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

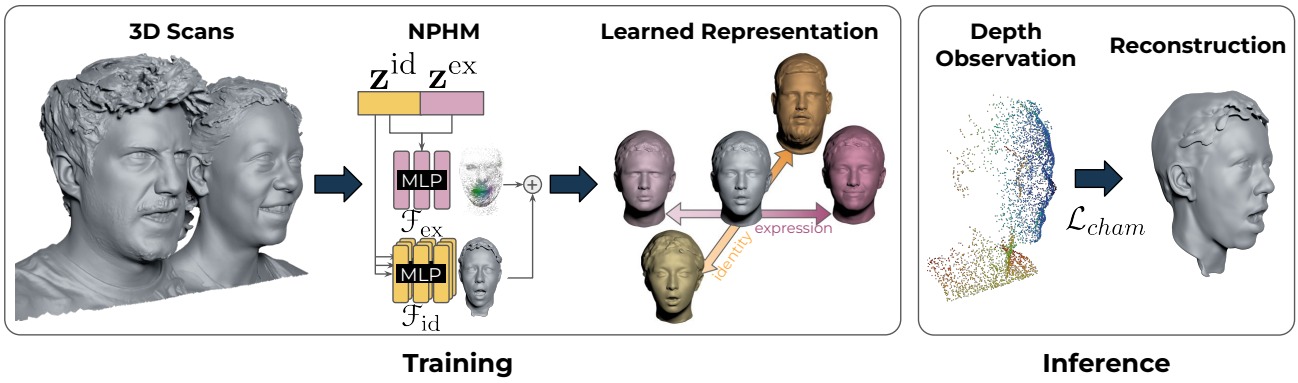

Figure 1. We propose to learn a neural parametric head model based on neural fields: first, we capture a large dataset of over 2200 high-fidelity head scans with varying shapes and expressions (left). We then non-rigidly register these scans to generate our training data. As a result of training, we 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 which 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 2200 head scans from 124 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 by a significant margin 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 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 [46].

Traditionally, these parametric models such as Blendshapes and 3D morphable models (3DMM), are based on a low-rank approximation of the underlying 3D mesh geometry. To this end, a given template mesh with a fixed

Website: https://simongiebenhain.github.io/NPHM

topology is non-rigidly registered to a series of 3D scans of human faces at training time. From this template registration, a parametric model can be computed using dimensionality 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 properly disentangle between shape identities and expression parameters. While these PCA-based models are excellent at regularizing out noise when fitting to noisy input point clouds, their inherent limitation lies in their inability to represent local surface detail and the reliance on a template mesh of fixed topology. As a result, fitted test-time models lack high-frequency surface details and are typically limited to the frontal facial regions, while for instance not including hair regions of the human head.

In this work, we propose neural parametric head models (NPHM), which represent complete human head geometry in a canonical space using a 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 to encode high-frequency geometric detail. 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 poses a strong geometry prior into our model, 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 2200 3D head scans from 124 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 characterizes significantly more detail than state-of-the-art models, representing even fine-scale details.

In sum, our contributions are as follows:

- We introduce a novel 3D dataset captured with a high-end capture rig, including over 2200 3D scans of human heads from 124 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 by a significant margin in terms of fitting accuracy.

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. The model was built upon PCA using 200 face scans, where correspondences were established via optical flow. 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 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 [19].

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 [37–39]. At the same time, graph convolutions [5, 15] 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. [22] implicitly define a surface as the decision boundary of a binary classifier and Mildenhall et al. [23] 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 [18, 20] and heads [44]. Following this paradigm, a number of neural parametric models have been proposed for bodies [9, 24, 25], faces [43], and —most closely related to our work— heads [32, 41, 42]. 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 [42, 43] 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

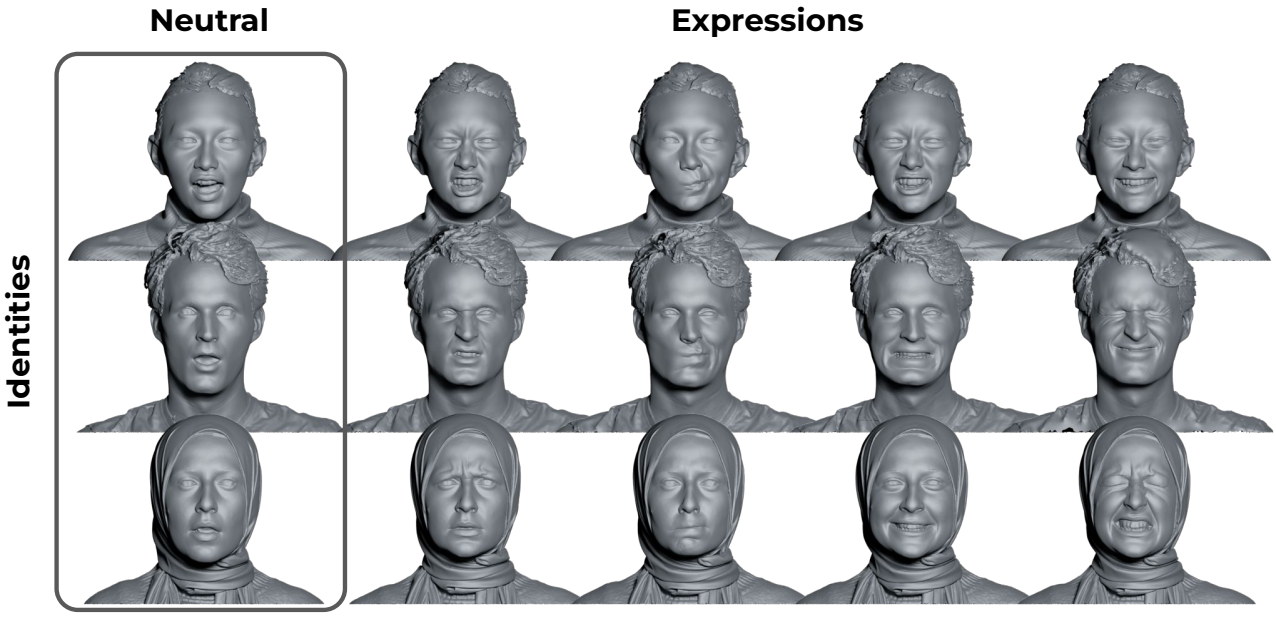

Figure 2. 3D head scans from our newly-captured dataset: for each person (rows), we first capture a neutral pose, followed by several scans in different expressions (columns). We aim to keep expressions consistent across different identities. Overall, our dataset has more than 2200 3D scans from 124 people.

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 [8, 12, 14, 30].

3. Dataset Acquisition

Num. Subjects	124 (90m/34f)
Total num. Scans	2280
Avg. num. Vertices/Faces	$\approx$ 1.5M/3.5M
Total Frames	437.760

Table 1. Statistics of our 3D scanning dataset.

Our dataset comprises 124 subjects, 25% female, and contains over 2200 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 or similar contraption, allowing for the capture of natural hair styles to a certain degree.

3.1. Capture Setup

Our setup is composed of two Artec Eva scanners [35], running the latest software including their upsampling algorithm, 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 reconstruction; each fused scan contains approximately 1.5M vertices and 3.5M triangles. During a capture session, we ask each participant to perform 20 different expressions, which are adopted from the FACS coded expression proposed in FaceWarehouse [7]. Most importantly, we capture a neutral expression with the mouth open, which later serves as canonical pose, as described in Section 4.

