# H3D-Net: Few-Shot High-Fidelity 3D Head Reconstruction

Eduard Ramon[1,2] Gil Triginer[1] Janna Escur[1] Albert Pumarola[3] Jaime Garcia[1]
Xavier Giro-i-Nieto[2,3] Francesc Moreno-Noguer[3]
[1]*Crisalix SA* [2]*Universitat Politècnica de Catalunya* [3]*Institut de Robòtica i Informàtica Industrial, CSIC-UPC*

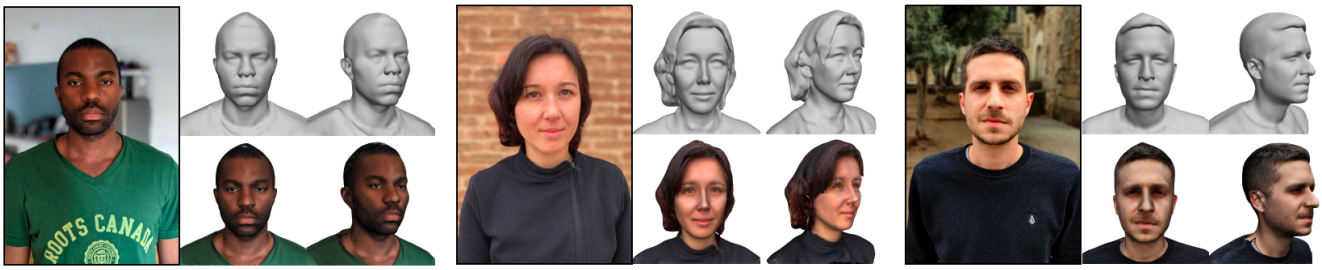

Figure 1. We introduce H3D-Net, a method for high-fidelity 3D head reconstruction in the wild. Our method estimates a signed distance function (SDF) of the head by optimizing a coordinate-based neural network on a small set of input images. This optimization process is constrained by a pre-trained probabilistic model of 3D head SDFs to obtain plausible shapes in few-shot setups. The figure shows the 3D head reconstruction of three scenes obtained with the proposed method from only three images with associated masks and camera poses.

## Abstract

*Recent learning approaches that implicitly represent surface geometry using coordinate-based neural representations have shown impressive results in the problem of multi-view 3D reconstruction. The effectiveness of these techniques is, however, subject to the availability of a large number (several tens) of input views of the scene, and computationally demanding optimizations. In this paper, we tackle these limitations for the specific problem of few-shot full 3D head reconstruction, by endowing coordinate-based representations with a probabilistic shape prior that enables faster convergence and better generalization when using few input images (down to three). First, we learn a shape model of 3D heads from thousands of incomplete raw scans using implicit representations. At test time, we jointly overfit two coordinate-based neural networks to the scene, one modelling the geometry and another estimating the surface radiance, using implicit differentiable rendering. We devise a two-stage optimization strategy in which the learned prior is used to initialize and constrain the geometry during an initial optimization phase. Then, the prior is unfrozen and fine-tuned to the scene. By doing this, we achieve high-fidelity head reconstructions, including hair and shoulders, and with a high level of detail that consistently outperforms both state-of-the-art 3D Morphable Models methods in the few-shot scenario, and non-parametric methods when large sets of views are available.*

## 1. Introduction

Recent learning based methods have shown impressive results in reconstructing 3D shapes from 2D images. These approaches can be roughly split into two main categories: model-based [3, 11, 25, 35, 36, 37, 44, 45, 49, 52] and model-free [9, 10, 16, 18, 21, 31, 33, 34, 39, 50, 51, 53]. The former incorporate prior knowledge using 3D Morphable Models (3DMMs) to limit the space of feasible solutions, making these approaches well suited for few-shot and one-shot shape estimation. However, most model-based methods produce shapes that usually lack geometric detail and cannot handle arbitrary topology changes.

On the other hand, model-free approaches based on discrete representations like voxels, meshes or point-clouds, have the flexibility to represent a wider spectrum of shapes, although at the cost of being computationally tractable only for small resolutions or being restricted to fixed topologies. These limitations have been overcome by neural implicit representations [8, 23, 24, 27, 32, 38, 41, 56], which can represent both geometry and appearance as a continuum, encoded in the weights of a neural network. [26, 54] have shown the success of such representations in learning detail-rich 3D geometry directly from images, with no 3D ground truth supervision. Unfortunately, the performance of these methods is currently conditioned to the availabil-

ity of a large number of input views, which leads to a time consuming inference.

