Learning Neural Parametric Head Models

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

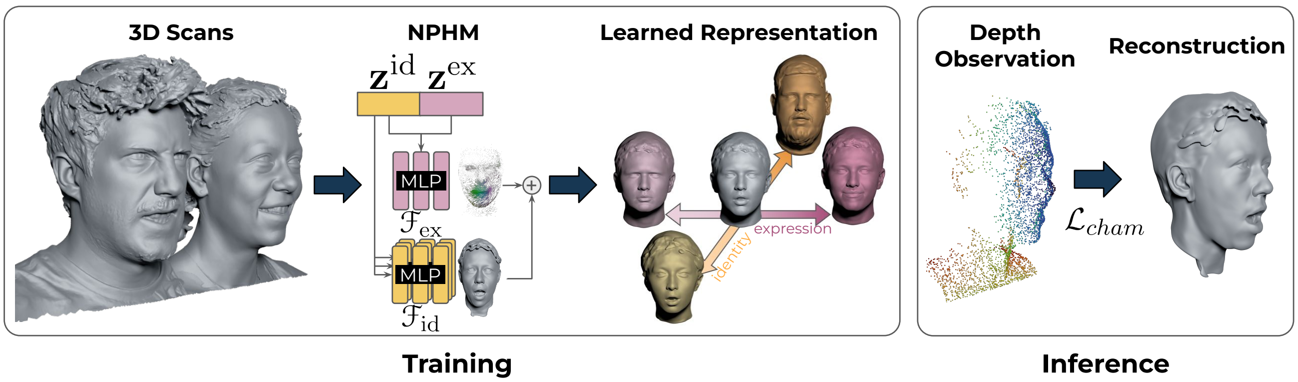


Figure 1. We propose to learn a neural parametric head model based on neural fields: first, we capture a large dataset of over 3700 high-fidelity head scans with varying shapes and expressions (left). We then non-rigidly register these scans to generate our training data. As a result of training, we obtain a disentangled latent that spans the space of shapes $\mathbf{z}^{\text{id}}$ and expressions $\mathbf{z}^{\text{ex}}$ (middle). At inference time, we can leverage the prior of our learned representation by fitting our model to a sparse input point cloud by solving for the latent codes (right).

Abstract

We propose a novel 3D morphable model for complete human heads based on hybrid neural fields. At the core of our model lies a neural parametric representation that disentangles identity and expressions in disjoint latent spaces. To this end, we capture a person's identity in a canonical space as a signed distance field (SDF), and model facial expressions with a neural deformation field. In addition, our representation achieves high-fidelity local detail by introducing an ensemble of local fields centered around facial anchor points. To facilitate generalization, we train our model on a newly-captured dataset of over 3700 head scans from 203 different identities using a custom high-end 3D scanning setup. Our dataset significantly exceeds comparable existing datasets, both with respect to quality and completeness of geometry, averaging around 3.5M mesh faces per scan[1]*. Finally, we demonstrate that our approach outperforms state-of-the-art methods in terms of fitting error and reconstruction quality.*

[1]We will publicly release our dataset along with a public benchmark for both neural head avatar construction as well as an evaluation on a hidden test-set for inference-time fitting.

1. Introduction

Human faces and heads lie at the core of human visual perception, and hence are key to creating digital replica of someones identity, likeliness, and appearance. In particular, 3D reconstruction of human heads from sparse inputs, such as point clouds, is central to a wide range of applications in the context of gaming, augmented and virtual reality, and digitization in our modern digital era. One of the most successful lines of research to address this challenging problem are parametric face models, which represent both shape identities and expressions featuring a low-dimensional parametric space. These Blendshape and 3D morphable models (3DMMs) have achieved incredible success, since they can be fitted to sparse inputs, regularize out noise, and provide a compact 3D representation. As a result, many practical settings could be realized, ranging from face tracking and 3D avatar creation to facial-reenactment applications [49].

Traditionally, 3DMMs, are based on a low-rank approximation of the underlying 3D mesh geometry. To this end, a template mesh with fixed topology is non-rigidly registered to a series of 3D scans. From this template registration, a 3DMM can be computed using dimensionality

Website: https://simongiebenhain.github.io/NPHM reduction methods such as principal component analysis (PCA). The quality of the resulting parametric space depends strongly on the quality of 3D scans, their registration, and the ability to disentangle identity and expression variations. While these PCA-based models exhibit excellent regularizing properties, their inherent limitation lies in their inability to represent local surface detail and the reliance on a template mesh of fixed topology, which inhibits the representation of diverse hair styles.

In this work, we propose neural parametric head models (NPHM), which represent complete human head geometry in a canonical space using an SDF, and morph the resulting geometry to posed space using a forward deformation field. By decoupling the human head representation into these two spaces, we are able to learn disentangled latent spaces – one of the core concepts of 3DMMs. Furthermore, we decompose the implicit geometry representation in canonical space into an ensemble of local MLPs. Each part is represented by a small MLP that operates in a local coordinate system centered around face keypoints. Additionally, we exploit face symmetry by sharing network weights of symmetric regions. This decomposition into separate parts imposes a strong geometry prior and helps to improve both generalization and provide higher levels of detail.

In order to train our model, we capture a new high-fidelity head dataset with a high-end capture rig, which is composed of over 3700 3D head scans from 203 different people. After rigidly aligning all scans in a canonical coordinate system, we train our identity network on scans in canonical expression. In order to train the deformation network, we non-rigidly register each scan against a template mesh, which we in turn use as training data for our neural deformation model. At inference time, we can then fit our model to a given input point cloud by optimizing for the latent code parameters for both expression and identity. In a series of experiments, we demonstrate that our neural parametric model outperforms state-of-the-art models and can represent complete heads, including fine details.
In sum, our contributions are as follows:

- We introduce a novel 3D dataset captured with a high-end capture rig, including over 3700 3D scans of human heads from 203 different identities.
- We propose a new neural-field-based parametric head representation, which facilitates high-fidelity local details through an ensemble of local implicit models.
- We demonstrate that our neural parametric head model can be robustly fit to range data, regularize out noise, and outperform existing models.