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 [19]. To this end, we deploy the MediaPipe [21] 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 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.

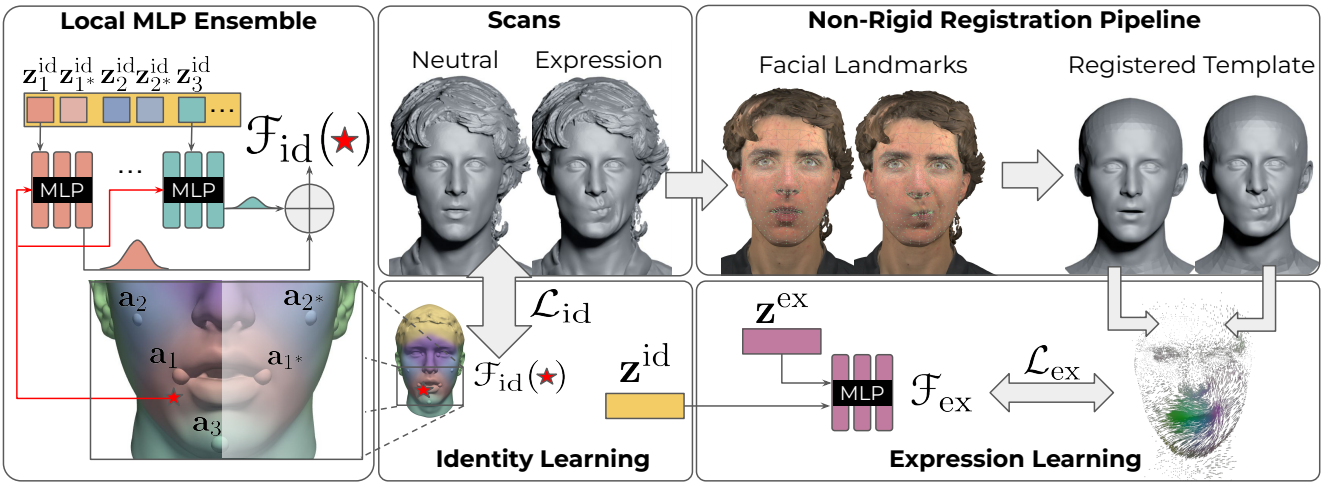

Figure 3. Method overview: at the core of our neural parametric head model lies a neural field representation that parameterizes shape and expressions in disentangled latent spaces. Specifically, we propose a local MLP ensemble that is anchored at face keypoints (left). We train this model by leveraging 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 registrer all scans to a common template (right).

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 [45] to mask out the respective areas of the scans.

Initialization. Given the 20 expression scans $\{S_j\}_{j=1}^{20}$ of a subject, we jointly estimate identity parameters $\mathbf{z}^{\text{id}} \in \mathbb{R}^{100}$, expression parameters $\{\mathbf{z}_j^{\text{ex}}\}_{j=1}^{20}$, and jaw poses $\{\theta_j\}_{j=1}^{20}$ of the FLAME model, as well as a shared scale $s \in \mathbb{R}$ and per-scan rotation and translation corrections $\{R_j\}_{j=1}^{20}$ and $\{t_j\}_{j=1}^{20}$. 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 $V(\Phi_j)$ its vertices. We jointly optimize for these parameters by minimizing

$$\underset{\Phi_1,\dots\Phi_{20}}{\arg\min} \sum_{j=1}^{20} \left[\|L_j - \hat{L}_j\|_1 + \lambda_d \cdot \sum_{v \in V(\Phi_j)} d(v, \mathcal{S}_j) + \mathcal{R}(\Phi_j) \right], \tag{1}$$

where $L_j \in \mathbb{R}^3$ denotes the back-projected 3D landmarks, $\hat{L}_j$ are the 3D landmarks from $V(\Phi_j)$, and $d(v, S_j)$ is the point-to-plane distance from v to its nearest neighbor in scan S_j. We refer to the supplemental for details of the regularization term $\mathcal{R}(\Phi)$.

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

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

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; see 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. 3. Thereby, we reduce the learning problem into smaller, more tractable ones, *e.g.* a network specialized on corners of an eye can generalize faster and with more detail, than a global one. 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 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 $\mathbf{z}_k^{\text{id}} \in \mathbb{R}^{d_{\text{loc}}}$. Together, the k-th region is represented by a small MLP

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

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

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:

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

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 $\mathbf{z}_{\text{glob}}^{\text{id}}$.

Since each local SDF focuses on a specific semantic region of the face, as defined by the anchors $\mathbf{a}$, we additionally introduce $f_0(x, \mathbf{z}_{\text{glob}}^{\text{id}}, \mathbf{z}_0^{\text{id}}) = \text{MLP}_0(x, \mathbf{z}_{\text{glob}}^{\text{id}}, \mathbf{z}_0^{\text{id}})$, 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 $a_0 := \mathbf{0} \in \mathbb{R}^3$.

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

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

using Gaussian kernels, similar to [12], 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

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

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 $\mathbf{Z}_j^{\text{id}} := \{\mathbf{z}_{\text{glob},j}^{\text{id}}, \mathbf{z}_{0,j}^{\text{id}}, \dots, \mathbf{z}_{K,j}^{\text{id}}\}$ for each j in the set of training indices J and network parameters θ_{pos} and $\theta_0, \dots, \theta_K$, by minimizing

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

where $\mathcal{L}_{\text{IGR}}$ is the loss introduced in [16] 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 [16, 34]. For training, we directly sample points and surface normals from our ground truth scans.

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

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

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, $\{\mathbf{z}^{\text{ex}}_{j,l}\}_{j \in J, l \in L}$, where j indexes identity and l indexes expressions. The deformation loss

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

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

5.1. Single-View Depth Map Reconstruction

In this experiment, we evaluate how well our method generalizes from our training dataset of 87 identities to new ones, and their unique expressions. Our test dataset consists of 6 female and 12 male heads. We fit our model to frontal single view depth maps, which are generated by rendering the unseen validation meshes. For ablations with respect to the number of points and noise level, we refer to our supplemental. In our evaluation, we isolate the reconstruction of identity and expression. The respective experiments are described in the following. We evaluate against the Basel Face Model (BFM) and FLAME as representatives of PCA-based approaches, and against NPMs as representative of neural field based morphable models. For the former two, we additionally provide the 68 facial landmarks that we obtained in our registration process, as described in 3.2.

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.

Identity Reconstruction. To evaluate the quality of our identity space, we fit against a single neutral expression for each identity, which we assume to be aligned in our canonical coordinate system. For the PCA-based baselines, we found an ICP loss in combination with a landmark reconstruction term to work best, while for our method, we minimize

$$\sum_{x \in X} |\mathcal{F}_{\text{id}}(x)| + \lambda^{\text{fit}}_{\text{glob}} \|\mathbf{z}^{\text{id}}_{\text{glob}}\|_2^2 + \lambda^{\text{fit}}_{\text{loc}} \sum_{k=1}^{K} \|\mathbf{z}^{\text{id}}_k\|_2^2 + \lambda^{\text{fit}}_{\text{sy}} \mathcal{L}_{\text{sy}}, \tag{14}$$

where X is the observed point cloud. For NPMs we simply omit the symmetry regularization $\mathcal{L}_{\text{sy}}$ local regularizer.

