network capacity. IMAvatar [53] learns personalized implicit fields of blendshapes, pose correctives, and skinning weights, then formulates the avatar with linear summation of blendshapes followed by linear blend skinning. However, their linear formulation limits the amount of expression deformations.

Geometry Anchored Implicit 3D Representation. Sparse local feature embedding attached on geometry has been demonstrated to be effective in improving the rendering quality of neural radiance field [26–28, 34, 54]. It also naturally supports neural radiance field editing since the modification on the geometry can be directly propagated to the rendering [11], which makes them a favorable representation to support the controllability for human avatar. We adapt this representation to head avatar and incorporate head specific priors. Differently from prior art, we leverage a CNN in UV space to learn local, per-vertex features that are expression-dependent, improving generalization of out-of-train expressions.

2D-based Head Avatars. There are numerous approaches that synthesize the head (or face) relying on 2D (explicit/implicit) representations, including 2D facial landmarks [43, 49, 50] and 2D warping fields [38, 39, 46]. Landmark-based avatar models [43, 49, 50] synthesize the face conditioned to the facial landmarks extracted with a, usually pre-trained, landmark detector. Specifically, an encoder is applied to extract an identity embedding from a reference image, a decoder is adapted by the identity code to animate the reference face with landmarks from the driving videos. X2Face [46] is the first approach to animate human heads by learning a dense warping field and producing the output video via image warping. MonkeyNet [38] and First Order Motion Model (FOMM) further propose to infer motion fields with self-learned keypoints, which significantly improves motion prediction and synthesizes higher quality renderings of heads. While most aforementioned methods can produce photorealistic results, they are not able to maintain geometry and multiview consistency due to their inherent 2D representation.

3. Method

Given a monocular RGB video containing M frames $\{\boldsymbol{I}_1, \boldsymbol{I}_2, ..., \boldsymbol{I}_M\}$, our method reconstructs a head avatar representation that can be rendered under arbitrary facial expressions, head poses, and camera viewpoints. We first preprocess the video to remove the background [19,32] and obtain camera and 3DMM parameters for each frame. More specifically, we use FLAME [24] as the 3DMM and denote the fitted face with shape $\boldsymbol{\beta}$, expressions $\boldsymbol{\psi}_i$, poses $\boldsymbol{\theta}_i$ (*i.e.*,

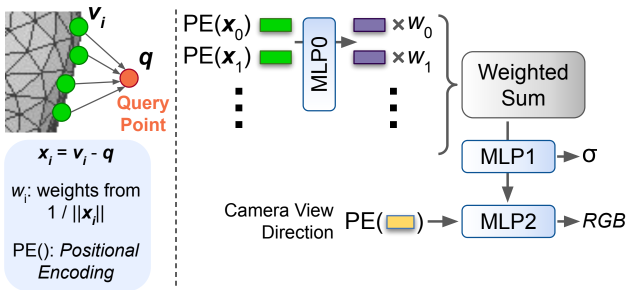

Figure 3. Illustration of Avatar Representation (Sec. 3.1). Given a query point, we find its k-Nearest-Neighbor (k-NN) vertices from the 3DMM. Then, we decode these vertices and features into a density and color with respect to the input camera view direction, via Multi-Layer-Perceptrons (MLPs) interleaved with inverse-distance based weighted sum.

neck, jaw, and eyes), where i is the frame index, with which a head mesh can be obtained via $\boldsymbol{V}_i(\boldsymbol{\beta}, \boldsymbol{\psi}_i, \boldsymbol{\theta}_i)$. Since $\boldsymbol{\beta}$ is fixed and does not depend on the pose or expressions for a given user, we omit it in the following sections for brevity.

An overview of our framework is shown in Fig. 2. We adopt the 3DMM-anchored neural radiance field (NeRF) as the core representation for our head avatar (Sec. 3.1), where local features are attached to the vertices of the deformable 3DMM mesh. During the inference, we first deform the 3DMM mesh based on the target configuration $\boldsymbol{V}_t = (\boldsymbol{\psi}_t, \boldsymbol{\theta}_t)$. Then, for an arbitrary 3D query point, we aggregate the features from neighboring vertices on $\boldsymbol{V}_t$ to estimate the local density and color by Multi-Layer-Perceptrons (MLPs), which are then integrated in the volumetric rendering formulation to generate the color image. To learn local features, we train CNN-based networks in the UV space to incorporate spatial context (Sec. 3.2). Our model is trained end-to-end with RGB supervisions (Sec. 3.3).

