MonoNPHM: Dynamic Head Reconstruction from Monocular Videos

Simon Giebenhain[1] Tobias Kirschstein[1] Markos Georgopoulos[2*] Martin Rünz[2]
Lourdes Agapito[3] Matthias Nießner[1]

[1]Technical University of Munich [2]Synthesia [3]University College London

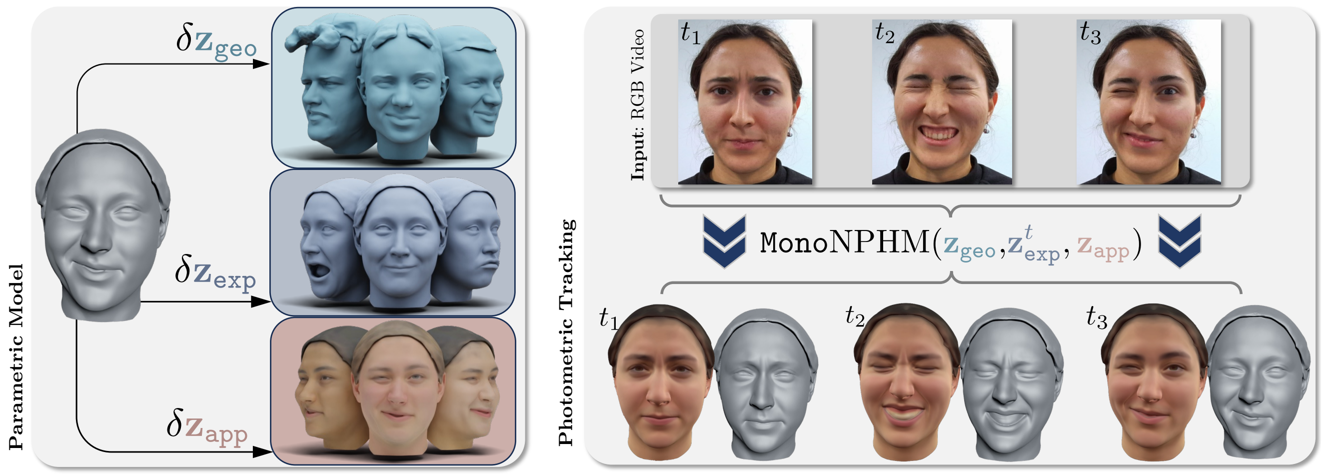


Figure 1. We present MonoNPHM, a neural-field-based parametric head model (left), for dynamic 3D head reconstruction in monocular videos (right). On the left we demonstrate the effect on a reconstructed human head, by individually varying shape (top box), expression (middle box) and appearance (bottom box). The right-hand side illustrates three input RGB frames (top row), and our reconstructed geometry (bottom row). We also show the reconstructed appearance under estimated lighting conditions, which is the basis of our reconstruction.

Abstract

We present Monocular Neural Parametric Head Models (MonoNPHM) for dynamic 3D head reconstructions from monocular RGB videos. To this end, we propose a latent appearance space that parameterizes a texture field on top of a neural parametric model. We constrain predicted color values to be correlated with the underlying geometry such that gradients from RGB effectively influence latent geometry codes during inverse rendering. To increase the representational capacity of our expression space, we augment our backward deformation field with hyper-dimensions, thus improving color and geometry representation in topologically challenging expressions. Using MonoNPHM as a learned prior, we approach the task of 3D head reconstruction using signed distance field based volumetric rendering. By numerically inverting our backward deformation field, we incorporated a landmark loss using facial anchor points that are closely tied to our canonical geometry representation. To evaluate the task of dynamic face reconstruction from monocular RGB videos we record 20 challenging Kinect sequences under casual conditions. MonoNPHM outperforms all baselines with a significant margin, and makes an important step towards easily accessible neural parametric face models through RGB tracking.

1. Introduction

Tracking, animation, and reconstruction of human faces and heads under complex facial movements are fundamental problems in many applications such as computer games, movie production, telecommunication, and AR/VR settings. In particular, obtaining high-fidelity 3D head reconstructions from monocular input videos is a common scenario in many practical settings, e.g., when only a commodity webcam is available.

Recovering the 3D head geometry throughout a monocular RGB video, however, is inherently under-constrained. The task is further complicated in the presence of depth ambiguity, complex facial movements, and strong lighting and shadow effects. Therefore, to disambiguate the 3D scene

Website: https://simongiebenhain.github.io/MonoNPHM
* Work done while MG was at Synthesia.

dynamics, it is common to introduce a set of assumptions about plausible facial structure, expressions, and appearance, often in the form of a model prior.

To regularize this otherwise heavily under-constrained problem, the most widely adopted model-prior, are 3D morphable models (3DMMs) [8], which capture shape, expression, and appearance variations through the use of principal component analysis (PCA) over a dataset of 3D scans that have been registered with a template mesh. Therefore, their expressiveness is often limited by the underlying (multi-)linear statistical model, the resolution of the template mesh, and its topology. Recent neural variants of mesh-based 3DMMs [28, 45, 66, 77, 81] and neural-field-based parametric face models [27, 48, 84, 85, 87] constitute more detailed model priors, but so far do not tackle 3D head reconstruction from monocular RGB videos.

In this work, we propose `Mono`NPHM, a neural parametric head model tailored towards monocular 3D reconstruction from RGB videos. We model an appearance field, coupled with a signed distance field (SDF) that represents the geometry, in *canonical space*. Facial expressions are represented using a backward deformation field that establishes correspondences from *posed space* into the canonical space. Additionally, we augment our backward deformation model using hyper-dimensions [58], in order to increase the dynamic capacity of our model. Building on top of our parametric model, we perform photometric 3D head tracking, by optimizing for latent geometry, appearance, and expression codes. To establish an RGB loss, we utilize SDF-based volumetric rendering [78] of rays in posed space which are backward-warped into canonical space. To account for different lighting conditions we incorporate spherical harmonics shading [64] into the volumetric rendering. Additionally, we find that a landmark loss is crucial for robust tracking through extreme facial movements. We use a discrete set of facial anchor points that is tightly coupled with our geometry representation [27]. We forward-warp the anchors by numerically inverting our backward deformation field using iterative root finding [17] and project them into image space to compute our landmark loss.