Figure 4 and Table 2 present qualitative and quantitative results, respectively. We observe that both neural field methods achieve much more accurate reconstructions. We further argue that our local conditioning allows us to model details better and capture statistically unlikely elements more reliably, *e.g.* see the beard of the second identity in Figure 4.

Method	L_1-Chamfer ↓	N. C. ↑	F-Score@1.5 ↑
BFM [29]	1.341e−2	0.936	0.319
FLAME [19]	0.640e−2	0.931	0.530
NPM [24]	0.257e−2	0.972	0.906
Ours	**0.204**e−2	**0.978**	**0.942**

Table 2. Identity fitting evaluation from a single depth map.

Expression Reconstruction. After fitting the identity, we optimize for all expression parameters, assuming the identity to be known. This time, the ICP loss was used for all methods; for more details, we refer to the supplementary material. Figure 5 and Table 3 show qualitative and quantitative comparisons with our baselines, respectively.

Method	L_1-Chamfer ↓	N. C. ↑	F-Score @ 1.5 ↑
BFM [29]	1.271e−2	0.937	0.508
FLAME [19]	0.679e−2	0.924	0.351
NPM [24]	0.360e−2	0.961	0.831
Ours	**0.350**e−2	**0.962**	**0.856**

Table 3. Expression fitting performance from a single depth map and initial neutral pose.

5.2. Real-World Tracking

Additionally, we evaluate our model in a real-world face tracking scenario. For this purpose, we fit our model against a depth video captured with a Kinect Azure, a commodity depth sensor. Figure 6 shows our results of a single frame and a comparison to the FLAME model. For details on our tracking approach and the full video, we refer to the supplemental.

5.3. Ablations

We ablate two main contributions of the proposed identity representation. First, we analyze the effect of the number of regions K in our local ensemble of SDFs, 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. Note that a much lower number of K, likely requires further architectural changes, i.e., deeper MLPs, since the regions that have

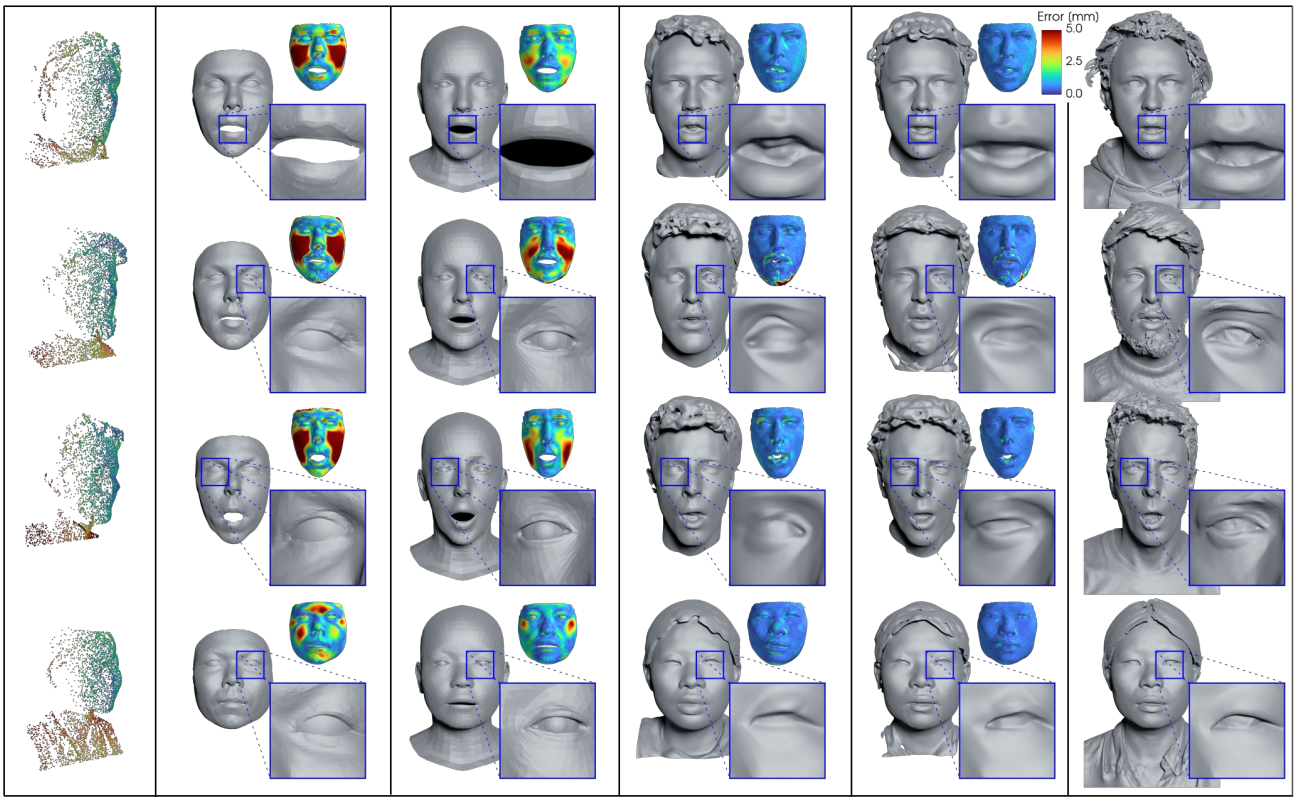

Figure 4. Model fitting: at inference time, we fit our model to sparse, partial input point clouds from single depth maps. We compare our method to widely-used state-of-the-art parametric face models, including the Basel face model (BMF) [13], the FLAME model [19], and neural parametric models (NPM) [24]. Note that our parametric model has significantly more surface detail and covers the entire human head, including the hair region.

to be covered are larger. Additionally, we confirm the benefit of sharing weights for symmetric keypoints by running experiments with and without symmetry constraints. Table 4 shows a quantitative evaluation of these two ablations supporting our design choices.

Method	L_1-Chamfer ↓	N. C. ↑	F-Score@1.5 ↑
NPM [24]	0.257	0.972	0.906
K=12, w/ sy.	0.286	0.968	0.879
K=26, w/ sy.	0.234	0.973	0.917
K=39, w/o sy.	0.227	0.976	0.921
Ours	**0.204**	**0.978**	**0.942**

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.

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 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.