3.1. Avatar Representation

An ideal representation for a head avatar should have the following properties: 1) Provides intuitive control to achieve the desired expression and head pose; 2) Requires a moderate amount of training data, *e.g.*, a short monocular video; 3) Produces expression-dependent rendering details; 4) Generalizes reasonably well to unseen expressions.

To this end, we propose the 3DMM-anchored neural radiance field (NeRF) as shown in Fig. 3. Inspired by local feature based neural radiance field [34, 47], we attach feature vectors $\boldsymbol{z}^j$ on each 3DMM vertex $\boldsymbol{v}_i^j$ to encode the local radiance fields that can be decoded with MLPs, where i denotes frame index and j denotes vertex index. In this way, the radiance field can be deformed according to vertex locations, hence can be intuitively controlled by the 3DMM expression and pose $(\boldsymbol{\psi}_i, \boldsymbol{\theta}_i)$. In addition, the 3DMM fitting on each frame provides a rough tracking across deformable face geometries, such that all the frames can contribute into the learning of a unified set of local per-vertex features. We will discuss model capacity and generalization in Sec. 3.2.

To decode the vertex features $\{\boldsymbol{z}^j\}$ into the radiance field for the frame i, given a 3D query point $\boldsymbol{q}$, we first find its k-Nearest-Neighbor (k-NN) vertices from the 3DMM mesh $\{\boldsymbol{v}_i^j\}_{j\in\mathcal{N}_k^{\boldsymbol{q}}}$ with attached features $\{\boldsymbol{z}^j\}_{j\in\mathcal{N}_k^{\boldsymbol{q}}}$. Then, we use two MLPs $\mathcal{F}_0$ and $\mathcal{F}_1$ with inverse-distance based weighted sum to decode local color and density. Formally,

$$
\begin{aligned}
\hat{\boldsymbol{z}}_i^j &= \mathcal{F}_0(\boldsymbol{v}_i^j - \boldsymbol{q}, \boldsymbol{z}^j) \\
\hat{\boldsymbol{z}}_i &= \textstyle\sum_j w^j \hat{\boldsymbol{z}}_i^j \\
\boldsymbol{c}_i(\boldsymbol{q}, \boldsymbol{d}_i), \sigma_i(\boldsymbol{q}) &= \mathcal{F}_1(\hat{\boldsymbol{z}}_i, \boldsymbol{d}_i),
\end{aligned}
\tag{1}
$$

where $w^j = \frac{d^j}{\sum_k d^k}$, $d^j = \frac{1}{\|\boldsymbol{v}_i^j - \boldsymbol{q}\|_2}$ with $j \in \mathcal{N}_k^{\boldsymbol{q}}$, and $\boldsymbol{d}_i$ denotes the camera view direction. Finally, we render the output image with volumetric rendering formulation as in vanilla NeRF [30] given the camera ray $\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{d}$:

$$
\boldsymbol{C}_i(\boldsymbol{r}) = \int_{t_n}^{t_f} T(t)\sigma_i(\boldsymbol{r}(t))\boldsymbol{c}_i(\boldsymbol{r}(t), \boldsymbol{d})\, \mathrm{d}t, \tag{2}
$$
$$
\text{where } T(t) = \exp\left(-\int_{t_n}^{t} \sigma_i(\boldsymbol{r}(s))\, \mathrm{d}s\right)
$$

To reduce misalignments caused by per-frame contents that cannot be captured by 3DMM (*e.g.*, 3DMM fitting errors), we additionally learn error-correction warping fields during training inspired from prior works on deformable NeRF [23, 45]. More specifically, we input the original query point and a per-frame latent code $\boldsymbol{e}_i$, which is randomly initialized and optimized during the training, into the error-correction MLPs $\mathcal{F}_\mathcal{E}$ to predict a rigid transformation, and apply it to the query point. The transformation is denoted as $\boldsymbol{q}' = \mathcal{T}_i(\boldsymbol{q}) = \mathcal{F}_\mathcal{E}(\boldsymbol{q}, \boldsymbol{e}_i)$. Then we use the warped query point $\boldsymbol{q}'$ to decode the color and density. Note that this warping field is disabled during testing. Please refer to the supplementary for detailed formulations of the warping field.