2. Related Work

3D morphable face and head models. The seminal work of Blanz and Vetter [1] was one of the first to introduce a model-based approach to represent variations in human faces using PCA. Since the scans were captured in constrained environments, the expressiveness of the model was relatively limited. As such, improvements in the registration [29], as well as the use of data captured in the wild [3, 4, 31], led to significant advances. Thereafter, more advanced face models were introduced, including multilinear models of identity and expression [2, 6], as well as models that combined linear shape spaces with articulated head parts [18], and localized approaches [23].

With the advent of deep learning, various works focused on extending face and head 3DMMs beyond linear spaces. To this end, convolutional neural network based architectures have been proposed to both regress the model parameters and reconstruct the face [16, 37–39, 42, 43]. At the same time, graph convolutions [5, 14] and attention modules [11] have been proposed to model the head mesh geometry.

Neural field representations. Neural field-based networks have emerged as an efficient way to implicitly represent 3D scenes. In contrast to explicit representations (e.g., meshes or voxel grids), neural fields are well-suited to represent geometries of arbitrary topology. Park et al. [26] proposed to represent a class-specific SDF with an MLP that is conditioned on a latent variable. Similarly, Mescheder et al. [21] implicitly define a surface as the decision boundary of a binary classifier and Mildenhall et al. [22] represent a radiance field using an MLP by supervising a photometric loss on the rendered images.

Building upon these approaches, a series of works focus on modeling deformations. These methods use a separate network to model the deformations that occur in a sequence (e.g., [27, 28]), and have been successfully applied to animation of human bodies [17, 19] and heads [46]. Following this paradigm, a number of neural parametric models have been proposed for bodies [9, 24, 25], faces [45], and —most closely related to our work— heads [32, 41, 44]. For instance, H3D-Net [32] and MoRF [41] proposed 3D generative models of heads, but do not account for expression-specific deformations. Recently, neural parametric models for human faces [44, 45] and bodies [9, 10, 24, 25] have explored combinations of SDFs and deformation fields, to produce complex non-linear deformations, while maintaining the flexibility of an implicit geometry representation. Our work is greatly inspired by these lines; however, the key difference is that we tailor our neural field representation specifically to human heads through an ensemble of local MLPs. Thereby, our work is also related to local conditioning methods for neural fields of arbitrary objects [8, 12, 13, 30], human bodies [25, 48] and faces [45]. Compared to ImFace [45], our model is more local, incorporates a symmetry prior, represents a complete head and models forward instead of backward deformations, which allows much faster animation.

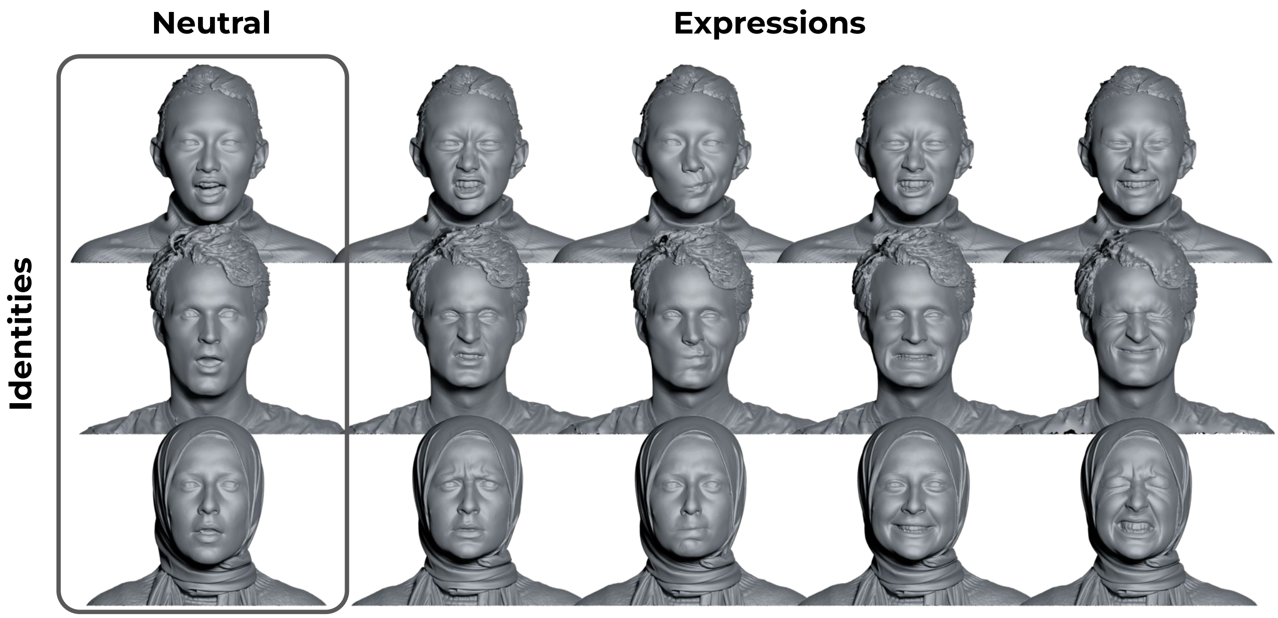


Figure 2. 3D head scans from our newly-captured dataset: for each person (rows), we first capture a neutral pose, followed by several scans in different expressions (columns). Overall, our dataset has more than 3700 3D scans from 203 people.

3. Dataset Acquisition

Our dataset comprises 203 subjects, 29% female, and contains over 3700 3D scans; see Table. 1. Our 3D head scans show great levels of detail and completeness, as shown in Fig. 2. Additionally, we do not require participants to wear a bathing cap as in the FaceScape dataset [43], allowing for the capture of natural hair styles to a certain degree. See Fig. 3 for a visual comparison of our new dataset to other 3D face datasets.