Compared to our strongest baselines we improve the reconstruction fidelity, measured by Chamfer distance, by 20%. To sum up our contributions are as follows:

- We introduce `Mono`NPHM, a neural parametric head model that jointly models appearance, geometry, and expression and is augmented with hyper-dimensions for an increased dynamics capacity.
- We tightly condition our appearance network on the underlying geometry, to allow for meaningful gradients during inverse rendering, which we formulate based on dynamic volume rendering of implicit surfaces.
- We introduce a landmark loss using discrete facial anchor points that are tightly coupled with our implicit geometry.

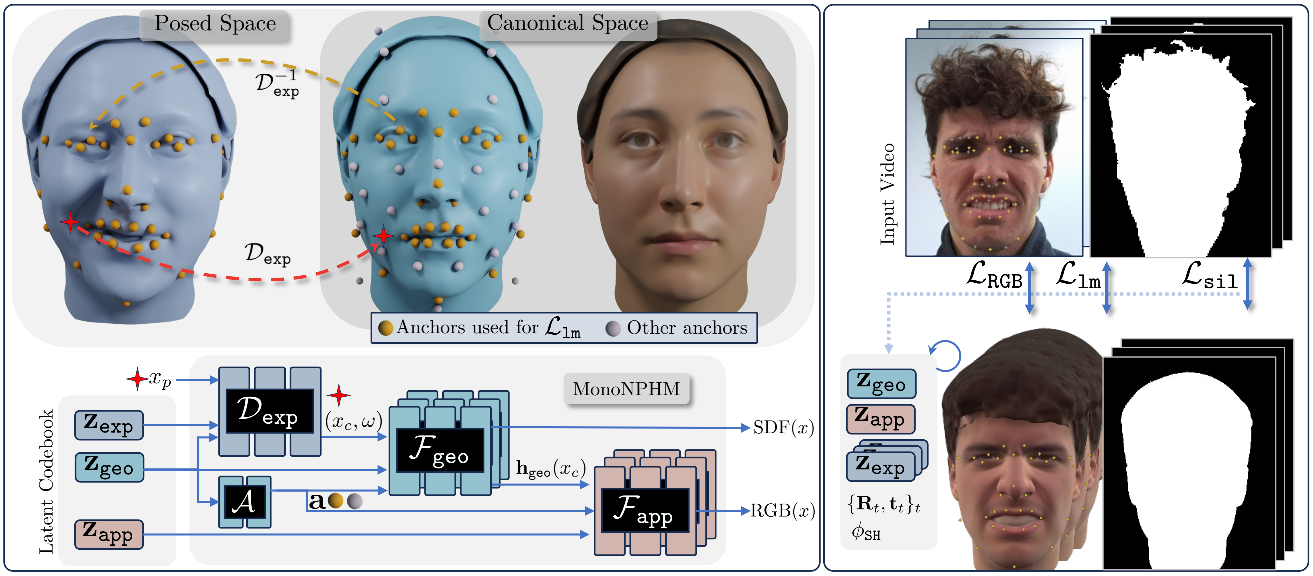
2. Related Work

Mesh-based face models Starting with the seminal work on 3DMMs [8], template-mesh-based PCA models [5, 9, 47, 61, 86] have been widely used for many application in computer graphics and vision. To relax the rigid linear assumption of PCA, subsequent efforts utilized variation auto-encoders(VAEs) [42], generative adversarial networks (GANs) [29], and diffusion models [31] to replace the PCA-basis underlying classical mesh-based 3DMMs [24, 28, 45, 55, 66, 77, 81].

3D face reconstruction from RGB Reconstructing the 3D geometry of a head from RGB images or videos is a fundamental problem in computer vision. Standard approaches optimize the parameters of a 3DMM based on the 2D input [4, 26, 75, 81]. Optimizing the parameters from arbitrary poses, especially in the presence of occlusions and strong shadows, is a very challenging problem. Learning-based methods address this issue by training neural networks to predict the face representation from the input images [18, 20, 49, 69, 74]. In order to model details beyond the 3DMM template, such as wrinkles, several efforts utilize shape-from-shading [25, 34, 71] while others model facial details as displacements maps [15, 21, 36, 46]. Instead of relying on a fixed-topology template mesh, we approach 3D reconstruction from RGB inputs using neural-field-based parametric head models, allowing for the representation of complete human heads with varying topologies.

Neural field-based face models Recent advances on neural fields [80], have shown impressive results on geometry reconstruction and generation [51, 56, 62, 78, 82], neural radiance fields (NeRFs) [16, 40, 52, 53], and dynamic scene reconstructions [3, 33, 44, 57, 58, 70]. Such techniques have been recently used in the context of 3D generative models [6, 12, 13, 79], and NeRF-based parametric models [10, 11, 32, 76, 92] to generate high-fidelity heads that can be rendered from different views. Others have focused on highly detailed geometry representation [27, 84, 85, 87, 88], design a diffusion prior for robust reconstruction from depth sensors [72], and facilitate few-shot 3D reconstruction from RGB images using a mixture of model-based fitting and test-time fine-tuning of model parameters [10, 48, 65]. Closer to our work, [48] is able to reconstruct an animatable head avatar from a single image in the wild. In this work, however, we focus on dynamic 3D reconstruction from monocular RGB videos by explicitly modeling the deformations, which allows us to obtain correspondences across the video.

Person-specific head avatars To escape the limitation of generalizing parameter space, methods for *person-specific*



a) MonoNPHM b) Tracking

Figure 2. **Method overview**: (a) Shows how MonoNPHM operates: First, points x_p in posed-space are backward-warped through $\mathcal{D}_{\text{exp}}$ into canonical space (indicated by red star and arrow). Our canonical geometry and appearance fields are conditioned on facial anchors $\mathbf{a}_c$ (yellow and gray points in canonical space). Instead of conditioning $\mathcal{F}_{\text{app}}$ on canonical coordinates (x_c, ω), we use hidden features $\mathbf{h}_{\text{geo}}$ extracted from the geometry network. (b) We approach tracking using SDF-based volumetric rendering [78] to build photometric and silhoutte terms. Additionally, we enforce a landmark loss by numerically inverting $\mathcal{D}_{\text{exp}}$ using iterative root finding (as indicated by the yellow arrow on the left).

avatars from monocular videos, have shown impressive results. These methods usually incorporate a 3DMM to introduce control in the neural implicit representation [2, 23, 50, 63, 89, 91, 94]. However, they lack generalization and as such, require to be trained per video, in contrast, to our method that generalizes across identities and expressions.