3.2. Predicting Expression-Dependent Features

While the proposed avatar representation (Sec. 3.1) enables intuitive controllability and convenience in learning, it still has limited capability for modeling complex expression-dependent variations due to the use of frame-shared vertex features $\{\boldsymbol{z}^j\}$.

To overcome this, we propose to predict the dynamic vertex features $\{\boldsymbol{z}_i^j\}$ conditioned on the 3DMM expression and pose $(\boldsymbol{\psi}_i, \boldsymbol{\theta}_i)$. A common practice for NeRF-based methods is to use MLP-based architectures for dynamic feature prediction [3, 15, 53, 54]. However, we find that this leads to blurry rendering results, presumably because of the limited model capacity due to the lack of spatial context (*i.e.*,

	Subject0 LPIPS / SSIM / PSNR	*Subject1* LPIPS / SSIM / PSNR	*Subject2* LPIPS / SSIM / PSNR	*Subject3* LPIPS / SSIM / PSNR	*Subject4* LPIPS / SSIM / PSNR
TPSMM [52]	0.192 / 0.852 / 22.60	0.205 / 0.830 / **16.38**	0.216 / 0.782 / 18.40	0.222 / 0.799 / 20.28	0.156 / 0.913 / 21.29
FOMM [39]	0.171 / 0.841 / **22.93**	0.179 / 0.827 / 16.02	0.202 / 0.777 / 18.98	0.186 / 0.798 / 22.28	0.122 / 0.915 / 23.94
NHA [17]	0.165 / 0.836 / 20.20	0.166 / 0.840 / 15.48	0.178 / 0.809 / 17.99	**0.153** / 0.798 / 21.31	0.091 / 0.926 / 23.78
IMAvatar [53]	0.207 / 0.852 / 21.26	0.187 / 0.848 / 15.98	0.265 / 0.729 / 15.80	0.214 / 0.782 / 20.37	0.142 / 0.897 / 20.63
NerFACE [15]	0.205 / 0.817 / 20.06	0.182 / 0.833 / 15.78	0.188 / 0.793 / 19.41	0.229 / 0.747 / 18.16	0.093 / 0.938 / 25.57
Ours-D	**0.144 / 0.864** / 21.92	**0.152 / 0.855** / 16.23	**0.141 / 0.841 / 20.42**	0.156 / **0.833 / 23.05**	**0.075 / 0.944 / 25.71**

Table 1. Quantitative Comparison with state-of-the-art (SOTA) approaches. *Subject4* is the data from NerFACE [15], while other subjects are from our dataset. Our method achieves superior results than prior SOTAs.

each vertex does not know the feature of its neighboring vertices). Based on this intuition, we propose to process the 3DMM expression and pose ($\boldsymbol{\psi}_i, \boldsymbol{\theta}_i$) with CNNs in the texture atlas space (UV space) to provide local spatial context.

Specifically, we design two variations of CNN-based architecture to learn expression-dependent vertex features $\{\boldsymbol{z}_i^j\}$. The first variant, denoted as *Ours-C*, trains a decoder consisting of transposed convolutional blocks to predict the feature map in UV space directly from 1-D codes of 3DMM expression and pose. We empirically find that such a model is effective in improving the overall rendering quality, however tends to fail and produce severe artifacts on out-of-training expressions (See discussions in Sec. 4.4). We then propose the second variant, denoted as *Ours-D*, that uses the 3D deformation of 3DMM in UV space as the input for feature prediction, and observe that resulting avatar models are more resilient to stretchy and unseen expressions. Specifically, we first compute the vertex displacements using 3DMM expression and pose as $\boldsymbol{D}_i = \boldsymbol{V}_i(\boldsymbol{\psi}_i, \boldsymbol{\theta}_i) - \boldsymbol{V}_{neutral}(\boldsymbol{0}, \boldsymbol{0})$. We then rasterize the vertex displacements $\boldsymbol{D}_i$ into UV space, and process it with a U-Net $\mathcal{F}_{\mathcal{D}}$. Finally, the output UV feature map is sampled back to mesh vertices $\boldsymbol{V}_i$, serving as the dynamic vertex features $\{\boldsymbol{z}_i^j\}$ (*i.e.*, the expression-dependent version of frame-shared vertex features $\{\boldsymbol{z}^j\}$ described in Sec. 3.1).