Num. Subjects	203 (144m/59f)
Total num. Scans	3720
Num. Vertices/Scan	$\approx$ 1.5M

Table 1. Statistics of our 3D scanning dataset.

3.1. Capture Setup

Our setup is composed of two Artec Eva scanners [35], that are rotated 360° around a subject's head using a robotic actuator. Each scan takes only 6 seconds, which is crucial to keep involuntary, non-rigid facial movements to a minimum. The scanners operate at 16 FPS, and are aligned through the scanning sequence and fused into a single mesh; each fused scan contains approximately 1.5M vertices and 3.5M triangles. Each participant is asked to perform 23 different expressions, which are adopted from the FACS coded expression proposed in FaceWarehouse [7], see our supplemntal for details. Importantly, we capture a neutral expression with the mouth open, which later serves as canonical expression, as described in Section 4.

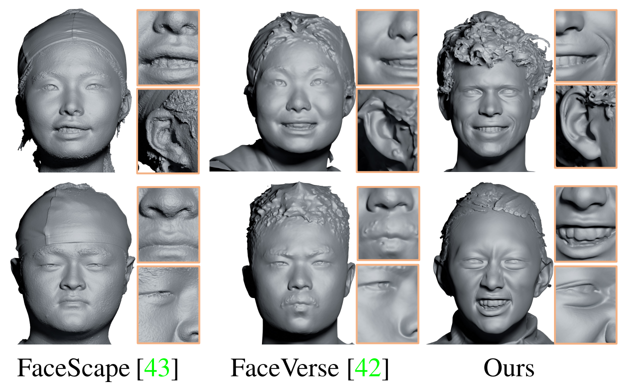


Figure 3. Compared to recent multi-view stero 3D face dataset, our data exhibits sharper details and less noise.

3.2. Registration Pipeline

Registering all head scans against a common template is a key requirement to effectively train our parametric head model. First, we start with a rigid alignment into our canonical coordinate system; second, we non-rigidly register all scans to a common template.

3.2.1 Rigid Alignment

We leverage 2D face landmark detectors to obtain a rigid transformation into the canonical coordinate system of the FLAME model [18]. To this end, we deploy the MediaPipe [20] face mesh detector and back-project a subset of 48 landmarks corresponding to iBUG68 annotations [33] to the 3D scan. Since not all viewing angles of the scanner's

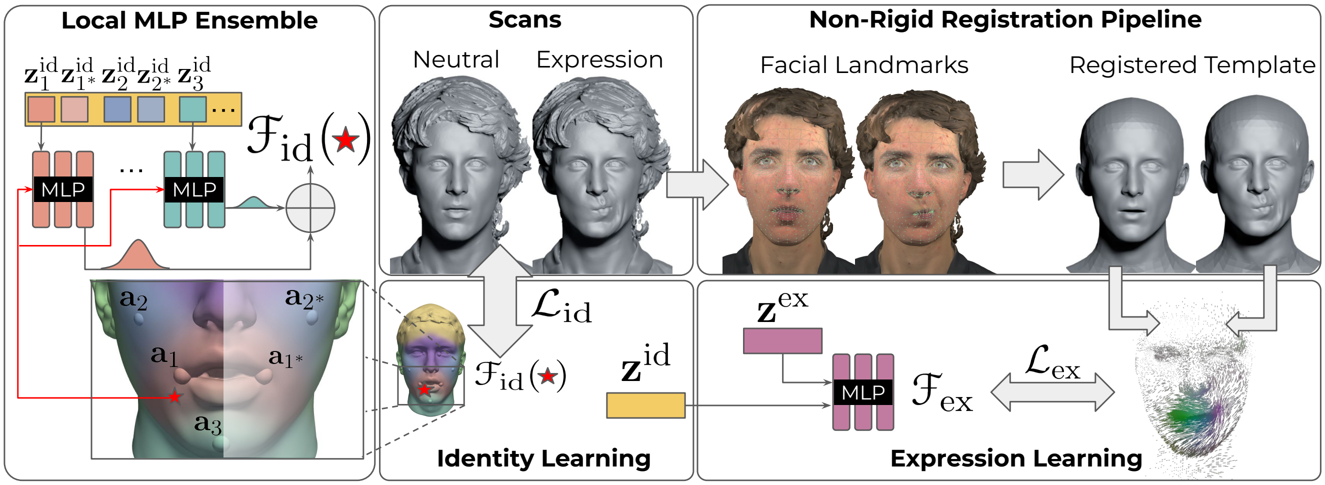


Figure 4. Method overview: at the core of our neural parametric head model lies a neural field representation that parameterizes shape and expressions in disentangled latent spaces. Specifically, we propose a local MLP ensemble that is anchored at face keypoints (left). We train this model by leveraging a set of high-fidelity 3D scans from our newly-captured dataset comprising various expressions for identity (middle). In order to obtain the ground truth deformation samples, we non-rigidly register all scans to a common template (right).

trajectories are suited for 2D facial landmark detection, we instead use frontal renderings of the colored meshes, which yields robust detection quality. Note that the initial landmark detection is the only time we use the scanner's color images. We then calculate a similarity transform using [40] to transform the detected landmarks to the average face of FLAME.

3.2.2 Non-Rigid Registration

As a non-rigid registration prior, we first constrain the non-rigid deformation to FLAME parameter space, before optimizing an offset for each vertex. Additionally, we back-project 2D hair segmentation masks obtained by FaRL [47] to mask out the respective areas of the scans.

Initialization. Given the 23 expression scans $\{S_j\}_{j=1}^{23}$ of a subject, we jointly estimate identity parameters $\mathbf{z}^{\text{id}} \in \mathbb{R}^{100}$, expression parameters $\{\mathbf{z}_j^{\text{ex}}\}_{j=1}^{23}$, and jaw poses $\{\theta_j\}_{j=1}^{23}$ of the FLAME model, as well as a shared scale $s \in \mathbb{R}$ and per-scan rotation and translation corrections $\{R_j\}_{j=1}^{23}$ and $\{t_j\}_{j=1}^{23}$. Updating the initial similarity transform is crucial to obtaining a more consistent canonical alignment.