A different line of work focuses on the construction of photo-realistic avatars from abundant multi-view video recordings [43, 50, 63, 68].

3. MonoNPHM

Our work aims at dynamic 3D Face reconstruction in monocular RGB videos. We approach this heavily underconstrained task using model-based photometric tracking through inverse SDF-based rendering. In this section we describe the construction of our underlying, neural field-based model, MonoNPHM, illustrated in Fig. 2, along with its disentangled parametric spaces for shape (Sec. 3.1), appearance (Sec. 3.2) and expression information (Sec. 3.3). In Sec. 4 we propose a model-based dynamic 3D reconstruction algorithm, based on MonoNPHM.

3.1. Canonical Geometry Representation

We represent the head geometry in canonical facial expression, as described by latent code $\mathbf{z}_{\text{geo}}$, using a neural SDF

$$\mathcal{F}_{\text{geo}} : \mathbb{R}^{3+d_{\text{geo}}} \to \mathbb{R}^1, x_c \mapsto \text{SDF}(x_c), \tag{1}$$

operating on points x_c in canonical space. Such an implicit representation provides the necessary topological flexibility to describe complete heads, including hair.

We follow NPHM [27] and compose $\mathcal{F}_{\text{geo}}$ as an ensemble of local MLPs

$$\mathcal{F}_{\text{geo}}(x_c, \mathbf{z}_{\text{geo}}) = \sum_{k \in \mathcal{N}_{x_c}} w_k(x_c, \mathbf{a}_c^k) f_{\text{geo}}^k(x_c - \mathbf{a}_c^k; \mathbf{z}_{\text{geo}}), \tag{2}$$

which are centered around facial *anchor* points $\mathbf{a}_c^k = \mathcal{A}(\mathbf{z}_{\text{geo}}) \in \mathbb{R}^{65\times 3}$, that are predicted by a small MLP $\mathcal{A}$ based on the geometry code $\mathbf{z}_{\text{geo}}$. Therefore, the anchor positions constitute an integral part of the pipeline and provide an important discrete structure which we leverage as a landmark loss for monocular tracking in Sec. 4.4.

For this purpose we design an anchor layout consisting of 65 points, s.t. the most important landmarks of common detectors coincide with anchor points, as shown in Fig. 2. To account for the increased number of anchors, we restrict the computation to the bipartite kNN-graph from x_c to its 8 nearest anchors $\mathcal{N}_{x_c}$, Compared to NPHM, which evaluates all local MLPs, in this case 65, this is more than an eight-fold reduction in memory. To account for the non-uniform spatial arrangement of anchors, we re-scale w_k for each neighborhood separately. Details are provided in our supplementary material.

3.2. Canonical Appearance Representation

We model appearance changes between subjects using separate latent codes $\mathbf{z}_{\text{app}}$, that condition a texture field $\mathcal{F}_{\text{app}}$.

To emphasize the dependence of appearance on the geometry, we incorporate a strong connection between the two networks, similar to PhoMoH [85]. Our motivations come from the fact that an appearance space, which is completely independent of the geometry, could reconstruct the observed color images without providing meaningful gradients for the latest geometry codes.

To this end, we condition our texture field $\mathcal{F}_{\text{app}}$, on features $\mathbf{h}_{\texttt{geo}}(x_c) \in \mathbb{R}^{16}$ which are extracted from the last layer of the geometry MLP $\mathcal{F}_{\texttt{geo}}$ using two narrow linear layers. As illustrated in Fig. 2, $\mathcal{F}_{\text{app}}$ follows the same local structure as our geometry network, i.e. local appearance MLPs

$$f^k_{\text{app}}(\mathbf{h}^k_{\texttt{geo}}(x_c); \mathbf{z}_{\text{app}}) \in [0, 255]^3 \tag{3}$$

are blended using the same weights as in Eq. (2). As we will show later in Sec. 5.4, removing the dependence of $\mathcal{F}_{\text{app}}$ on spatial coordinates x_c and using features $\mathbf{h}_{\texttt{geo}}(x_c)$ instead, is beneficial for RGB-based 3D reconstruction.

3.3. Representing Dynamics

While both previous components operate in canonical space, it is the task of our deformation network

$$\mathcal{D}_{\text{exp}} : \mathbb{R}^{3+d_{\text{exp}}+d_{\texttt{geo}}} \to \mathbb{R}^3, x_p \mapsto x_c \tag{4}$$

to *backward*-warp points x_p in posed space into canonical coordinates x_c. Such a formulation implies that all changes in the geometry and appearance fields between two expressions can be explained through a deformation of space.

To relieve this strong assumption, we relax the formulation by adding *hyper-dimensions*, or *ambient dimensions*, [58] to the output of the deformation network, i.e. $\mathcal{D}_{\text{exp}}(x_p; \mathbf{z}_{\text{exp}}, \mathbf{z}_{\texttt{geo}}) = (x_c, \omega) \in \mathbb{R}^{3+h}$, where h is the number of hyper-dimensions (in practice we use $h = 2$). Consequently, $\mathcal{F}_{\texttt{geo}}$ is provided with canonical coordinates *and* hyper-coordinates ω, which increase the dynamic capacity of the overall network. Fig. 6 demonstrates the topological issues that arise without using hyper-dimensions.

Following previous work [27, 54], $\mathcal{D}_{\text{exp}}$ is conditioned on both $\mathbf{z}_{\text{exp}}$ *and* $\mathbf{z}_{\texttt{geo}}$ since the identity information is relevant to find correct correspondences between x_p and x_c. Note that NPHM [27] uses forward deformations, which we ablate to perform inferior compared to formulation.

3.4. Training

We train all model components and latent codes end-to-end using an auto-decoder formulation [56]. Given a public dataset consisting of high-quality textured 3D scans [27][1], we sample points x_p near the mesh surface and pre-compute $\text{SDF}(x_p)$ and $\text{RGB}(x_p)$ values for direct supervision of our geometry and color fields. Conceptually, we optimize for model parameters Θ and latent codes $\mathcal{Z}$

$$\begin{aligned}\underset{\mathcal{Z},\Theta}{\text{argmin}} \quad \sum_{s\in S, e\in E_s} &\lambda_{\text{SDF}}|\mathcal{F}_{\texttt{geo}}(\mathcal{D}_{\text{exp}}(x_p)) - \text{SDF}(x_p)| + \\ &\lambda_{\text{RGB}}|\mathcal{F}_{\text{app}}(\mathbf{h}_{\texttt{geo}}(x_c)) - \text{RGB}(x_p)| + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}},\end{aligned} \tag{5}$$

where $\mathcal{L}_{\textbf{reg}}$, among others, imposes regularization on all latent codes, supervises anchor predictions, and regularizes deformations and predicted hyper-dimensions to be small. We provide more details about the training and network architectures in our supplementary document.