3.3. Training Schema

Our model is trained on monocular RGB videos mainly with the photometric loss, where we penalize the l_2 distance between the rendering and the ground truth images. Formally $\mathcal{L}_{rgb} = \sum_i \sum_{\boldsymbol{r}} \|\boldsymbol{C}_i(\boldsymbol{r}) - \boldsymbol{I}_i(\boldsymbol{r})\|_2$, where $\boldsymbol{r}$ denotes the camera ray of each pixel and i denotes frame index. To regularize the learning of error-correction warping field $\mathcal{T}(\boldsymbol{q})$, we adopt a elastic loss $\mathcal{L}_{elastic}$ similar to Nerfies [33], and a magnitude loss defined as $\mathcal{L}_{mag} = \sum_{\boldsymbol{q}} \|\boldsymbol{q} - \mathcal{T}(\boldsymbol{q})\|_2^2$. The total loss is defined as:

$$\mathcal{L} = \mathcal{L}_{rgb} + \lambda_{elastic}\mathcal{L}_{elastic} + \lambda_{mag}\mathcal{L}_{mag}, \quad (3)$$

where we set $\lambda_{elastic} = 10^{-4}$ at the beginning and decay to 10^{-5} after 155k iterations, and $\lambda_{mag} = 10^{-2}$. Please see supplementary for more details.

4. Experiments

We train and evaluate our method on casually captured monocular RGB videos (Sec. 4.1) and show that our method achieves superior rendering quality than prior state-of-the-art monocular RGB head avatars (Sec. 4.2). Then we verify our key observations on architectural choices to design a good avatar model, in terms of rendering quality and expression robustness (Sec. 4.4).

4.1. Datasets and Metrics

Datasets. Following the prior art [15], we captured monocular RGB videos of various subjects with smartphones or webcams for training and evaluation. Each video is 1-2 minutes long (around 1.5k-2k frames at 30 FPS) with the first 1000-1500 frames as training data and the rest frames for evaluation if not otherwise specified. For the training clip, the subjects are asked to first keep a neutral expression and rotate their heads, then perform different expressions during the head rotation, with extreme expressions included. For the testing clip, the subjects are asked to perform freely without any constraints. To demonstrate that our method also works for common talking head videos, we also include a video from NerFACE [15], which has significantly less variability in expressions when compared to our capture protocol. We mask out background [32] for each video, and obtain 3DMM and camera parameters with 3DMM fitting similar to that in NHA [17]. Please refer to the supplementary for examples of data. Please note that we collect relatively shorter training videos compared to related work. This favors a better user experience while still synthesizing high-quality avatars with more personalized and detailed expressions.

Metrics. Following the prior art [15], we use the following standard image quality metrics for quantitative evaluations: the Learned Perceptual Image Patch Similarity (LPIPS) [51], Structure Similarity Index (SSIM) [44], and Peak Signal-to-Noise Ratio (PSNR).

Figure 4. Qualitative Comparison to prior state-of-the-art monocular head avatars. Note how our approach more faithfully reconstructs the ground truth expressions while preserving most of the high frequency details.

4.2. Comparisons with State-of-Art

We compare our method with five state-of-the-art methods of different types with publicly available implementations from the authors: NerFACE [15], IMAvatar [53], NHA [17], FOMM [39], and TPSMM [52]. For subject-specific methods (*i.e.* Ours, NerFACE [15], IMAvatar [53], and NHA [17]), we train the avatar for each subject separately with training frames of each video, then drive and render the trained avatar with 3DMM and camera parameters of testing frames. For few-shot methods (*i.e.*, FOMM [39], and TPSMM [52]), we use the first frame of each video, which shows a frontal head, as the source image, then use testing frames as driving images in self-reenactment manner. Finally, the generated images of each method are compared with ground truth testing frames quantitatively (See Tab. 1) and qualitatively (See Fig. 4).

As shown in Fig. 4, FOMM [39] and TPSMM [52] struggle with large head rotations and fail to produce 3D consistent results. NHA [17] uses explicit mesh surface with neural textures. The fix mesh topology makes it hard to handle challenging hair in *Subject 3*. Also, NHA uses linear 3DMM blendshapes, making it hard to capture complex expressions (*e.g.*, *Subject 1 & 3*) and wrinkles (*e.g.*, *Subject 4*). IMAvatar [53] uses an occupancy field to model geometry, making it hard to handle challenging hair. Despite learning a personalized blendshape field, the linear formulation still limits their model capacity of handling complex expressions. In addition, we observed training instability of IMAvatar on the captured data, converging to oversmooth results. NerFACE [15] works relatively well on data with easy expressions (*i.e.*, common talking heads in *Subject 4*), but struggles on our challenging data and estimates incorrect geometry (*e.g.*, distorted head for *Subject 3*) and blurry rendering. In contrast, our method gives superior results on all the aspects discussed above. Tab. 1 further quantitatively confirms the good rendering quality of our method.