Let Φ_j denote all parameters affecting the j-th FLAME model and V_{Φ_j} its vertices. We jointly optimize for these parameters by minimizing

$$\underset{\Phi_1,\dots\Phi_{23}}{\arg\min}\sum_{j=1}^{23}\Big[\lambda_l\|L_j-\hat{L}_j\|_1+d(V_{\Phi_j},\mathcal{S}_j)+\mathcal{R}(\Phi_j)\Big], \quad (1)$$

where $L_j \in \mathbb{R}^{68\times 3}$ denotes the back-projected 3D landmarks, $\hat{L}_j$ are the 3D landmarks from V_{Φ_j}, $d(V_{\Phi_j}, S_j)$ is the mean point-to-plane distance from V_{Φ_j} to its nearest neighbors in scan S_j, and $\mathcal{R}(\Phi_j)$ regularizes FLAME parameters.

Fine tuning. Once the initial alignment has been obtained, we upsample the mesh resolution by a factor of 16 for the face region, and perform non-rigid registration using ARAP [36] for each scan individually.

Let V be the upsampled vertices, which we aim to register to the scan $\mathcal{S}$. We seek vertex-specific offsets $\{\delta_v\}_{v\in V}$, and auxiliary, vertex-specific rotation $\{R_v\}_{v\in V}$ from the ARAP term. Therefore, we solve

$$\underset{\substack{\{\delta_v\}_{v\in V}\\ \{R_v\}_{v\in V}}}{\arg\min}\sum_{v\in V}\Big[d(\hat{v},\mathcal{S})+\sum_{u\in\mathcal{N}_v}\|R(v-u)-(\hat{v}-\hat{u})\|_2^2\Big], \quad (2)$$

using the L-BFGS optimizer, where $\hat{v} = v + \delta_v$, $\mathcal{N}_v$ denotes all neighboring vertices, and $d(\hat{v}, \mathcal{S})$ is as before. See the supplemental for more details.

4. Neural Parametric Head Models

Our neural parametric head model separately represents geometry in a canonical space and facial expression as forward deformations; see Sections 4.1 and 4.2, respectively.

4.1. Identity Representation

We represent a person's identity-specific geometry implicitly in its canonical space as a SDF. Compared to template-mesh-based approaches, this offers the necessary flexibility that is required to model a complete head with hair. In accordance with related work on human body modeling, *e.g*. [9,24,25], we choose a canonical expression with an open mouth to avoid topological issues. While a canonical coordinate system already reduces the dimensionality of the learning problem at hand, we further tailor our neural identity representation to the domain of human heads; as described below.

4.1.1 Local Decomposition

Instead of globally conditioning the SDF network on a specific identity, we exploit the structure of the human face to impose two important geometric priors. First, we embrace the fixed composition of human faces by decomposing the SDF network into an ensemble of several smaller local MLP-based networks, which are defined around certain facial anchors, as shown in Fig. 4. Thereby, we reduce the learning problem into smaller, more tractable ones.We choose facial anchor points as a trade-off between the relevance of an area and spatial uniformity. Second, we exploit the symmetry of the face by only learning SDFs on the left side of the face, which are shared with the right half after flipping spatial coordinates accordingly. More specifically, we divide the face into $K = 2K_{\text{symm}} + K_{\text{middle}}$ regions, which are centered at facial anchor points $\mathbf{a} \in \mathbb{R}^{K\times 3}$. We use $\mathcal{M}$ to denote the index set anchors lying on the symmetry axis, and $\mathcal{S}$ and $\mathcal{S}^*$ for symmetric regions on the left and right side respectively, such that for $k \in \mathcal{S}$ there is a $k^* \in \mathcal{S}^*$ that corresponds to the symmetric anchor point.

In addition to a global latent vector $\mathbf{z}_{\text{glob}} \in \mathbb{R}^{d_{\text{glob}}}$, the k-th region is equipped with a local latent vector $\mathbf{z}^{\text{id}}_k \in \mathbb{R}^{d_{\text{loc}}}$. Together, the k-th region is represented by a small MLP

$$f_k : \mathbb{R}^{d_{\text{glob}}+d_{\text{loc}}+3} \to \mathbb{R} \tag{3}$$

$$(x, \mathbf{z}^{\text{id}}_{\text{glob}}, \mathbf{z}^{\text{id}}_k) \mapsto \text{MLP}_{\theta_k}([x - \mathbf{a}_k, \mathbf{z}^{\text{id}}_{\text{glob}}, \mathbf{z}^{\text{id}}_k]), \tag{4}$$

that predicts SDF values for points $x \in \mathbb{R}^3$, where $[\cdot]$ denotes the concatenation operator.

In order to exploit face symmetry, we share the network parameters and mirror the coordinates for each pair (k, k^*) of symmetric regions:

$$f_{k^*}(x, \mathbf{z}^{\text{id}}_{\text{glob}}, \mathbf{z}^{\text{id}}_{k^*}) := f_k(\text{flip}(x - a_{k^*}), \mathbf{z}^{\text{id}}_{\text{glob}}, \mathbf{z}^{\text{id}}_{k^*}), \tag{5}$$

where $\text{flip}(\cdot)$ represents a flip of the coordinates along the face symmetry axis.

4.1.2 Global Blending

In order to facilitate a decomposition that helps generalization, it is crucial that reliable anchor positions $\mathbf{a}$ are available. To this end, we train a small MLP_{pos} that predicts $\mathbf{a}$ from the global latent $\mathbf{z}^{\text{id}}_{\text{glob}}$.