[1] We use the version 2 release containing a total of 473 identities

4. 3D Dynamic Face Reconstruction

Our main goal is tracking heads in the parametric space of MonoNPHM, in the case of a single, monocular RGB input video. In such a challenging scenario, it is essential that a strong, but expressive, model-prior can guide the optimization through the often under-constrained task. We conceptually visualize this task in Fig. 2. Given a video sequence of RGB frames $\{I_1, \dots I_T\}_{t=1}^T$, associated silhouettes $\{S_1, \dots S_T\}_{t=1}^T$ and 2D facial landmarks $\{L_1, \dots, L_T\}_{t=1}^T$ we aim to reconstruct model parameters $\phi = \{\mathbf{z}_{\text{app}}, \mathbf{z}_{\texttt{geo}}\} \cup \{\mathbf{z}_{\text{exp}}\}_{t=1}^T$, composed of time-invariant codes $\mathbf{z}_{\texttt{geo}}$ and $\mathbf{z}_{\text{app}}$, as well as, per frame expression codes $\mathbf{z}^t_{\text{exp}}$. We solve the tracking task by minimizing the energy

$$\underset{\phi,\zeta,\Pi}{\text{argmin}} \sum_{t=1}^{T} \mathcal{L}^t_{\texttt{RGB}} + \lambda_{\texttt{sil}}\mathcal{L}^t_{\texttt{sil}} + \lambda_{\texttt{lm}}\mathcal{L}^t_{\texttt{lm}} + \lambda_{\texttt{reg}}\mathcal{L}^t_{\texttt{reg}} \tag{6}$$

with respect to latent codes ϕ, head poses $\Pi = \{\mathbf{R}_t, \mathbf{t}_t\}_{t=1}^T$, as well as, lighting parameters $\zeta \in \mathbb{R}^9$ of a 3-band spherical harmonics approximation [64].

The data term of our energy contains a pixel-level RGB loss $\mathcal{L}_{\texttt{RGB}}$ and silhouette loss $\mathcal{L}_S$, as explain in Secs. 4.1 to 4.3, and a landmark loss $\mathcal{L}_{\texttt{lm}}$ for coarse guidance of the expression (see Sec. 4.4). We describe our regularization term and optimization strategy in Sec. 4.5 and Sec. 4.6, respectively.

4.1. Rendering Formulation

To relate the 3D neural field, parameterized by the latent codes, with the 2D observations, we perform volumetric rendering in posed space. Given intrinsic camera parameters K, we transfer the head pose into an extrinsics matrix $E_t = [\mathbf{R}_t|\mathbf{t}_t]$, and shoot a ray $r_p(\tau) = o + \tau \cdot d$ into the scene at time t. Samples along the ray are warped into canonical space $r_c(\tau) = \mathcal{D}_{\text{exp}}(r_p(\tau); \mathbf{z}^t_{\text{exp}}, \mathbf{z}_{\texttt{geo}})$. Consequently, we infer SDF values $\mathcal{F}_{\texttt{geo}}(r_c(\tau); \mathbf{z}_{\texttt{geo}})$ in caonical space. For volume rendering, we rely on the formulation of NeuS [78] to transfer SDF values along a ray into rendering densities

$\sigma(r_c(\tau))$. In total, predicted RGB values $c(\tau)$ along a ray are aggregated into pixel colors

$$\hat{I}_t(r) = \int_{\tau_n}^{\tau_f} w(\tau)c(\tau)d\tau \quad (7)$$

using volume rendering [37, 52]. Here, rendering weights $w(\tau)$ and the accumulated transmittance $T(\tau)$ are defined as follows:

$$w(\tau) = T(\tau)\sigma(r_c(\tau)), \text{ and } T(\tau) = e^{-\int_{\tau_n}^{\tau} \sigma(r_c(s))ds}. \quad (8)$$

4.2. Spherical Harmonics Shading

To bridge the domain gap between our albedo appearance space, and in-the-wild lighting effects, we include a 3-bands spherical harmonics as a simple approximation for the scene lighting [64]. Thus, we obtain shaded RGB predictions

$$c(\tau) = \mathrm{SH}_\zeta(n(\tau))\mathcal{F}_{\texttt{app}}(\mathbf{h}_{\texttt{geo}}(r_c(\tau)); \mathbf{z}_{\texttt{app}}) \quad (9)$$

by multiplying predicted colors with the spherical harmonics term, parameterized by $\zeta \in \mathbb{R}^9$. For this, we use world space normals $n(\tau) = \mathbf{R}_t \nabla_{x_p} \mathcal{F}_{\texttt{geo}}(x_c; \mathbf{z}_{\texttt{geo}})$, where the dependence on x_p is included in the relation $x_c = \mathcal{D}_{\texttt{exp}}(x_p; \mathbf{z}_{\texttt{exp}}, \mathbf{z}_{\texttt{geo}})$. We show the importance of accounting for lighting effects in Sec. 5.4.

4.3. Rendering Losses

The most important term in our inverse rendering is the color loss $\mathcal{L}^t_{\texttt{RGB}} = \mathrm{MAE}(\hat{I}_t, I_t)$, which measures the average L1-loss over all pixels in the foreground region, between predicted image colors $\hat{I}_t$ and observed images I_t.

Additionally, we supervise the silhouette S_t using an average binary cross-entropy loss $\mathcal{L}^t_{\texttt{sil}} = \mathrm{BCE}(\hat{S}_t(r), S_t(r))$ over all pixels, with predicted forground $\hat{S}_t = \int_{\tau_n}^{\tau_f} w(\tau)d\tau$.

4.4. Landmark Loss

Next to the above-mentioned rendering losses we observe that the optimization can get stuck in local minima for extreme mouth movements. We address this issue by incorporating a landmark loss, a common practice in face tracking.