4.3. Driving the Avatar

After training, the learned avatar model can be driven by the same subject under different capture conditions, *e.g.* hair

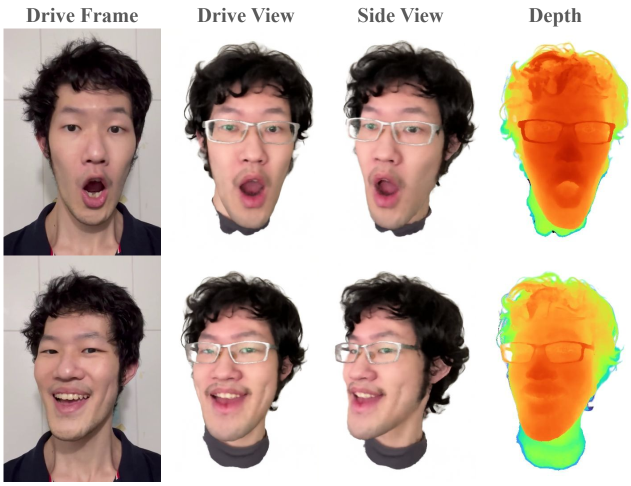

Figure 5. Results on driving the learned avatar by the same subject under different capture conditions. Our method produces faithful expressions, multi-view consistent rendering, and good geometry.

Figure 6. Results on expressions out of the training distribution with different amount of training data. *Ours-D* more robustly handles unseen expressions and degrades less with fewer training data.

style, illumination, glasses. We show the driving results in Fig. 5. Our avatar faithfully reproduces the expressions of the driving frame, while also achieving multi-view consistent renderings and generating high quality geometry.

4.4. Ablation Study: Expression Features

Learning good local features is crucial for improving model capacity and capturing high frequency details, without losing the regularization from 3DMM (Sec. 3.2). We investigate and compare three alternative approaches and our two model variations for feature learning:

Static Features. We use frame-shared (hence expression-shared) vertex features $\{z^j\}$. As shown in Fig. 7, the expression-shared features struggle to capture strong expression-dependent variations, leading to incorrect geometry (*e.g.*, incorrect cheek silhouette), blurry details (*e.g.*, teeth), and inferior LPIPS scores in Tab. 2.

3DMM Codes. We extend *Static Features* by directly concatenating 3DMM expression and pose codes $(\psi_i, \boldsymbol{\theta}_i)$ to the vertex features $\{z^j\}$. This allows the model to have more capacity to capture expression-dependent variations. Although median-level expression characteristics are recognizable (*e.g.*, cheek silhouette in Fig. 7), the results are even

	Subject0	*Subject1*	*Subject2*	*Subject3*
Full Training Data				
Static Features	0.1559	0.1586	0.1552	0.1688
3DMM Codes	0.1599	0.1620	0.1746	0.1738
3DMM Codes MLP	0.1568	0.1551	0.1505	0.1686
Ours-C	**0.1417**	**0.1457**	**0.1383**	**0.1550**
Ours-D	<u>0.1439</u>	<u>0.1523</u>	<u>0.1415</u>	<u>0.1559</u>
First 50% Training Data				
Ours-C	0.2038	0.1483	0.1580	0.1566
	+0.0621	+0.0026	+0.0197	+0.0016
Ours-D	0.1711	0.1511	0.1516	0.1558
	+0.0272	-0.0012	+0.0101	-0.0001

Table 2. Quantitative ablations with baselines in LPIPS scores (lower is better). **Bold** denotes the best while <u>underline</u> denotes the second best.

blurrier for components less depend on expressions such as hairs, possibly due to the limited model capacity of shallow MLPs without spatial context.