Since each local SDF focuses on a specific semantic region of the face, as defined by the anchors $\mathbf{a}$, we additionally introduce $f_0(x, \mathbf{z}^{\text{id}}_{\text{glob}}, \mathbf{z}^{\text{id}}_0) = \text{MLP}_0(x, \mathbf{z}^{\text{id}}_{\text{glob}}, \mathbf{z}^{\text{id}}_0)$, which operates in the global coordinate system, hence covering all SDF values far away from any anchor in $\mathbf{a}$. To clarify the notation, we set $a_0 := \mathbf{0} \in \mathbb{R}^3$.

Finally, we blend all local fields f_k into a global field

$$\mathcal{F}_{\text{id}}(x) = \sum_{k=0}^{K} w_k(x, a_k) f_k(x, \mathbf{z}^{\text{id}}_{\text{glob}}, \mathbf{z}^{\text{id}}_k), \tag{6}$$

using Gaussian kernels, similar to [12,48], where

$$w^*_k(x, a_k) = \begin{cases} e^{\frac{-||x-a||_2}{2\sigma}}, & \text{if } k > 0 \\ c, & \text{if k = 0} \end{cases} \tag{7}$$

$$\text{and } w_k(x, a_k) = \frac{w^*_k(x, a_k)}{\sum_{k'} w^*_{k'}(x, a_{k'})} \tag{8}$$

We use a fixed isotropic kernel with standard deviation σ and a constant response c for f_0.

4.2. Expression Representation

In contrast to our local geometry representation, we model expressions only with a globally conditioned deformation field; *e.g*. a smile will effect the cheeks corners of the mouth and eye region. In this context, we define $\mathbf{z}^{\text{ex}} \in \mathbb{R}^{d_{\text{ex}}}$ as a latent expression description. Since such a deformation field is defined in the ambient Euclidean space, it is crucial to additionally condition the deformation network with an identity feature. By imposing an information bottleneck on the latent expression description, the deformation network is then forced to learn a disentangled representation of expressions.

More formally, we model deformations using an MLP

$$\mathcal{F}_{\text{ex}}(x, \mathbf{z}^{\text{ex}}, \hat{\mathbf{z}}^{\text{id}}) : \mathbb{R}^{d_{\text{ex}}+d_{\text{id-ex}}} \to \mathbb{R}^3. \tag{9}$$

Rather than directly feeding all identity information into $\mathcal{F}_{\text{ex}}$ directly, we first project the information to a lower dimensional representation

$$\hat{\mathbf{Z}}^{\text{id}} = W[\mathbf{z}^{\text{id}}_{\text{glob}}, \mathbf{z}^{\text{id}}_0, \dots \mathbf{z}^{\text{id}}_K, \mathbf{a}_1, \dots, \mathbf{a}_K], \tag{10}$$

using a single linear layer W, where $d_{\text{id-ex}}$ denotes the dimensionality of the interdependence of identity and expression.

4.3. Training Strategy

Our training strategy closely follows NPMs [24] and sequentially trains the identity and expression networks in an auto-decoder fashion.

Identity Representation For the identity space, we jointly train latent codes $\mathbf{Z}^{\text{id}}_j := \{\mathbf{z}^{\text{id}}_{\text{glob},j}, \mathbf{z}^{\text{id}}_{0,j}, \dots, \mathbf{z}^{\text{id}}_{K,j}\}$ for each j

in the set of training indices J and network parameters θ_{pos} and $\theta_0, \dots, \theta_K$, by minimizing

$$\mathcal{L}_{\text{id}} = \sum_{j \in J} \mathcal{L}_{\text{IGR}} + \lambda_a \|\hat{\mathbf{a}}_j - \mathbf{a}_j\|_2^2 + \lambda_{\text{sy}} \mathcal{L}_{\text{sy}} + \lambda_{\text{reg}}^{\text{id}} \|\mathbf{Z}_j^{\text{id}}\|_2^2, \quad (11)$$

where $\mathcal{L}_{\text{IGR}}$ is the loss introduced in [15] which enforces SDF values to be zero on the surface and contains an Eikonal term. This ensures consistency between surface normals and SDF gradients and is in similar spirit to [15, 34]. For training, we directly sample points and surface normals from our ground truth scans.

Additionally, we supervise anchor predictions $\mathbf{a}_j$ using the corresponding vertices from our registrations $\hat{\mathbf{a}}_j$. The last two terms serve regularization purposes, where

$$\mathcal{L}_{\text{sy}} = \sum_{k \in \mathcal{S}} \|\mathbf{z}_k^{\text{id}} - \mathbf{z}_{k^*}^{\text{id}}\|_2^2 \quad (12)$$

enforces the local latent description of symmetric regions to be close, and the final term encourages a well-behaved distribution of both global and local latent descriptions centered around zero.

Expression Representation Once the identity representation is learned, we optimize for network parameters θ_{ex}, W and latent expression codes, $\{\mathbf{z}_{j,l}^{\text{ex}}\}_{j \in J, l \in L}$, where j indexes identity and l indexes expressions. The deformation loss

$$\mathcal{L}_{\text{ex}} = \sum_{\substack{i,j \in J,L \\ x \in X_{j,l}}} \|\mathcal{F}_{\text{ex}}(x, \mathbf{z}_{j,l}^{\text{ex}}, \hat{\mathbf{z}}_j^{\text{id}}) - \delta(x)_{j,l}\|_2^2 + \lambda_{\text{reg}}^{ex} \|\mathbf{z}_{j,l}^{\text{ex}}\|_2^2 \quad (13)$$

directly supervises the deformation field using samples $x \in X_{j,l}$, which have been precomputed from the registration. See the supplemental for more details.

5. Results

We aim to evaluate how well our method generalizes from our training dataset of 87 identities to unseen ones, and their unique expressions. Our test dataset consists of 6 female and 12 male identities in 23 expressions each. We fit our model and baselines to frontal single view depth maps, which are generated by rendering the unseen validation meshes and randomly sampling 5000 points. For ablations with respect to the number of points and noise level, as well as for a demonstration of real-world tracking with NPHM using a commodity depth sensor, we refer to the supplementary material. In our evaluation, we isolate the reconstruction of identity and expression in section 5.1 and 5.2, respectively.