For this purpose, we exploit the structure of the underlying NPHM model that is offered through its anchor points $\mathcal{A}(\mathbf{z}_{\texttt{geo}}) = \mathbf{a}_c$. We determine the anchor positions $\mathbf{a}_p^t$ in posed space that satisfy

$$0 = \mathbf{a}_c^t - \mathcal{D}_{\texttt{exp}}(\mathbf{a}_p^t; \mathbf{z}_{\texttt{exp}}, \mathbf{z}_{\texttt{geo}}) \quad (10)$$

using iterative root finding [17], i.e. the backward deformation field is inverted through a numerical procedure. To coarsely guide $\mathbf{z}_{\texttt{exp}}$ during tracking we enforce

$$\mathcal{L}_{\text{lm}} = \mathrm{MSE}(\pi_{K,E_t}(\mathbf{a}_p^t), L_t), \quad (11)$$

which measures the screen-space distance between detected landmarks L_t and projected posed anchors, where π_{K,E_t} denotes a perspective projection using camera intrinsics K and extrinsics E_t.

4.5. Regularization

We encourage the latent codes to stay within a well-behaved parameter range, which is also enforced during training:

$$\mathcal{L}_{\text{prior}} = \|\mathbf{z}_{\texttt{geo}}\|^2 + \lambda_{\texttt{app}}\|\mathbf{z}_{\texttt{app}}\|^2 + \frac{\lambda_{\texttt{exp}}}{T}\sum_t \|\mathbf{z}_{\texttt{exp}}\|^2. \quad (12)$$

Additionally, we use the symmetry loss from NPHM [27] on the local latent codes contained in $\mathbf{z}_{\texttt{geo}}$ and $\mathbf{z}_{\texttt{app}}$, and enforce temporal smoothness on time-dependent parameters

$$\mathcal{L}_{\text{smooth}} = \mathrm{TV}(\mathbf{z}_{\texttt{exp}}) + \lambda_{\text{rot}}\mathrm{TV}(\mathbf{R}_t) + \lambda_{\text{trans}}\mathrm{TV}(\mathbf{t}_t). \quad (13)$$

4.6. Optimization Strategy

We optimize Eq. (6) using stochastic gradient descent (SGD) and the Adam optimizer [41]. We initialize all latent codes as zeros, ζ is initialized as uniform lighting from all directions, and head poses $\mathbf{R}_t, \mathbf{t}_t$ are initialized from a tracked FLAME model.

We start our optimization by separately optimizing the first frame, and then optimize for the remaining frames sequentially in a frame-by-frame fashion, where $\mathbf{z}_{\texttt{geo}}$, $\mathbf{z}_{\texttt{app}}$, and ζ remain frozen. This strategy provides good estimates over all parameters and serves as initialization for our main stage, where we optimize over *all* parameters jointly. For each optimization step a random timestep t and random rays for the rendering losses are sampled. Our smoothness loss is computed between the neighboring frames $t-1$ and $t+1$.

5. Results

To evaluate our goal of dynamic face reconstruction, we record 20 Kinect sequences in a casual setting, for a lack of publicly available alternatives. The RGB sensor serves as input, while the depth sensor allows for a geometric evaluation. We record 5 participants (3 female, 2 male) under a wide range of facial expressions, emotions and include one talking sequence. All participants signed agreements compliant with GDPR requirements. We record each sequence for 12 seconds at 15 frames per second, resulting in 180 frames per sequence.

5.1. Metrics

We report unidirectional L_1-Chamfer distance in meters from the back-projected depth map to the reconstructions, which cover the complete head. Similarly, we report the unidirectional cosine-similarity of normals. Additionally, we measure the recall [73], i.e. the percentage of ground truth points that are covered by at least one point on the reconstruction w.r.t. to a given threshold distance.

Evaluation Protocol To eliminate any remaining depth ambiguity, we optimize for a similarity transform from reconstructed mesh to ground truth point cloud using ICP [7].

Input RGB DECA [20] MICATracker [93] NHA [30] IMavatar [89] HeadNeRF [32] Ours

Figure 3. Comparison of the 3D reconstruction quality from monocular RGB videos against our baselines. We show error maps with color-coded point-to-mesh distance from ground truth Kinect depth to the reconstructed meshes.

To exclude the sensor noise of the regions inside the mouth and eyes and to account for differences between the compared methods, we remove these regions, as well as the hair and neck region, using facial segmentation [90]. We visualize the resulting ground truth point clouds in Fig. 3, which are color-coded according to the Chamfer distance.

5.2. Baselines

Mesh-Based Baselines: 3DMMs are the most common model prior for 3D face tracking. Therefore, we compare against DECA [20] and the MICA tracker [93]. The former is a CNN that is trained in a self-supervised fashion on in-the-wild images to predict FLAME [47] parameters. The latter is a state-of-the-art face tracker, inspired by Face2Face [75]. Additionally, we compare against Neural Head Avatar (NHA) [30], which learns person-specific face offsets and expression dependent neural textures.

Field-Based Baselines: IMavatar [89] uses neural fields to explain details beyond an underlying FLAME model, which is used as guidance during optimization. HeadNeRF [32] is a NeRF-based [52] neural 3DMM. To achieve high-fidelity appearance it relies on a screenspace CNN.

5.2.1 Implementation Details

Training MonoNPHM We implement our model in Pytorch [59] and utilize PytorchGeometric [22] to restrict

Method	L_1-Chamfer ↓	N. C. ↑	Recall@2.5mm ↑
DECA [20]	0.0034	0.917	0.644
MICATracker [93]	0.0030	0.932	0.654
NHA [30]	0.0055	0.872	0.490
IMavatar [89]	0.0054	0.888	0.625
HeadNeRF [32]	0.0049	0.883	0.504
Ours	**0.0024**	**0.940**	**0.785**

Table 1. Quantitative comparison of 3D face reconstruction from RGB videos. The chamfer distance is reported in meters.

computations in canonical space to the k nearest anchors. We follow NPHM [27] and use 64 dimensions for the global parts of $\mathbf{z}_{\text{app}}$ and $\mathbf{z}_{\text{exp}}$. For the local codes we use 32 dimensions. For the expression codes $\mathbf{z}_{\text{exp}}$ we use 100 dimensions. We train our model for 2500 epochs, use a batch size of 64, and a learning rate of $5e^{-4}$ for the networks and $2e^{-3}$ for the latent codes. We train on the updated release of the NPHM dataset [27] using 4 NVIDIA RTX2080 GPUs with 12GB of VRAM taking roughly 52 hours until convergence. More details are provided in our supplementary.