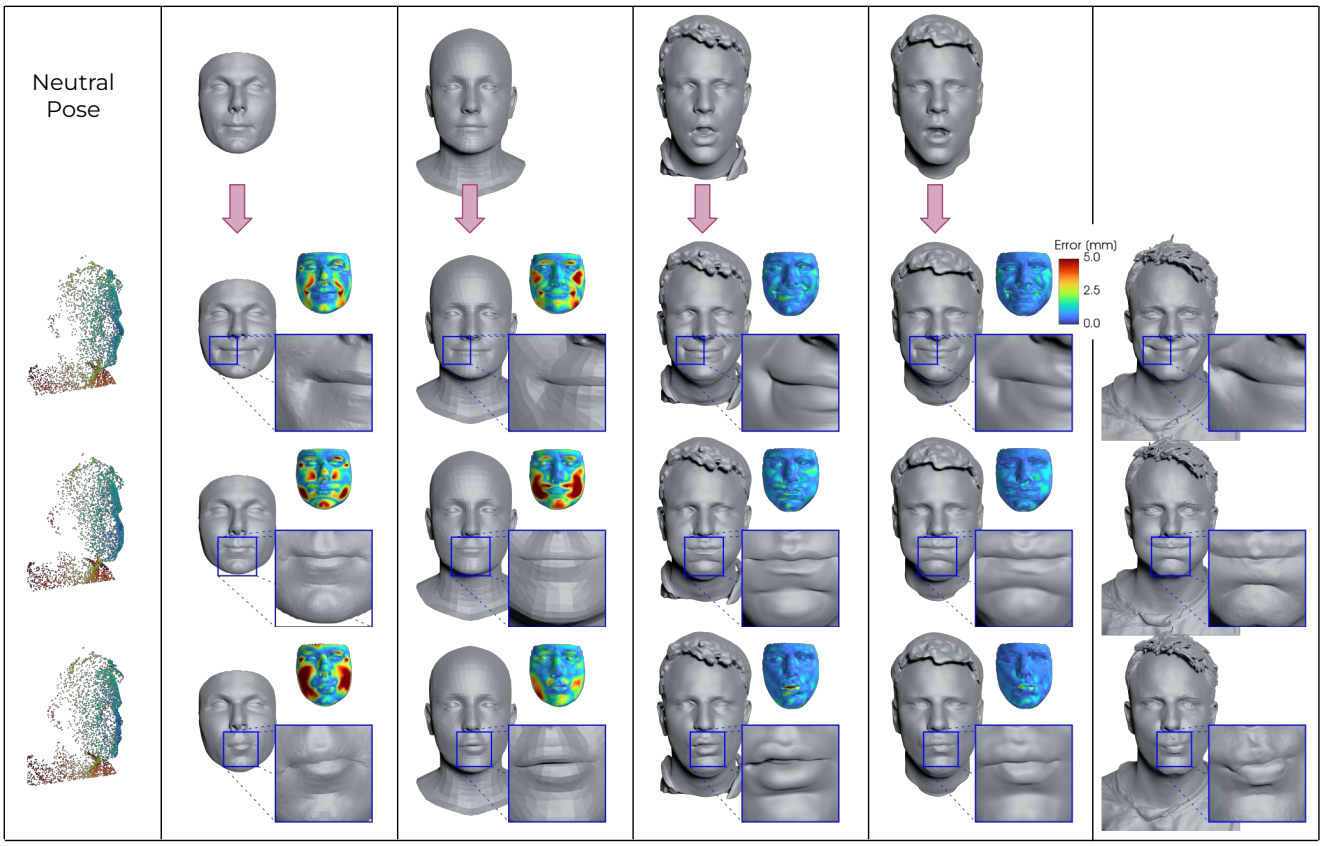

Figure 5. Comparison on fitting expressions to sparse input point clouds: from a sparse set of depth observations from a frontal view (left), we compare against the Basel face model (BFM) [13], the FLAME model [19], neural parametric models (NPM) [24], and our method against the respective ground truth scans. Note that our model is able to reconstruct significantly more surface detail than the baselines.

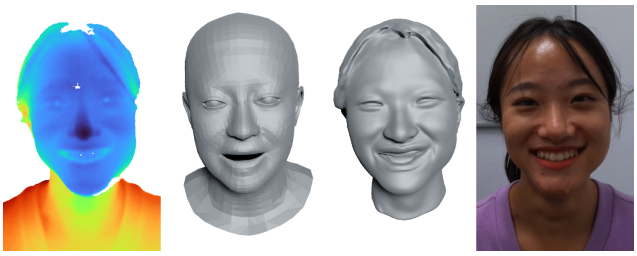

Figure 6. Real-world tracking. For a single frame we show from left to right: the depth map obtained from a commodity depth sensor, FLAME and our reconstructions, and an image as reference.

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 2200 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", and the German Research Foundation (DFG) Research Unit "Learning and Simulation in Visual Computing". We would like to thank Maximilian Knörl and Tim Walter for the help with scanning, and Angela Dai for the video voice over.

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

[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, 11

[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. 3

[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, 4

[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

[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. 3, 5

[13] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018. 7, 8, 15

[14] 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. 3

[15] 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

[16] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 5, 13

[17] 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. 12

[18] 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

[19] 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, 15

[20] 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

[21] 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

[22] 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

[23] 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

[24] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. *arXiv preprint arXiv:2104.00702*, 2021. 2, 4, 5, 6, 7, 8, 13, 14

[25] Pablo Palafox, Nikolaos Sarafianos, Tony Tung, and Angela Dai. Spams: Structured implicit parametric models. *CVPR*, 2022. 2, 4

[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, 13

[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

[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. 3

[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. 5, 13

[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, 12

[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. 3, 15

[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] 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, 15

[43] 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, 15

[44] 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

[45] 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

[46] 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

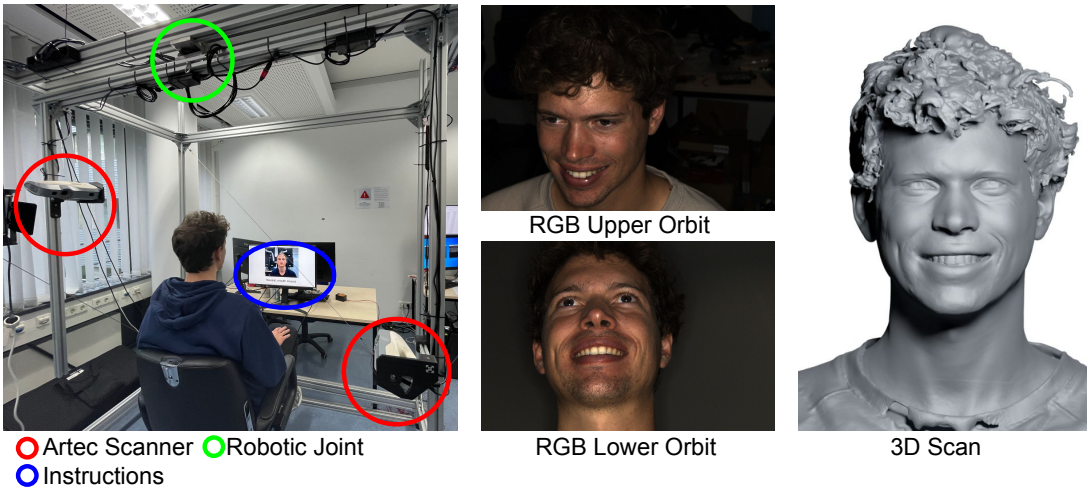

Figure 7. Our custom capture setup (left). Participants are seated on a height-adjustable chair. Two Artec Eva scanners are rotated around the head using a robotic actuator. A screen presents instructions for the 20 different expressions to be performed. Next to the 3D scans (right), the scanners also capture 1.3 megapixel RGB images (middle).

Appendix

Our supplementary material is structured as follows: In Section A, we provide additional details about our capture set-up and dataset. Section B describes additional details of our non-rigid registration, training and hyperparameters. In Section C, we conduct additional ablation experiments, with respect to the quality of the observed depth map. Finally, Section D provides details about our depth map fitting and tracking approach.

For additional visual results we refer to our supplemental video and project page[2]. All of our code and data will be available for research purposes.

A. Dataset

High quality data is of fundamental importance for every learning algorithm. We therefore decided to capture a high quality dataset of 3D head scans. In the following, we provide details about our custom capture set-up and the dataset.

For more samples of our dataset, we refer to Figure 14.

A.1. Capture Set-Up

Figure 7 shows our custom capture set-up, which is built inside of an aluminium cube with an edge length of two meters. We use a robotic actuator[3] to rotate an inverted U-shape around a participant's head.