Mesh-Based Baselines. We evaluate against the Basel Face Model (BFM) and FLAME as representatives of existing template-based PCA-models. Furthermore, we compare against a PCA model with delta expressions [1] trained on our registered meshes and a local variant thereof. For the local PCA model we utilize the same facial anchors as in NPHM to divide each neutral registered mesh into regions, which are separately represented by local PCA models. To obtain a final prediction we use the same blending scheme as described in Section 4.1.2. For all these models we additionally provide the 68 facial landmarks as input.

Implicit Baselines. We compare against ImFace [45] as a neural backward deformation baseline. We evaluate a version trained on the FaceScape dataset [43] and one that we train on our dataset using their preprocessing (denoted as ImFace*). Additionally, we compare against NPMs [24], isolating the effect of our proposed identity representation.

Metrics. To evaluate the quality of the reconstructions, we report L_1-Chamfer distance, normal consistency (N. C.), and F-Score with a threshold of 1.5mm.

5.1. Identity Reconstruction

To separately evaluate the quality of our identity space, we fit against a single neutral expression scan for each identity. These scans are aligned to each method's canonical coordinate system. We assist baselines that use a closed mouth in their canonical space, i.e. baselines not trained on our data, by optimizing these over all scans instead. More details on the optimization strategy for each model can be found in the supplemental.

Figure 5 and Table 2 present qualitative and quantitative results, respectively. We observe that all neural field methods consistently achieve more faithful reconstructions and further note that the proposed local conditioning allows NPHM to reconstruct details and statistically unlikely elements more reliably.

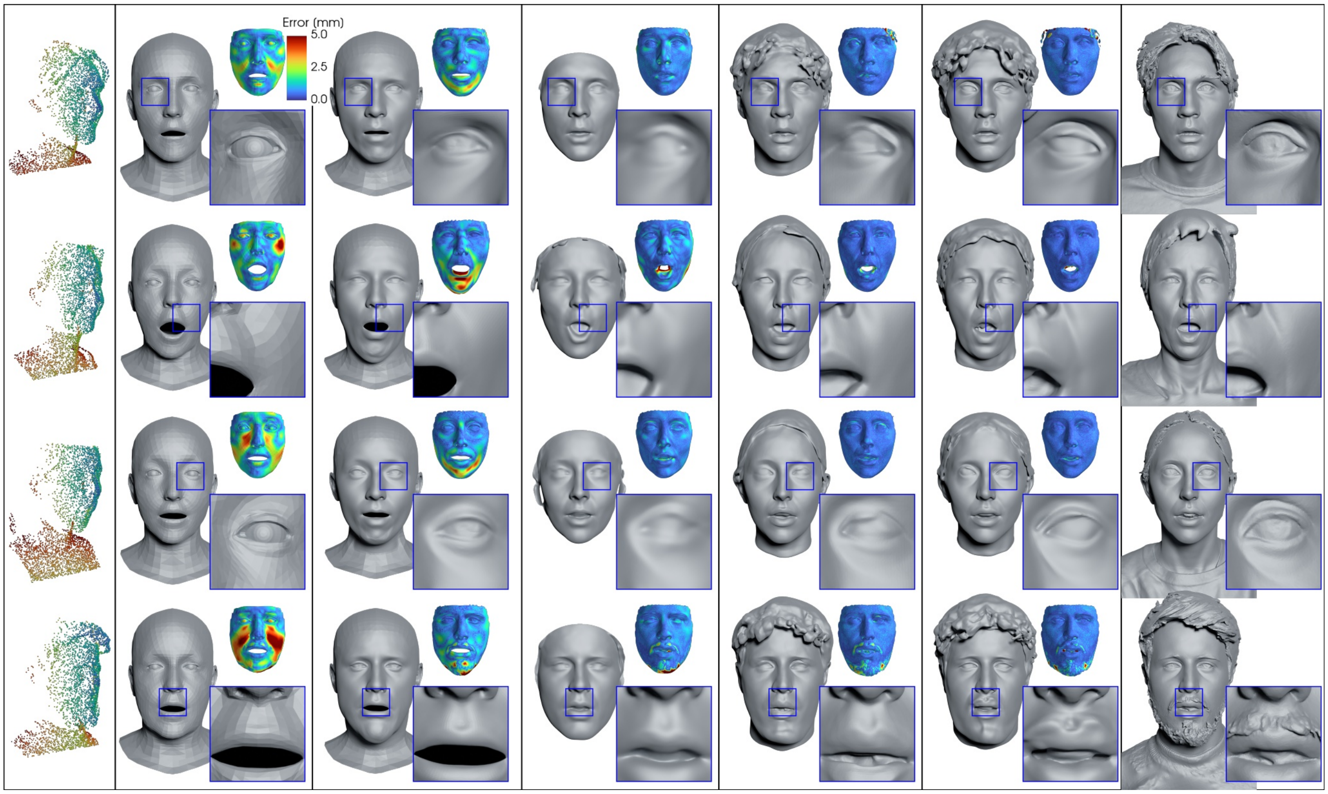
Method	L_1-Chamfer ↓	N. C. ↑	F-Score@1.5 ↑
BFM [29]	1.341e−2	0.936	0.319
FLAME [18]	0.640e−2	0.931	0.530
Global PCA [1]	0.563e−2	0.954	0.571
Local PCA [1]	0.416e−2	0.960	0.756
ImFace [45]	0, 404e−2	0.954	0.832
ImFace* [45]	0.312e−2	0.971	0.883
NPM [24]	0.200e−2	0.975	0.947
Ours	**0.182**e−2	**0.978**	**0.954**

* trained on our data

Table 2. Identity fitting to a single depth map in neutral expression.