Data Pre-Processing We perform several common preprocessing steps to remove parts of the observed images that are not included in our learned prior. Namely, we rely on face detection [19], facial landmark detection [35], semantic segmentation to remove the torso [90], as well as, video matting [39] to remove the background. For all baselines,

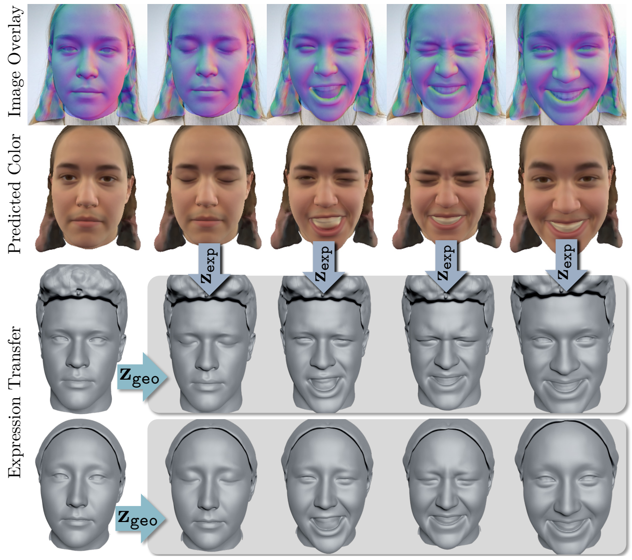


Figure 4. **Cross-Reenactment:** We show five reconstructions from same sequence (top two rows). A 50-50 overlay (top row) of reconstructed normals and the RGB images shows accurate image-space-alignment. The bottom two rows show cross-reenactment by transferring expression codes to two other subjects.

we follow their proposed pre-processing pipeline.

Tracking For each step of SGD we randomly sample 500 rays. During volume rendering, we randomly sample 32 coarse samples, and additional 32 samples using importance sampling. We start with a large variance for the NeuS [78] rendering, which is decayed over time to concentrate tightly around the surface. We perform 250 optimization steps for the first frame, and 60 steps per frame otherwise. To build forward correspondences for our landmark loss, we use 5 random initializations for iterative root finding. More details are provided in the supplementary.

Our optimization operates at roughly 1.2 frames per minute. As a comparison, the MICA tracker can track 2 frames per minute using the default settings, and IMavatar operates at roughly 0.4 frames per minute.

5.3. Tracking Results

We compare MonoNPHMto our baselines by fitting each model to all the 20 monocular RGB sequences individually. We report quantitative and qualitative results in Tab. 1 and

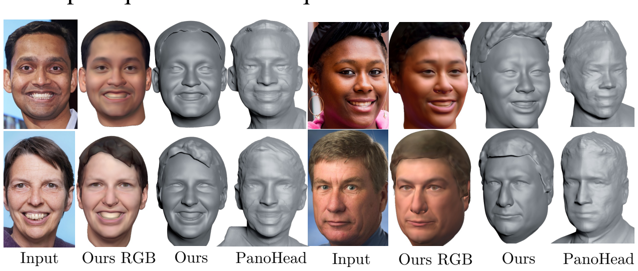


Figure 5. **Single-Image Reconstructions** on FFHQ [38] images.

Fig. 3, respectively. For results on the complete sequences, we kindly refer to our supplementary video. Fig. 3 shows that MonoNPHM reconstructs important details about the face shape and expressions, that significantly help to recognize the identity and interpret the reconstructed emotion correctly. Compared to the 3DMM-based approaches, DECA and the MICA tracker, MonoNPHM is capable of reconstructing complete heads, including the mouth inside and hair. IMavatar, on the other hand, suffers from its increased representational capacity compared to FLAME, due to the difficulty of task. Both NHA and HeadNeRF employ high capacity neural networks for high-fidelity renderings. Consequently, the geometry of these approaches is underconstrained. The quantitative evaluation reported in Tab. 1 confirms these findings.

Additionally, we show qualitative results for five frames of the same sequence in Fig. 4, to demonstrate temporal consistency and the alignment of our reconstructed geometry against the input sequence in screen space. Alongside, we show the predicted color images $\hat{I}$ of MonoNPHM, to give further insights into our rendering loss $\mathcal{L}_{\text{RGB}}$, which mainly drives our optimization. Finally, we perform cross-reenactment by transferring the reconstructed latent codes $\mathbf{z}_{\text{exp}}$ to the identity codes $\mathbf{z}_{\text{geo}}$ and $\mathbf{z}_{\text{app}}$ from another participant. Visually, the resulting reenactments capture the contents of the original expression to a high degree.

Single-Image Reconstruction Furthermore, in Fig. 5, we demonstrate MonoNPHM's 3D reconstruction capabilities on single in-the-wild images from the FFHQ dataset [38]. This indicates that our learned prior is strong enough for sparse observations, extreme lighting conditions and diverse identities. Additionally, we include a qualitative comparison against PanoHead [1], a recent SOTA 3D generative head model trained on the FFHQ dataset.

5.4. Ablations

	Method	L_1-Ch. ↓	N. C. ↑	Recall ↑
tracking	w/ sphere tracing	0.0033	0.905	0.674
	w/o spher. harm.	0.0028	0.923	0.718
	w/o $\mathcal{L}_{\text{lm}}$	0.0027	0.939	0.745
architecture	NPHM_{app}	0.0028	0.926	0.724
	w/ $\|\mathbf{a}\| = 39$	0.0027	0.934	0.761
	w/o color comm.	0.0028	0.933	0.735
	w/ global MLP	0.0026	**0.940**	0.768
	Ours	**0.0024**	**0.940**	**0.785**

Table 2. Ablations on single components of our tracking approach (first 3 rows), and of our architecture (second section).

We support several of our claims by ablations on the same 20 Kinect sequences. Quantitative results are reported in Tab. 2 andqualitative results in our supplemental.