We place two Artec Eva scanners opposite of each other, with complementary viewing angles on the ends of the inverted U-shape. The height and angles of the scanners are adjusted to obtain an optimal coverage, while avoiding extreme step angles which decrease scanning accuracy.

A.2. Details

During the six seconds of a 360° rotation, each scanner roughly produces 95 frames. Each frame captures range measurements obtained by analyzing a structured light projection using a stereo camera pair. Additionally, a third camera captures RGB images every fifth frame, as depicted in Figure 7. Note that we currently do not use the captured RGB input, except for face landmark detection.

We process the individual 3D measurements of each frame using the provided software of Artec. First, we align the individual frames of the upper and lower scanner using a global registration algorithm. The individual frames are then fused into a single 3D mesh. Subsequently, we further use a hole filling algorithm and remove disconnected parts, for simplicity. However, the unprocessed fused meshes we be additionally released.

The RGB data, including camera parameters and poses will be released alongside the captured 3D scans. We believe that the raw images can be of value to the community and future research projects, e.g. for creating textured 3DMMs or (multi) image reconstruction tasks.

A.3. Expressions

As mentioned in the main paper, our 20 facial expressions are adapted from FaceWarehouse [7]. We illustrate the different expression that we capture in figure 15. As mentioned before the neutral, open mouthed expressions is of special importance, since it serves as our canonical ex-

[2] https://simongiebenhain.github.io/NPHM

[3] We use an acuator of the TUAKA series of Sumitomot Drive Technologies: https://us.sumitomodrive.com/en-us/actuators

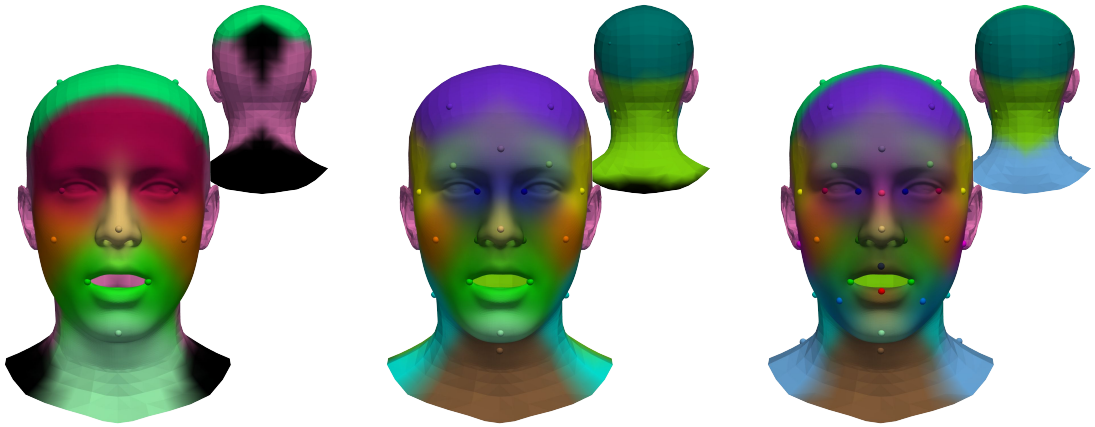

Figure 8. Anchor Layout: Each anchor is assigned a unique color, except for symmetric pairs which share colors. We calculate vertex colors by blending in the same fashion, as for the ensemble of local MLPs. Consequently, the colors show the influence that each local MLP has on its surrounding. Black denotes the color of f_0. Anchor points were chosen as vertices of the average over all registrations.

pression.

A.4. GDPR

Due to the sensitivity of the captured data, all participants in our dataset signed an agreement form compliant with GDPR. Please note that GDPR compliance includes the right for every participant to request the timely deletion of their data. We will enforce these rights in the distribution of our dataset.

B. Implementation Details

We implement our approach – including registration, training, and inference – in PyTorch and, unless otherwise mentioned, run all heavy computations on the GPU, for which we use an Nvidia GTX 3090.

B.1. Non-Rigid Registration

In Equations 1 and 2 of the main paper, we use the point-to-plane distance $d(v, \mathcal{S})$ from a point $v \in \mathbb{R}^3$ to a surface $\mathcal{S} \subset \mathbb{R}^3$. To make our energy terms more robust, we filter this distance based on a distance δ_d and normal threshold δ_n, such that

$$d^*(v, \mathcal{S}) = \begin{cases} 0, & \text{if } d(v, \mathcal{S}) > \delta_d, \\ 0, & \text{if } \langle n(v), n(s) \rangle > \delta_n, \\ d(v, \mathcal{S}), & \text{otherwise,} \end{cases} \tag{15}$$

where

$$d(v, \mathcal{S}) = \min_{s \in \mathcal{S}} |\langle v - s, n(s) \rangle| \tag{16}$$

is the unfiltered point to plane distance and $n(v)$ and $n(s)$ denote the vertex normals of v in the template mesh and the normals of its nearest neighbor in the target $\mathcal{S}$, respectively.

FLAME Fitting We regularize our optimization in FLAME parameter space using

$$\mathcal{R}(\Phi_j) = \lambda_{\text{id}} \frac{\|\mathbf{z}^{\text{id}}\|_2^2}{20} + \lambda_{\text{ex}} \|\mathbf{z}^{\text{ex}_j}\|_2^2 + \lambda_{\text{jaw}} \|\theta_j\|_2^2 + \lambda_{\text{rigid}} (\|R_j\|_2^2 + \|t_j\|_2^2). \tag{17}$$

We use $\lambda_{\text{id}} = 1/5000$, $\lambda_{\text{ex}} = 1/3000$ to regularize the identity and expression parameters respectively. For the jaw angle and the rigid parameters we regularize with $\lambda_{\text{jaw}} = 1/10$ and $\lambda_{\text{rigid}} = 1/10$. Since the point to plane distance initially gives an unreliable signal, despite our filtering we down-weight the point to plane distance with $\lambda_d = 1/15$ for the first 300 iterations. For all remaining one of the 2000 iterations we set $\lambda_d = 1$. We solve Equation 1 using the Adam [17] optimizer with a learning rate of $4e^{-3}$, which is decayed by a factor of 5 for the final 500 iterations.

Finetuning We exponentially decay the weight λ_{ARAP} of the ARAP [36] term with a factor of 0.99. We start with $\lambda_{\text{ARAP}} = 10.0$, but do not decay below $\lambda_{\text{ARAP}} = 0.1$. On average our unoptimized implementation converges after 400-500 iterations of the L-BFGS optimizer and takes roughly 4 minutes on a single GPU.

B.2. Data Preparation and Training

Identity Training To train $\mathcal{F}_{\text{id}}$, we use the loss

$$\mathcal{L}_{\text{IGR}} = \sum_{x \in \delta X} \lambda_s |\mathcal{F}_{\text{id}}(x)| + \lambda_s \left(1 - \langle \nabla \mathcal{F}_{\text{id}}(x), n(x) \rangle\right) \quad (18)$$
$$+ \sum_{x \in X \cup \delta X} \lambda_{\text{eik}} (\|\nabla \mathcal{F}_{\text{id}(x)}|_2 - 1) + \lambda_0 \sum_{x \in X} \exp(-\alpha |\mathcal{F}_{\text{id}}(x)|)$$

introduced in [16] and [34], where we omit the conditioning of $\mathcal{F}_{\text{id}}$ for simplicity. Here δX denotes samples on the surface and X denotes samples in space. We choose $\lambda_s = 2$, $\lambda_n = 0.3$, $\lambda_{\text{eik}} = 0.1$ and $\lambda_0 = 0.01$. For the additional hyperparameters mentioned in Equation 11 we set $\lambda^{\text{id}}_{\text{reg}} = 0.005$, $\lambda_a = 7.5$ and $\lambda_{\text{sy}} = 0.005$.