5.2. Expression Reconstruction

To evaluate each model's expression space, we fit it to multiple expressions of the same person with the task of recovering one identity code per subject and one expression code per expression. For the neural forward deformation



Input FLAME [18] Local PCA [1] ImFace* [45] NPM [24] Ours GT Scan

Figure 5. Model fitting: at inference time, we fit our model to sparse, partial input point clouds from single depth map. We compare our method to widely-used state-of-the-art parametric face models, including FLAME [18], a local PCA [1], ImFace [45] and neural parametric models (NPM) [24]. Our parametric model has significantly more surface detail and covers the entire head, including the hair region.

models, NPM and NPHM, we utilize iterative root finding [10] to fit the expression codes. For simplicity, we keep the identity code fixed after fitting to the neutral scan. For all other models we jointly solve for expression and identity codes. Figure 6 and Table 3 show qualitative and quantitative comparisons with our baselines, respectively. Owing to the abiility of backward deformations to directly connect the observed with the canonical space, ImFace reliably reconstructs expressions. Nevertheless, it still suffers from blurry reconstructions, compared to both NPM and NPHM.

See our supplemental for more details and an additional comparison of jointly fitting identity and expression when only a single depth observation is available.

Method	L_1-Chamfer ↓	N. C. ↑	F-Score@1.5 ↑
BFM [29]	$1.271\text{e}{-2}$	0.937	0.508
FLAME [18]	$0.679\text{e}{-2}$	0.924	0.351
Global PCA [1]	$0.515\text{e}{-2}$	0.956	0.606
Local PCA [1]	$0.535\text{e}{-2}$	0.950	0.641
ImFace [45]	$0.369\text{e}{-2}$	0.959	0.824
ImFace* [45]	$0.321\text{e}{-2}$	**0.971**	0.879
NPM [24]	$0.299\text{e}{-2}$	0.962	0.891
Ours	$\mathbf{0.272}\text{e}{-2}$	0.969	**0.913**

* trained on our data

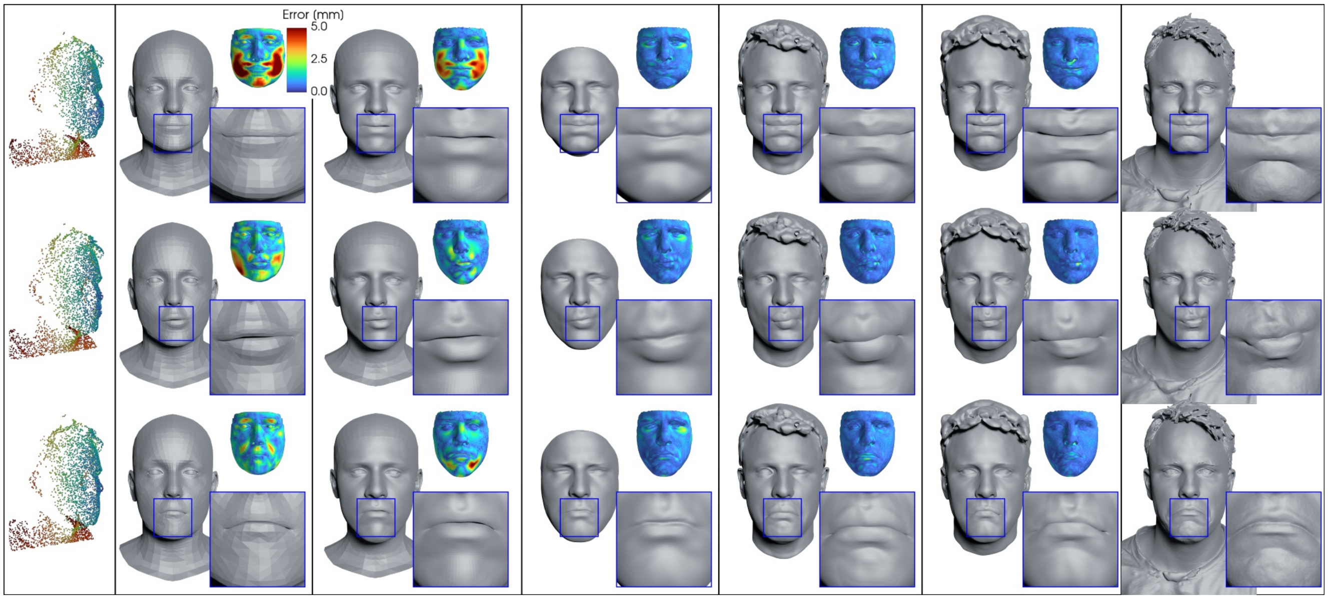
Table 3. Expression fitting on 23 single depth maps per person.

5.3. Ablations

We ablate two main contributions of the proposed identity representation, by fitting identity codes to a neutral scan without involving expressions. First, we analyze the effect of the number of regions K of our ensemble, by comparing against NPM [24], which effectively would be an ensemble of size 1, and against versions with 12 and 26 regions and adjusted number of latent dimensions. Additionally, we confirm the benefit of sharing weights for symmetric keypoints. Table 4 shows a quantitative evaluation of these two ablations supporting our design choices.

5.4. Limitations

In our experiments, we show that NPHM can reconstruct high-quality human heads; however, at the same time, we believe that there are still several limitations and opportunities for future work. For instance we focus solely on the



Input FLAME [18] Local PCA [1] ImFace* [45] NPM [24] Ours GT Scan

Figure 6. Comparison on fitting expressions to sparse input point clouds: from a sparse set of depth observations of different expressions from a frontal view (left), we compare FLAME [18], a local PCA [1], ImFace [45], neural parametric models (NPM) [24], and our method against the respective ground truth scans.