Tracking Algorithm. Firstly, we show that NeuS-style volume rendering [78], instead of the sphere-tracing-based render for implicit surfaces of IDR [82], is essential for the

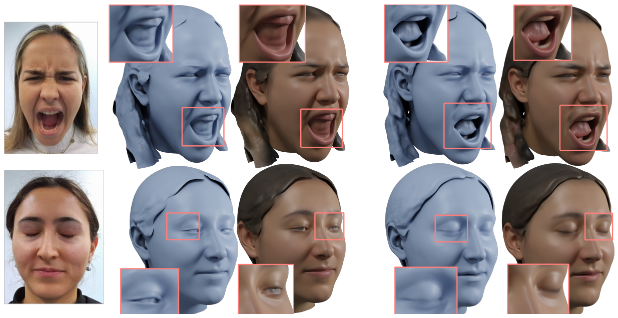


Figure 6. **Effect of Hyper-Dimensions**: Without the addition of hyper-dimensions, the backward deformation model cannot accurately reconstruct geometry and color for topologically challenging expressions, e.g. the opening of mouth and closing of eyes.

success of our tracking approach. Second, the use of spherical harmonics is crucial to account for lighting conditions that vastly differ from our training data. Otherwise, the RGB loss is dominated by lighting effects that our model cannot explain. Lastly, we note that without using the landmark loss $\mathcal{L}_{\texttt{lm}}$ our optimization often performs similarly well, but tends to get stuck in local minima, if large and fast mouth movements, e.g. during shouting, are encountered.

Difference to NPHM [27]. To analyze differences to the NPHM model, we train an extended version of NPHM, denoted as $\text{NPHM}_{\texttt{app}}$, which predicts color in the canonical space identical to MonoNPHM, and uses our anchor layout. The main remaining differences are our use of backward (instead of forward) deformations and hyper dimensions. Additionally, we ablate the effect of our anchor layout by training MonoNPHM using NPHM's anchor layout consisting of $|a| = 39$ anchor points. This leads to a less effective landmarks loss and slightly reduced model capacity.

The effect of $\mathbf{h}_{\texttt{geo}}$. Furthermore, we show the significance of removing the communication channel between geometry and color networks. To this end, we train a model that uses canonical coordinates x_c instead of $\mathbf{h}_{\texttt{geo}}(x_c)$ as input to the local color MLPs defined in Eq. (3). We hypothesize that for such a model, gradients through our most important loss $\mathcal{L}_{\texttt{RGB}}$ are less informative for $\mathbf{z}_{\texttt{geo}}$ and $\mathbf{z}_{\texttt{exp}}$.

Local vs Global MLPs. Additionally, we ablate the effect of using the local MLP ensemble from NPHM [27] against a simpler architecture, that represents $\mathcal{F}_{\texttt{geo}}$ and $\mathcal{F}_{\texttt{app}}$ using a global MLP. To account for good tracking of extreme expressions, we find that the landmark loss is equally important for this global architecture. To this end, we include the anchor prediction MLP $\mathcal{A}$ into this ablation experiment, such that the usage of $\mathcal{L}_{\texttt{lm}}$ becomes viable. Doing so, we are able to associate $\mathbf{a}_c$ with $\mathbf{z}_{\texttt{geo}}$, and achieve good tracking performance, with slightly fewer geometric details, as reflected in the metrics.

Effect of Hyper-Dimensions. Finally, in Fig. 6 we highlight the importance of using hyper-dimensions for correctly representing topologically challenging expressions, such as the closing of the eyes and opening of the mouth. Note that the metrics are barely affected, since mouth and eye regions are excluded due to strong sensor noise.

6. Limitations

In our experiments, we show that MonoNPHM can reconstruct high-quality human heads from monocular videos; however, at the same time, we believe that there are still several limitations and opportunities for future work. For instance, while spherical harmonics can be used to account for simple lighting conditions without increasing the model complexity, we believe that reconstructions could be improved by addressing lighting and shadows more thoroughly. Possible options are the inclusion of a more advanced shading model during volume rendering [67], image-space delighting [83], as well as, CNN-based image encoders [14, 20]. Another limitation is our tracking speed. While this is partially explained by our unoptimized implementation that runs a full optimization for each frame, we believe that several advances can be made, e.g. using CNN-based initialization [60], coarse-to-fine optimization, faster neural-field backbones [16, 53] and second-order optimization for tracking [75].

7. Conclusion

In this work we have introduced MonoNPHM, a neural-field-based parametric face model, that represents faces using an SDF and texture field in canonical space, and represents movements using backward deformations, augmented with hyper-dimensions. We enforce a tight communication between appearance and geometry to facilitate efficient inverse rendering. By including explicit control points in our implicit geometry representation, we have developed a highly accurate 3D face tracking algorithm based on volumetric rendering for implicit surfaces. MonoNPHM achieves significantly more accurate 3D reconstruction on challenging monocular RGB videos, compared to all our baselines. We believe that our work makes the use of neural parametric head models much more accessible for many downstream tasks. We hope that our work inspires more research to explore the use of neural-field-based parametric models and develop the necessary toolsets that are already available for classical 3DMMs.

Acknowledgements

This work was funded by Synthesia and supported by the ERC Starting Grant Scan2CAD (804724), the German Research Foundation (DFG) Research Unit "Learning and Simulation in Visual Computing". We would like to thank our research assistants Mohak Mansharamani and Kevin Qu, and Angela Dai for the video voice-over.

References

[1] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20950–20959, 2023. 7

[2] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20364–20373, 2022. 3

[3] Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O'Toole, and Changil Kim. HyperReel: High-fidelity 6-DoF video with ray-conditioned sampling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[4] Haoran Bai, Di Kang, Haoxian Zhang, Jinshan Pan, and Linchao Bao. Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 362–371, 2023. 2

[5] Linchao Bao, Xiangkai Lin, Yajing Chen, Haoxian Zhang, Sheng Wang, Xuefei Zhe, Di Kang, Haozhi Huang, Xinwei Jiang, Jue Wang, Dong Yu, and Zhengyou Zhang. High-fidelity 3d digital human head creation from rgb-d selfies. *ACM Transactions on Graphics*, 2021. 2

[6] Alexander W. Bergman, Petr Kellnhofer, Wang Yifan, Eric R. Chan, David B. Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. In *NeurIPS*, 2022. 2

[7] P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. 5

[8] 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

[9] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2): 233–254, 2018. 2

[10] Marcel C. Bühler, Kripasindhu Sarkar, Tanmay Shah, Gengyan Li, Daoye Wang, Leonhard Helminger, Sergio Orts-Escolano, Dmitry Lagun, Otmar Hilliges, Thabo Beeler, and Abhimitra Meka. Preface: A data-driven volumetric prior for few-shot ultra high-resolution face synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3402–3413, 2023. 2

[11] 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 Trans. Graph.*, 41(4), 2022. 2