Furthermore, we train for $15,000$ epochs with a learning rate of 0.0005 and 0.001 for the network parameters and latent codes, respectively. Both learning rates are decayed by a factor of 0.5 every $3,000$ epochs. We use a batch size of 16 and $|\delta X| = 500$ points sampled on the surface. Samples X are obtained by adding Gaussian noise with $\sigma = 0.01$ to surface points and adding some points sampled uniformly in a bounding box. Additionally, we use gradient clipping with a cut-off value of 0.1 and weight decay with a factor of 0.01.

Since this loss only requires samples on the surface directly, we precompute $2,000,000$ points sampled uniformly on the surface of the 3D scans, after removing the lower part of the scan, which we determine using a plane spanned by three vertices on the neck of our registered template mesh. Since our focus lies on the front part of the face, 80% of these points are sampled on the front and 20% on the back and neck. The frontal area is determined by a region on our registered meshes, which covers the face, ears, and forehead. We additionally sample surface normals.

Training the identity network takes about 12 hours until convergence on a single GPU.

Expression Training For the training of $\mathcal{F}_{\text{ex}}$, we follow NPMs [24] and precompute samples of the deformation field, which can be used for direct supervision of $\mathcal{F}_{\text{ex}}$.

More specifically, let $\mathcal{M}$ and $\mathcal{M}'$ be a neutral and expression scan. For a point $x \in \mathcal{M}$, we determine the corresponding point $x' \in \mathcal{M}'$ using barycentric coordinates and construct samples of the deformation $\delta(x) = x' - x$. While strictly speaking the deformation is only defined for points on the surface, we compute field values close to the surface by offsetting along the normal direction, *i.e.* $\delta(x + \alpha n(x)) = x' + \alpha n(x') - (x + \alpha n(x))$, where we sample $\alpha \sim \mathcal{N}(0, \tau_i \mathbb{I}_3)$ twice with standard deviations $\tau_1 = 0.02$ and $\tau_2 = 0.004$. Overall, we sample $2,000,000$ points per expression.

For the expression training we use $\lambda^{\text{ex}}_{\text{reg}} = 5e^{-5}$ and a learning rate of $5e^{-4}$ and $1e^{-3}$ for the network and latent codes, respectively. We train for $2,000$ epochs with a learning rate decay of 0.5 every 600 epochs, gradient clipping at 0.025 and weight decay strength $5e^{-4}$. We use 1000 samples to compute $\mathcal{L}_{\text{ex}}$ and a batch size of 32.

Training the expression network until convergence takes about 8 hours on a single GPU.

B.3. Architectural Details

B.3.1 NPMs

In the main paper, we compare our proposed method against our implementation of NPMs [24]. Our method replaces the identity network $\mathcal{F}_{\text{id}}$. Instead of $\mathcal{F}_{\text{id}}$, NPMs uses the original architecture of DeepSDF [26] with 8 layers, a hidden dimensionality of 1024 and $\mathbf{Z}_{\text{id}} = 512$ dimensions for the latent vector.

The expression latent dimension is $d_{\text{ex}} = 200$ and the MLP has 6 hidden layers with 512 hidden units.

B.3.2 NPHMs

Our default choice for the number of anchor points is $K = 39$, of which $K_{\text{symm}} = 16$ are symmetric. This leads to 7 anchor points lying directly on the symmetry axis, and hence parameters of $16 + 7 = 23$ local DeepSDFs have to be optimized. Figure 8 depicts the arrangement of the anchor points.

The identity latent space is composed of the shared global part $\mathbf{z}^{\text{id}}_{\text{glob}} \in \mathbb{R}^{d_{\text{glob}}}$ with $d_{\text{glob}} = 64$ and local latent vectors $\mathbf{z}^{\text{id}}_k \in \mathbb{R}^{d_{\text{loc}}}$ with $d_{\text{loc}} = 32$. Our local MLPs have 4 hidden layers with 200 hideen units each and follow the DeepSDF [26] architecture. Note that the total number of latent identity dimensions $d_{\text{id}} = (K + 1) * d_{\text{loc}} + d_{\text{glob}} = 1344$.

Furthermore, we use $\sigma = 0.1$ and $c = e^{-0.2/\sigma^2}$ to blend the ensemble of local MLPs. Figure 8 illustrates the resulting influence that the individual local MLPs have on the final prediction.

Anchor Points In the main paper, we ablated the number of face anchor points. Figure 8 shows a comparison of the different anchor layouts that we ablated. For a lower number of anchors, we increase d_{loc} such that d_{id} is roughly preserved.

For the ablation of our symmetry prior, we keep the exact same anchor layout; however, do not share network weights for symmetric anchors and do not mirror coordinates.

B.4. Evaluation

Since we quantitatively compare models that represent vastly different regions of the human head, we restrict the calculations of our metrics to the face region. This also aligns with the fact, that each model only observes a single, frontal depth map, i.e. other parts of the head can only be estimated roughly.

To this end, we determine the facial area by all points which are closer than 1cm to a region defined on our registered template mesh. Within this region we sample 1,000,000 points with their corresponding normals on the ground truth as well as on each reconstruction. Using these sampled points and normals, we compute all of our metrics.

C. Additional Ablations

The experiments in the main paper were restricted to single view depth maps with 5000 points. Here, we present a thorough evaluation with respect to the number of input points and with respect to artificial Gaussian noise.

Number of Points: Figure 9 shows how the number of observed points effect the reconstructions quantitatively. We evaluate on 250, 500, 1000, 2500, 5000, and 10000 points, respectively. Figure 12 illustrates the effect qualitatively.

Noise: Similarly, we ablate against additive Gaussian noise with standard deviations of 0.0mm, 0.3mm, 0.75mm and 1.5mm. Quantitative and qualitative results are presented in Figures 10 and 11, respectively.

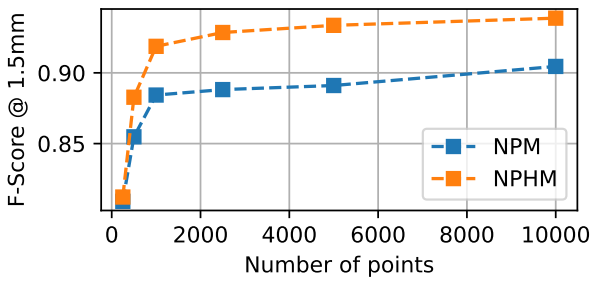

Figure 9. Ablation with respect to the number of points in the input point cloud.

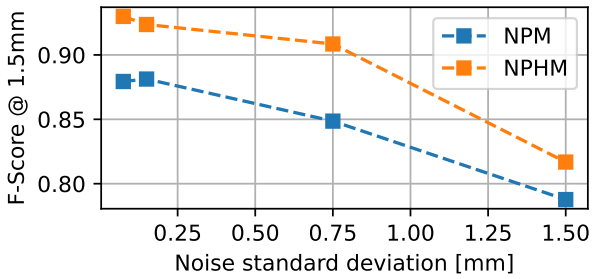

Figure 10. Robustness of our method to noise in the input point cloud.