Method	L_1-Chamfer ↓	N. C. ↑	F-Score@1.5 ↑
NPM [24]	0.254	0.972	0.906
K=12, w/ sy.	0.289	0.966	0.876
K=26, w/ sy.	0.237	0.971	0.913
K=39, w/o sy.	0.230	0.974	0.917
Ours	**0.206**	**0.976**	**0.938**

Table 4. Effect of the number of anchor points K and symmetry on identity reconstruction performance. NPM represents the extreme case of using exactly 1 anchor point. Note that to be consistent with the original version, NPM differs to the other models in both width and depth of the underlying MLP.

geometry of heads while omitting any information about appearance. This makes our model ill-suited for fitting to RGB images using dense photometric terms. Here, an interesting future avenue would be to explore learning appearance, anchored on top of the geometric base model. In fact, as part of our dataset we also provide the RGB frames captured during the 3D scanning process, which should facilitate learning such a texture model.

Another limitation is that currently we do not capture open hair, which limits general diversity; however, compared to other existing face models such as 3D morphable models, we significantly expand the application domain by covering the entirety of the human head. In the future, we still would like to cover a broader range of hairstyles.

6. Conclusion

We have introduced neural parametric head models, a neural representation which disentangles identity and expressions of human heads, by representing geometry in canonical space and modelling expressions as forward deformations. For our identity representation we have proposed and validated a local representation that is tailored towards human head. To train our model, we introduce a new dataset of over 3700 high-fidelity 3D scans. Once trained, our model can be fitted to sparse input point clouds, for instance, captured by a commodity range sensor. Compared to existing methods, such as widely-used PCA-based techniques, our model represents significantly more detail while being able to regularize out noise of the underlying point cloud inputs. Overall, we believe that our method is an important step towards high-fidelity face capture and our newly-introduced dataset opens up opportunities to further explore learning priors for neural face models.

Acknowledgements

This work was supported by the ERC Starting Grant Scan2CAD (804724), the German Research Foundation (DFG) Grant “Making Machine Learning on Static and Dynamic 3D Data Practical”, the German Research Foundation (DFG) Research Unit “Learning and Simulation in Visual Computing”, and Synthesia. We would like to thank Maximilian Knörl and Tim Walter for the help with scanning, and Angela Dai for the video voiceover.

References

[1] 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, 6, 7, 8

[2] Timo Bolkart and Stefanie Wuhrer. A groupwise multilinear correspondence optimization for 3d faces. In *Proceedings of the IEEE international conference on computer vision*, pages 3604–3612, 2015. 2

[3] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3d face morphable models" in-the-wild". In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 48–57, 2017. 2

[4] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5543–5552, 2016. 2

[5] Giorgos Bouritsas, Sergiy Bokhnyak, Stylianos Ploumpis, Michael Bronstein, and Stefanos Zafeiriou. Neural 3d morphable models: Spiral convolutional networks for 3d shape representation learning and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7213–7222, 2019. 2

[6] Alan Brunton, Timo Bolkart, and Stefanie Wuhrer. Multilinear wavelets: A statistical shape space for human faces. In *European Conference on Computer Vision*, pages 297–312. Springer, 2014. 2

[7] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. 20(3):413–425, mar 2014. 3

[8] Rohan Chabra, Jan E. Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX*, page 608–625, Berlin, Heidelberg, 2020. Springer-Verlag. 2

[9] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J. Black, Andreas Geiger, and Otmar Hilliges. gdna: Towards generative detailed neural avatars. *CoRR*, abs/2201.04123, 2022. 2, 5

[10] 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, 7

[11] Zhixiang Chen and Tae-Kyun Kim. Learning feature aggregation for deep 3d morphable models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13164–13173, 2021. 2

[12] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4857–4866, 2020. 2, 5

[13] Simon Giebenhain and Bastian Goldluecke. Air-nets: An attention-based framework for locally conditioned implicit representations. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021. 2

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

[15] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 6

[16] Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, and Hao Li. Learning formation of physically-based face attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[17] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. *arXiv preprint arXiv:2206.08929*, 2022. 2

[18] 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, 6, 7, 8

[19] 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. 2

[20] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019. 3

[21] 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

[22] 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

[23] Thomas Neumann, Kiran Varanasi, Stephan Wenger, Markus Wacker, Marcus Magnor, and Christian Theobalt. Sparse localized deformation components. *ACM Trans. Graph.*, 32(6), nov 2013. 2

[24] 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, October 2021. 2, 5, 6, 7, 8

[25] Pablo Palafox, Nikolaos Sarafianos, Tony Tung, and Angela Dai. Spams: Structured implicit parametric models. *CVPR*, 2022. 2, 5

[26] 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

[27] 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

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

[29] 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, 6, 7

[30] 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

[31] Stylianos Ploumpis, Haoyang Wang, Nick Pears, William AP Smith, and Stefanos Zafeiriou. Combining 3d morphable models: A large scale face-and-head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10934–10943, 2019. 2

[32] 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

[33] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 397–403, 2013. 3

[34] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020. 6

[35] Janujah Sivanandan, Eugene Liscio, and P Eng. Assessing structured light 3d scanning using artec eva for injury documentation during autopsy. *J Assoc Crime Scene Reconstr*, 21:5–14, 2017. 3

[36] Olga Sorkine and Marc Alexa. As-Rigid-As-Possible Surface Modeling. In Alexander Belyaev and Michael Garland, editors, *Geometry Processing*. The Eurographics Association, 2007. 4

[37] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1126–1135, 2019. 2

[38] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7346–7355, 2018. 2

[39] Luan Tran and Xiaoming Liu. On learning 3d face morphable model from in-the-wild images. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):157–171, 2019. 2

[40] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04):376–380, 1991. 4

[41] 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*, pages 1–9, 2022. 2

[42] 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)*, June 2022. 2, 3

[43] 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)*, June 2020. 2, 3, 6

[44] 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

[45] 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, 6, 7, 8

[46] 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. 2

[47] 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. *arXiv preprint arXiv:2112.03109*, 2021. 4

[48] 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 (CVPR)*, June 2022. 2, 5

[49] 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. In *Computer graphics forum*, volume 37, pages 523–550. Wiley Online Library, 2018. 1