[12] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *arXiv*, 2020. 2

[13] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 2

[14] Prashanth Chandran, Gaspard Zoss, Paulo F. U. Gotardo, and Derek Bradley. Continuous landmark detection with 3d queries. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 16858–16867. IEEE, 2023. 8

[15] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9429–9439, 2019. 2

[16] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 8

[17] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 5

[18] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 2

[19] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020. 6

[20] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG), Proc. SIGGRAPH*, 40(4):88:1–88:13, 2021. 2, 6, 8

[21] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4): 1–13, 2021. 2

[22] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. 6

[23] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, 2021. 3

[24] Stathis Galanakis, Alexandros Lattas, Stylianos Moschoglou, and Stefanos Zafeiriou. Fitdiff: Robust monocular 3d facial shape and reflectance estimation using diffusion models, 2023. 2

[25] 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):1–15, 2016. 2

[26] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting

for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1155–1164, 2019. 2

[27] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Learning neural parametric head models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 4, 5, 6, 8

[28] Shunwang Gong, Lei Chen, Michael Bronstein, and Stefanos Zafeiriou. Spiralnet++: A fast and highly efficient mesh convolution operator. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[29] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[30] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022. 6

[31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 2

[32] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 6

[33] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *arXiv preprint arXiv:2305.06356*, 2023. 2

[34] Luo Jiang, Juyong Zhang, Bailin Deng, Hao Li, and Ligang Liu. 3d face reconstruction with geometry details from a single image. *IEEE Transactions on Image Processing*, 27 (10):4756–4770, 2018. 2

[35] Haibo Jin, Shengcai Liao, and Ling Shao. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *International Journal of Computer Vision*, 129(12):3174–3194, 2021. 6

[36] James T. Kajiya and Brian P Von Herzen. Ray tracing volume densities. *SIGGRAPH Comput. Graph.*, 18(3):165–174, 1984. 2

[37] James T. Kajiya and Brian P Von Herzen. Ray tracing volume densities. In *Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques*, page 165–174, New York, NY, USA, 1984. Association for Computing Machinery. 5

[38] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[39] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson W. H. Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition, 2022. 6

[40] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2

[41] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. 5

[42] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 2

[43] Tobias Kirschstein, Simon Giebenhain, and Matthias Nießner. Diffusionavatars: Deferred diffusion for high-fidelity 3d head avatars, 2023. 3

[44] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), 2023. 2

[45] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Fitme: Deep photorealistic 3d morphable model avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8629–8640, 2023. 2

[46] Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 394–403, 2023. 2

[47] 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, 6

[48] Connor Z. Lin, Koki Nagano, Jan Kautz, Eric R. Chan, Umar Iqbal, Leonidas Guibas, Gordon Wetzstein, and Sameh Khamis. Single-shot implicit morphable faces with consistent texture parameterization. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 2

[49] Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5891–5900, 2020. 2

[50] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 3

[51] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 2

[52] 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, 5, 6

[53] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 2, 8

[54] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12695–12705, 2021. 4

[55] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, and Stefanos Zafeiriou. Relightify: Relightable 3d faces from a single image via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[56] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2, 4

[57] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2

[58] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), 2021. 2, 4

[59] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6

[60] Dario Pavllo, David Joseph Tan, Marie-Julie Rakotosaona, and Federico Tombari. Shape, pose, and appearance from a single image via bootstrapped radiance field inversion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[61] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 2

[62] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision*, pages 523–540. Springer, 2020. 2

[63] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians, 2023. 3

[64] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, page 497–500, New York, NY, USA, 2001. Association for Computing Machinery. 2, 4, 5

[65] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5620–5629, 2021. 2

[66] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3D faces using convolutional mesh autoencoders. In *European Conference on Computer Vision (ECCV)*, pages 725–741, 2018. 2

[67] Anurag Ranjan, Kwang Moo Yi, Jen-Hao Rick Chang, and Oncel Tuzel. Facelit: Neural 3d relightable faces. In *CVPR*, 2023. 8

[68] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars, 2023. 3

[69] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019. 2

[70] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023. 2

[71] Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M Seitz. Total moving face reconstruction. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 796–812. Springer, 2014. 2

[72] Jiapeng Tang, Angela Dai, Yinyu Nie, Lev Markhasin, Justus Thies, and Matthias Niessner. Dphms: Diffusion parametric head models for depth-based tracking, 2023. 2

[73] Maxim Tatarchenko*, Stephan R. Richter*, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, 2019. 5

[74] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 1274–1283, 2017. 2

[75] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2016. 2, 6, 8

[76] Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Gotardo. Morf: Morphable radiance

fields for multiview neural head modeling. In *ACM SIGGRAPH 2022 Conference Proceedings*, New York, NY, USA, 2022. Association for Computing Machinery. 2

[77] Lizhen Wang, Zhiyua Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2022)*, 2022. 2

[78] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021. 2, 3, 4, 7

[79] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4563–4573, 2023. 2

[80] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *Computer Graphics Forum*, 2022. 2

[81] 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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[82] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 2, 7

[83] Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics (TOG)*, 2022. 8

[84] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12803–12813, 2021. 2

[85] Mihai Zanfir, Thiemo Alldieck, and Cristian Sminchisescu. Phomoh: Implicit photorealistic 3d models of human heads. *CoRR*, abs/2212.07275, 2022. 2, 4

[86] Longwen Zhang, Zijun Zhao, Xinzhou Cong, Qixuan Zhang, Shuqi Gu, Yuchong Gao, Rui Zheng, Wei Yang, Lan Xu, and Jingyi Yu. Hack: Learning a parametric head and neck model for high-fidelity animation. *ACM Trans. Graph.*, 42(4), 2023. 2

[87] Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. Imface: A nonlinear 3d morphable face model with implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[88] Mingwu Zheng, Haiyu Zhang, Hongyu Yang, Liming Chen, and Di Huang. Imface++: A sophisticated nonlinear 3d morphable face model with implicit neural representations, 2023. 2

[89] Yufeng Zheng, Victoria Fernández Abrevaya, Xu Chen, Marcel C. Bühler, Michael J. Black, and Otmar Hilliges. I M avatar: Implicit morphable head avatars from videos. *CoRR*, abs/2112.07471, 2021. 3, 6

[90] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022. 6

[91] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[92] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. In *European Conference on Computer Vision*, 2022. 2

[93] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision (ECCV)*. Springer International Publishing, 2022. 6

[94] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars, 2022. 3