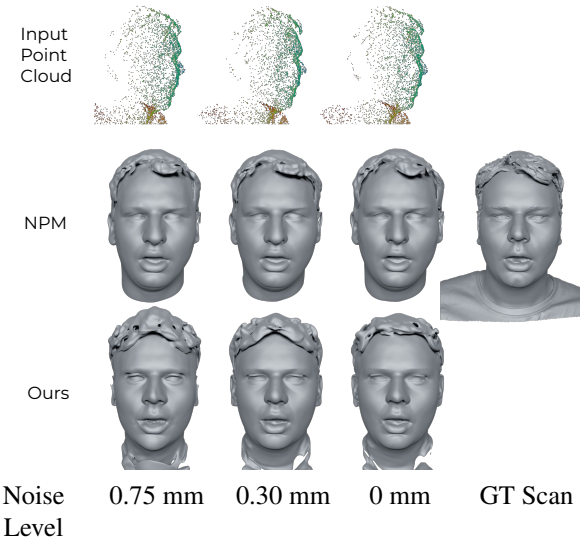

Figure 11. Qualitative comparison of NPMs [24] and our method with respect to noise in the input point cloud. We perturb the points by applying random Gaussian noise with different standard deviations.

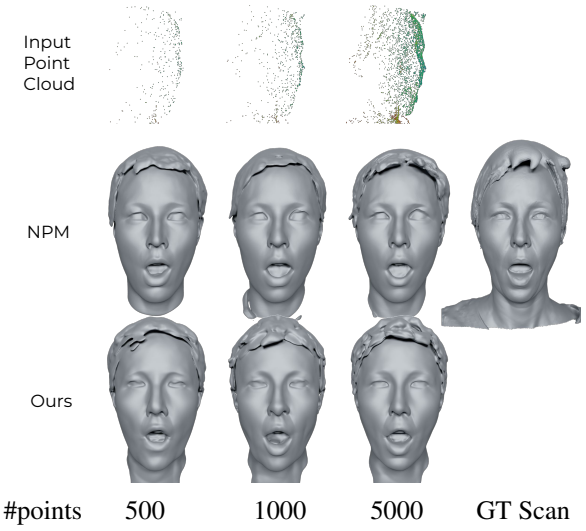

Figure 12. Qualitative comparison of NPMs [24] and our method with respect to the number of points in the input point cloud.

C.1. Deformation Consistency

Furthermore, we illustrate the behaviour of our expression network $\mathcal{F}_{\text{ex}}$ in figure 13, by assigning a distinctive UV-map as colors to each vertex. To be more specific, we assign vertex colors by projecting a UV-map parallel to the "depth-dimension". We then fix vertex colors and deform the mesh using $\mathcal{F}_{\text{ex}}$. The results show that semantic consistency is preserved well, which is a direct consequence of our train-

ing strategy. i3DMMs [42] and ImFace [43] exhibit fewer consistent correspondences since they model backward deformations and do not rely on direct supervision from deformations.

D. Fitting

D.1. Mesh-Based Models

For both BFM [13] and FLAME [19], we optimize Equation 1, *i.e.* we resort to facial landmarks and jointly optimize identity and expression parameters over multiple expressions.

D.2. Field-Based Model, Identity Fitting

For our field based methods we optimize Equation 14 directly for identity parameters using the Adam optimizer for 400 iterations. The optimization procedure starts with a learning rate of 0.01 and is decayed by a factor of 5 after epochs 150, 300 and 350.

NPHM For our model we use $\lambda^{\text{fit}}_{\text{glob}} = 0.004$ and $\lambda^{\text{fit}}_{\text{loc}} = 0.01$ to regularize the global and local identity components. Additionally, we encourage symmetry $\lambda^{\text{fit}}_{\text{sy}} = 1.0$ for the first half of iterations and then set $\lambda^{\text{fit}}_{\text{sy}} = 0.0$.

NPM We use the exact same hyperparameters as for our model. However, the local regularization and symmetry prior have no effect.

D.3. Field-Based Model, Expression Fitting

After fitting the identity code $\mathbf{z}^{\text{id}}$ of a person from a neutral scan we optimize an ICP-style loss for expression parameters

$$\arg\min_{\mathbf{z}^{\text{ex}}} \sum_{x \in \mathcal{S}} \left| \mathcal{F}_{\text{ex}}(x, \mathbf{z}^{\text{ex}}, \mathbf{z}^{\text{id}}) \right| + \lambda^{\text{fit}}_{\text{ex}} \|\mathbf{z}^{\text{ex}}\|_2^2, \tag{19}$$

where $\mathcal{S}$ are $100,000$ points sampled uniformly on the surface $\{x \in \mathbb{R}^3 : \mathcal{F}_{\text{id}}(x, \mathbf{z}^{\text{id}}) = 0\}$, which we extract using marching cubes. Here we use $\lambda^{\text{fit}}_{\text{ex}} = 0.005$ and an initial learning rate of 0.001 for the Adam optimizer. We optimize for 400 epochs and decay the learning rate by a factor of 10 after epochs 200 and 300.

D.4. Tracking

For our tracking results on a commodity depth sensor, we include a total variation prior along the temporal axis over estimated head pose and expression parameters. More specifically, we add

$$\mathcal{L}_{TV}(\phi) = \sum_{t=1}^{T} \|\phi(t+1) - \phi(t)\| \tag{20}$$

to the respective optimization problems, where $\phi(t)$ denotes any of the time dependent optimization parameters, *i.e.* expression and pose.

Otherwise, we follow the same strategy as before, *i.e.* optimize for identity on a single frame and then optimize pose and expression parameters for each time step. To align the coordinates system of the back-projected depth map into our canonical coordinate system, we calculate the similarity transform using [40] from detected landmarks to the the landmarks of the average FLAME face.

To stabilize the optimization, we also include landmarks at the mouth and eye corners, as well as on the top and bottom of the lips, which we denote as $\mathbf{a}_t \in \mathbb{R}^{8\times 3}$ for each time step.

For the identity fitting on a chosen frame t_{can}, the landmarks serve as additional supervision for $\mathbf{z}^{\text{id}}_{\text{glob}}$ including the term

$$\|\text{MLP}_{\text{pos}}(\mathbf{z}^{\text{id}}_{\text{glob}}) - \mathbf{a}_{t_{\text{can}}}\|_1.$$

In this stage, we additionally estimate normals using a Sobel filter and use them as additional supervision signal, as in Equation 18.