3DMM Codes MLP. To further increase the model capacity, we use a more sophisticated MLP architecture inspired from Zheng et al. [54], which is a MLP-based conditional Variational AutoEncoder (cVAE), to predict the expression-dependent vertex features $\{z_i^j\}$ from the 3DMM expression and pose codes $(\psi_i, \boldsymbol{\theta}_i)$. More specifically, the cVAE is conditioned on 3DMM codes and vertex coordinates on neutral face mesh $\boldsymbol{V}_{neutral}$. Then, the cVAE encodes the frame index into a latent code and decodes it into the vertex features $\{z_i^j\}$. Although the overall sharpness of results is improved (*e.g.*, glasses and hair in Fig. 7), this model still produces blurry details such as teeth, eyes, and wrinkles. Tab. 2 further confirms that this more sophisticated MLP architecture still produces inferior results compared to CNN-based methods *Ours-C* and *Ours-D*.

Ours-C *vs.* Ours-D. By replacing MLPs with CNNs in UV space (Please refer to Sec. 3.2 for more details), both variations of *Ours* achieve superior rendering quality than MLP-based baselines, as shown in Fig. 7. Moreover, we find that *Ours-D* is more robust to expressions that are outside the training distribution compared to *Ours-C* (See Fig. 6). To further investigate the model robustness to out-of-training expressions, we train *Ours-C* and *Ours-D* with the first 50% of training frames, to simulate the scenario with less expression coverage during training. As shown in Tab. 2 and Fig. 6, *Ours-D* degrades less than *Ours-C*, indicating enhanced robustness.

4.5. Robustness to Expression Extrapolation

We qualitatively test our method on expression extrapolation setting by artificially manipulating the expression code for a given test frame. In particular, we double the value of the expression code and compare the results with

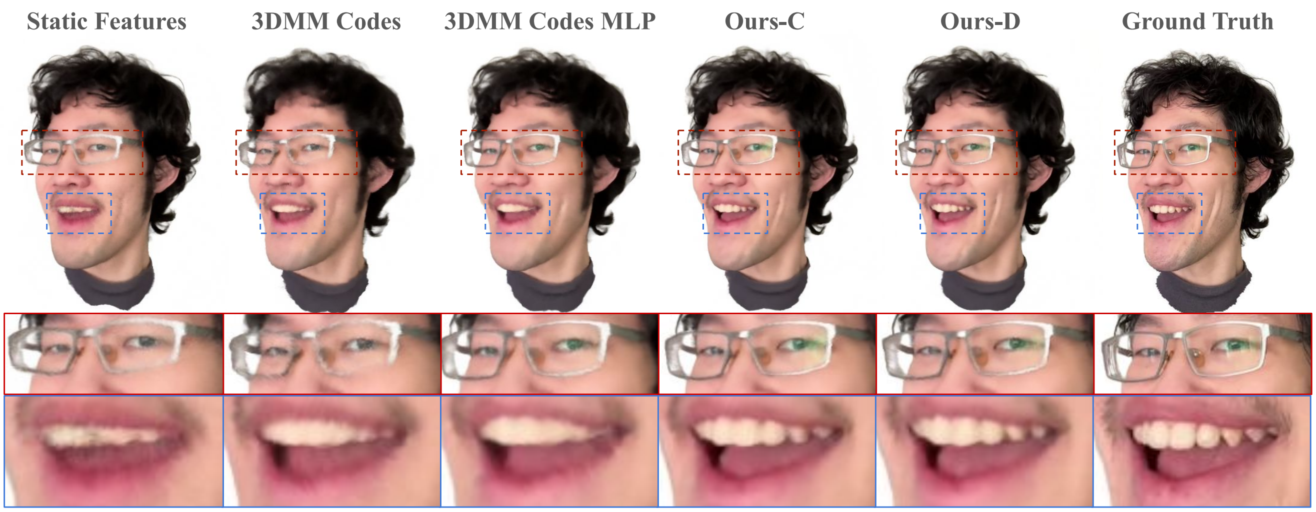

Figure 7. Comparison between different designs for local vertex feature learning. See Sec. 4.4 for more details. "Static feature" struggles on capturing personalized expressions. "3DMM Codes" improves the personalization but suffers from overall blurriness. "3DMM Codes MLP" further improves the sharpness, but still cannot present the details. Overall, our convolution-based methods lead to superior renderings on areas such as cheek silhouette, glasses frames and reflections, and teeth.