During expression fitting, we instead use

$$\sum_{t=1}^{T} \|\mathcal{F}_{\text{ex}}(\text{MLP}_{\text{pos}}(\mathbf{z}^{\text{id}}_{\text{glob}}), \mathbf{z}^{\text{ex}}_t, \mathbf{z}^{\text{id}}_{\text{glob}}) - \mathbf{a}_t\|_1. \tag{21}$$

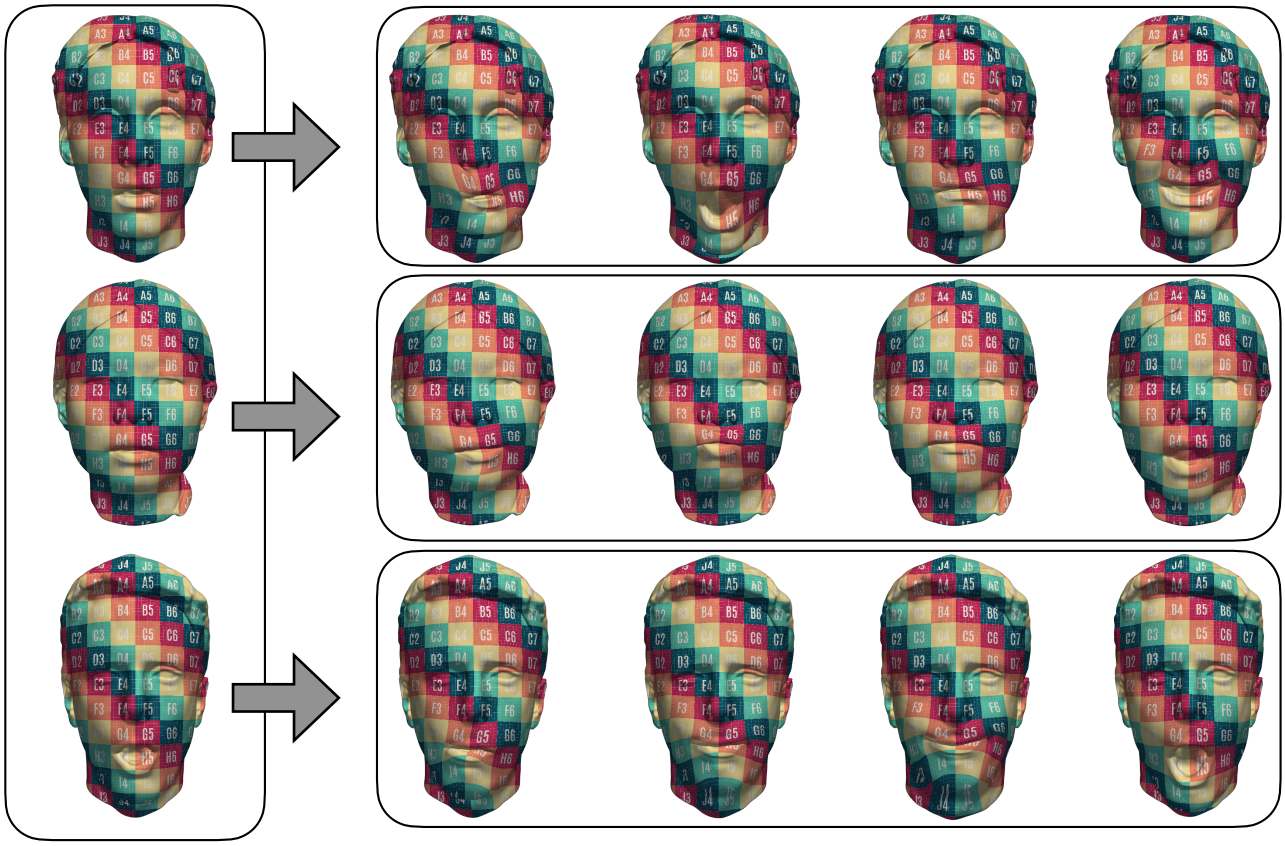

Figure 13. Deformation Consistency: We show surface correspondences between neutral and posed meshes from our test set.

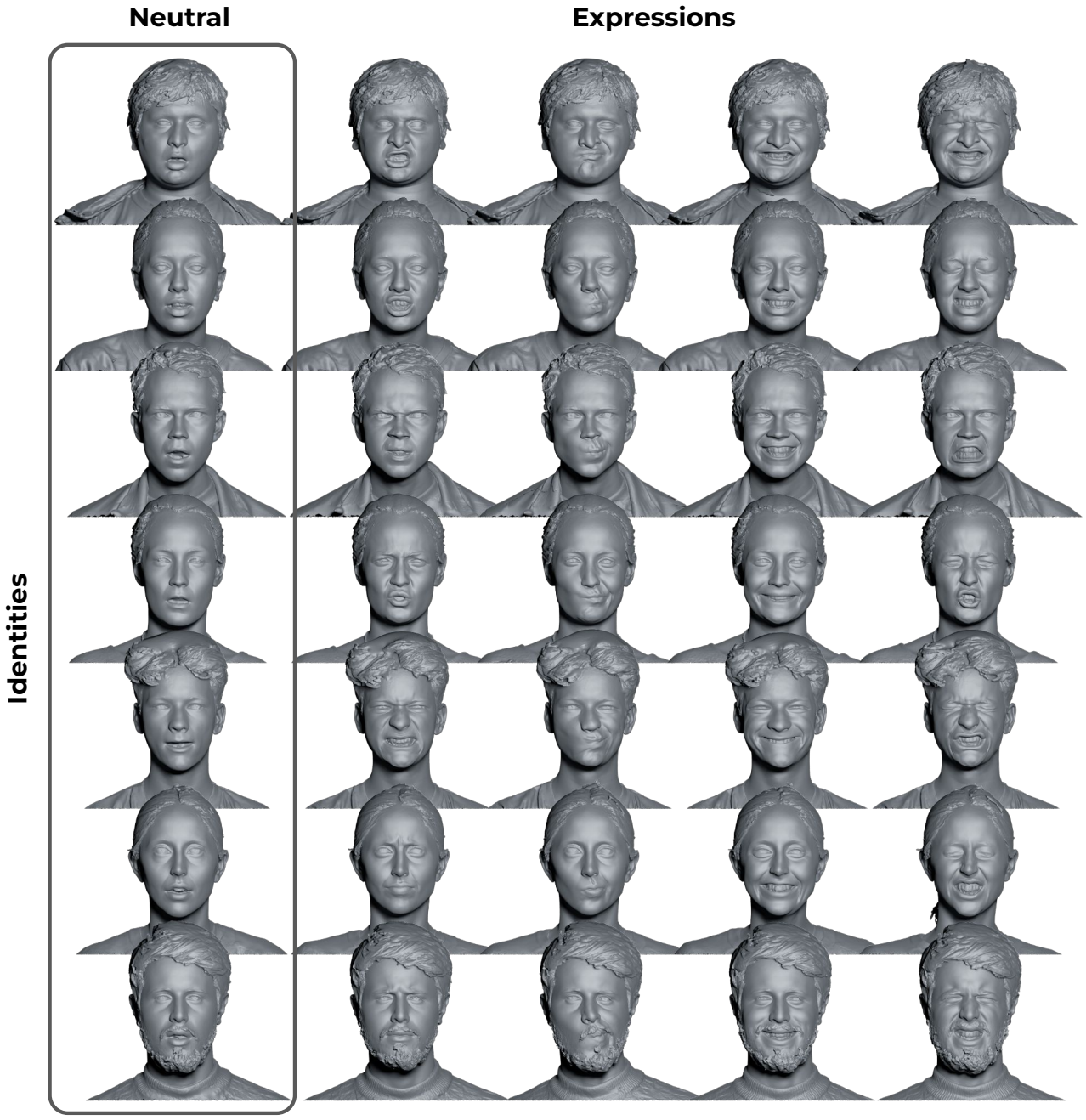

Figure 14. Additional 3D head scans from our newly-captured dataset. Here, we show how different participants perform expressions in their own unique ways.

Figure 15. We capture 20 expressions for each participant, and included three bonus expressions for the latest 50 participants. Here we show two subjects performing all expressions.