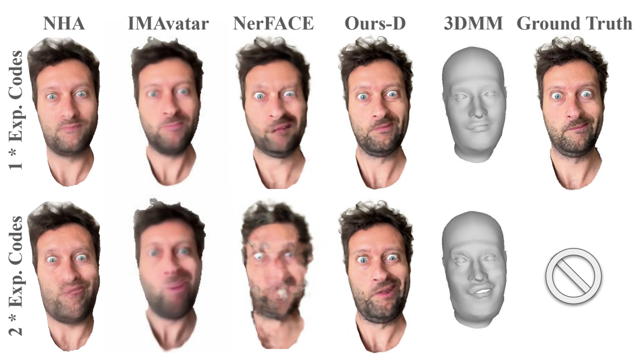

Figure 8. Expression extrapolation results and comparisons. Compared to other work, our method performs well on extreme and out-of-training expression.

prior works. As shown in Fig. 8, NHA [17] gives reasonable expression extrapolation results for both geometry and appearance. However, it cannot faithfully capture the out-of-3DMM details when compared to the ground truth. IMAvatar [53] gives reasonable geometry for extrapolation, but struggles to produce sharp extrapolated appearances. NerFACE [15] fails in producing reasonable renderings. In contrast, our method not only faithfully captures out-of-3DMM details, but also generalizes to extrapolated expressions.

5. Discussion

In this work, we presented a framework to learn high-quality controllable 3D head avatars from monocular RGB videos. The core of our method is a 3DMM-anchored neural radiance field decoded from features attached on 3DMM vertices. The vertex features can be either predicted from 3DMM expression and pose codes, or vertex displacements, via CNNs in UV space, where the former favors quality and the latter favors robustness. We experimentally demonstrate that it is possible to learn high-quality avatars with faithful, highly non-rigid expression deformations purely from monocular RGB videos, leading to a superior rendering quality of our method compared to other approaches.

Limitations. Compared to the state-of-the-art approaches, our proposed framework learns portrait avatars with superior rendering quality. Nevertheless, our method still inherits the disadvantages of NeRF [30] on time-consuming subject-specific training and slow rendering. Our method relies on the expression space of a 3DMM, thus cannot capture components that are completely missing in the 3DMM, such as the tongue. Moreover, extending our method to include the upper body or integrating into full body avatars are also interesting future directions.

Ethical Considerations. The rapid progress on avatar synthesis facilitates numerous downstream applications, but also raises concerns on ethical misuses. On the synthesis side, it would be ideal to actively use watermarking and avoid driving avatars with different identities. On the viewing side, forgery detection [2, 25, 35, 36, 41] is an active research field with promising progress. However, it is still hard to guarantee reliable detection of fake visual materials at the current stage. Encouraging the use of cryptographical signatures may also be a potential solution to ensure the authenticity of visual material.

References

[1] MetaHuman - Unreal Engine. https://www.unrealengine.com/en-US/metahuman, 2. Accessed: 2022-10-17. 1

[2] Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. Detecting deep-fake videos from appearance and behavior. In *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2020. 8

[3] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. RigNeRF: Fully Controllable Neural 3D Portraits. pages 20364–20373, 2022. 2, 3, 4

[4] Ziqian Bai, Zhaopeng Cui, Xiaoming Liu, and Ping Tan. Riggable 3D Face Reconstruction via In-Network Optimization. pages 6216–6225, June 2021. 2

[5] Ziqian Bai, Zhaopeng Cui, Jamal Ahmed Rahim, Xiaoming Liu, and Ping Tan. Deep Facial Non-Rigid Multi-View Stereo. pages 5850–5860, 2020. 2

[6] Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W Sumner, and Markus Gross. High-Quality Passive Facial Performance Capture Using Anchor Frames. In *ACM SIGGRAPH 2011 Papers*, pages 1–10. 2011. 1

[7] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pages 187–194, 1999. 2

[8] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason Saragih. Authentic Volumetric Avatars from a Phone Scan. *ACM Transactions on Graphics (TOG)*, 41(4), 2022. 1

[9] Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. Real-Time Facial Animation With Image-Based Dynamic Avatars. 35(4), 2016. 2

[10] Bindita Chaudhuri, Noranart Vesdapunt, Linda Shapiro, and Baoyuan Wang. Personalized Face Modeling for Improved Face Reconstruction and Motion Retargeting. In *ECCV 2020: 16th European Conference on Computer Vision*, pages 142–160, 2020. 2

[11] Chong Bao and Bangbang Yang, Zeng Junyi, Bao Hujun, Zhang Yinda, Cui Zhaopeng, and Zhang Guofeng. Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[12] Ruofei Du, Ming Chuang, Wayne Chang, Hugues Hoppe, and Amitabh Varshney. Montage4D: Real-time Seamless Fusion and Stylization of Multiview Video Textures. *Journal of Computer Graphics Techniques*, 8(1):1–34, Jan. 2019. 1

[13] Ruofei Du, David Li, and Amitabh Varshney. Geollery: A Mixed Reality Social Media Platform. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, number 685 in CHI. ACM, May 2019. 1

[14] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3D Morphable Face Models—Past, Present, and Future. 39(5):1–38, 2020. 2

[15] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. pages 8649–8658, 2021. 2, 3, 4, 5, 6, 8

[16] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of Personalized 3D Face Rigs From Monocular Video. *ACM Transactions on Graphics (TOG)*, 35(3):28, 2016. 2

[17] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural Head Avatars From Monocular RGB Videos. pages 18653–18664, 2022. 2, 5, 6, 8

[18] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul Debevec, and Shahram Izadi. The Relightables: Volumetric Performance Capture of Humans With Realistic Relighting. *ACM Transactions on Graphics*, Nov. 2019. 1

[19] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul Debevec, and Shahram Izadi. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Trans. Graph.*, 38(6), nov 2019. 3

[20] Zhenyi He, Ruofei Du, and Ken Perlin. CollaboVR: A Reconfigurable Framework for Multi-user to Communicate in Virtual Reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality*, ISMAR, pages 542–554. IEEE, Nov. 2020. 1

[21] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar Digitization From a Single Image for Real-Time Rendering. 36(6):1–14, 2017. 2

[22] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. Dynamic 3D Avatar Creation From Hand-Held Video Input. 34(4):1–14, 2015. 2

[23] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. NeuMan: Neural Human Radiance Field From a Single Video. 2022. 4

[24] 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):194:1–194:17, 2017. 2, 3

[25] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020. 8

[26] Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. Learning implicit templates for point-based clothed human modeling. In *ECCV*, 2022. 3

[27] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020. 3

[28] 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 Transactions on Graphics (TOG)*, 40(6):1–16, 2021. 3

[29] Abhimitra Meka, Rohit Pandey, Christian Haene, Sergio Orts-Escolano, Peter Barnum, Philip David-Son, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, et al. Deep Relightable Textures: Volumetric Performance Capture With Neural Rendering. *ACM Transactions on Graphics (TOG)*, 39(6):1–21, 2020. 1

[30] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2, 4, 8

[31] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. Holoportation: Virtual 3D Teleportation in Real-Time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST)*. ACM, Oct. 2016. 1

[32] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4):1–21, 2021. 3, 5

[33] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable Neural Radiance Fields. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5865–5874, 2021. 5

[34] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural Body: Implicit Neural Representations With Structured Latent Codes for Novel View Synthesis of Dynamic Humans. pages 9054–9063, 2021. 3, 4

[35] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018. 8

[36] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019. 8

[37] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019. 1

[38] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019. 3

[39] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First Order Motion Model for Image Animation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. 3, 5, 6

[40] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. FML: Face Model Learning From Videos. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10812–10822, 2019. 2

[41] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020. 8

[42] Zach Waggoner. *My Avatar, My Self: Identity in Video Role-Playing Games*. McFarland, 2009. 1

[43] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-Shot Video-to-Video Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. 13(4):600–612, 2004. 5

[45] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022. 4

[46] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on Computer Vision (ECCV)*, pages 670–686, 2018. 3

[47] Bangbang Yang, Chong Bao, Junyi Zeng, Hujun Bao, Yinda Zhang, Zhaopeng Cui, and Guofeng Zhang. Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In *European Conference on Computer Vision*, pages 597–614. Springer, 2022. 4

[48] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. FaceScape: A Large-Scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction. pages 601–610, 2020. 2

[49] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference on Computer Vision*, pages 524–540. Springer, 2020. 3

[50] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9459–9468, 2019. 3

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features As a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 5

[52] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022. 5, 6

[53] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13545–13555, 2022. 3, 4, 5, 6, 8

[54] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured Local Radiance Fields for Human Avatar Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15893–15903, 2022. 3, 4, 7

[55] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications. volume 37, pages 523–550, 2018. 2