In this work we introduce H3D-Net, a hybrid scheme that combines the strengths of model-based and model-free representations by incorporating prior knowledge into neural implicit models for category-specific multi-view reconstruction. We apply this approach to the problem of few-shot full head reconstruction. In order to build the prior, we first use several thousands of raw incomplete scans to learn a space of Signed Distance Functions (SDF) representing 3D head shapes [27]. At inference, this learnt shape prior is used to initalize and guide the optimization of an Implicit Differentiable Renderer (IDR) [54] that, given a potentially reduced number of input images, estimates the full head geometry. The use of the learned prior enables faster convergence during optimization and prevents it from being trapped into local minima, yielding 3D shape estimates that capture fine details of the face, head and hair from just three input images (see Figure 1).

We exhaustively evaluate our approach on a mid-resolution Multiview-Stereo (MVS) public dataset [29] and on a high-resolution dataset we collected with a structured-light scanner, consisting of 10 3D full-head scans. The results show that we consistently outperform current state-of-the-art, both in a few-shot setting and when many input views are available. Importantly, the use of the prior also makes our approach very efficient, achieving competitive results in terms of accuracy about 20$\times$ faster than IDR [54]. Our key contributions can be summarized as follows:

- We introduce a method for reconstructing high quality full heads in 3D from small sets of in-the-wild images.
- Our method is the first to use implicit functions for reconstructing 3D humans heads from multiple images and also to rival parametric and non-parametric models in 3D accuracy at the same time.
- We devise a guided optimization approach to introduce a probabilistic shape prior into neural implicit models.
- We collect and release a new dataset[1] containing high-resolution 3D full head scans, images, masks and camera poses for evaluation purposes, which we dub H3DS.

## 2. Related work

**Model-based**. 3D Morphable Models [5, 6, 20, 28, 30, 46, 47, 48] have become the *de facto* representation used for few-shot 3D face reconstruction in-the-wild given that they lead to light-weight, fast and robust systems. Adopting 3DMMs as a representation, the 3D reconstruction problem boils down to estimating the small set of parameters that best represent a target 3D shape. This makes it possible to obtain 3D reconstructions from very few images [3, 11, 35, 52] and even a single input [36, 37, 44, 45, 49]. Nevertheless, one of the main limitations of morphable models is their lack of expressiveness, specially for high frequencies. This issue has been addressed by learning a post processing that transfers the fine details from the image domain to the 3D geometry [21, 37, 45]. Another limitation of 3DMMs is their inability to represent arbitrary shapes and topologies. Thus, they are not suitable for reconstructing full heads with hair, beard, facial accessories and upper body clothing.

**Model-free**. Model-free approaches build upon more generic representations, such as voxel-grids or meshes, in order to gain expressiveness and flexibility. Voxel-grids have been extensively used for 3D reconstruction [10, 14, 16, 18, 53] and concretely for 3D face reconstruction [16]. Their main limitation is that memory requirements grow cubically with resolution, and octrees [14] have been proposed to address this issue. On the other hand, meshes [9, 17, 21, 50] are a more efficient representation for surfaces than voxel-grids, and are suitable for graphics applications. Meshes have been proposed for 3D face reconstruction [9, 21] in combination with graph neural networks [7]. However, similarly to 3DMMs, meshes are also usually restricted to fixed topologies and are not suitable for reconstructing other elements beyond the face itself.

**Implicit representations**. Recently, implicit representations for surface modelling [22, 23, 27] have been proposed to jointly address the memory limitations of voxel grids and the topological rigidity of meshes. These representations model surfaces as a level-set of a coordinate-based continuous function, *e.g.* a signed distance function or an occupancy function. Such functions, usually implemented as multi-layer perceptrons (MLPs), can theoretically express any shape with infinite resolution and a fixed memory footprint. Implicit methods for 3D reconstruction can be divided in those that, at inference time, perform a single forward pass of a previously trained model [8, 23, 38], and those that overfit a model to a set of input images through an optimization process using implicit differentiable rendering [26, 54]. In the later, given that the inference is an optimization process, the obtained 3D reconstructions are more accurate. However, they are slow and require an important number of multi-view images, failing in few-shot setups as those we consider in this work.

**Priors for implicit representations**. Building priors for implicit representations of surfaces has been addressed with two main purposes. The first consists in speeding up convergence of methods that perform an optimization at inference time [40] using meta-learning techniques [12]. The second is to find a space of implicit functions that represent the shape of a certain category using auto-decoders [27, 55]. However, [27, 55] have been used to solve tasks using 3D

[1]Project page: https://crisalixsa.github.io/h3d-net

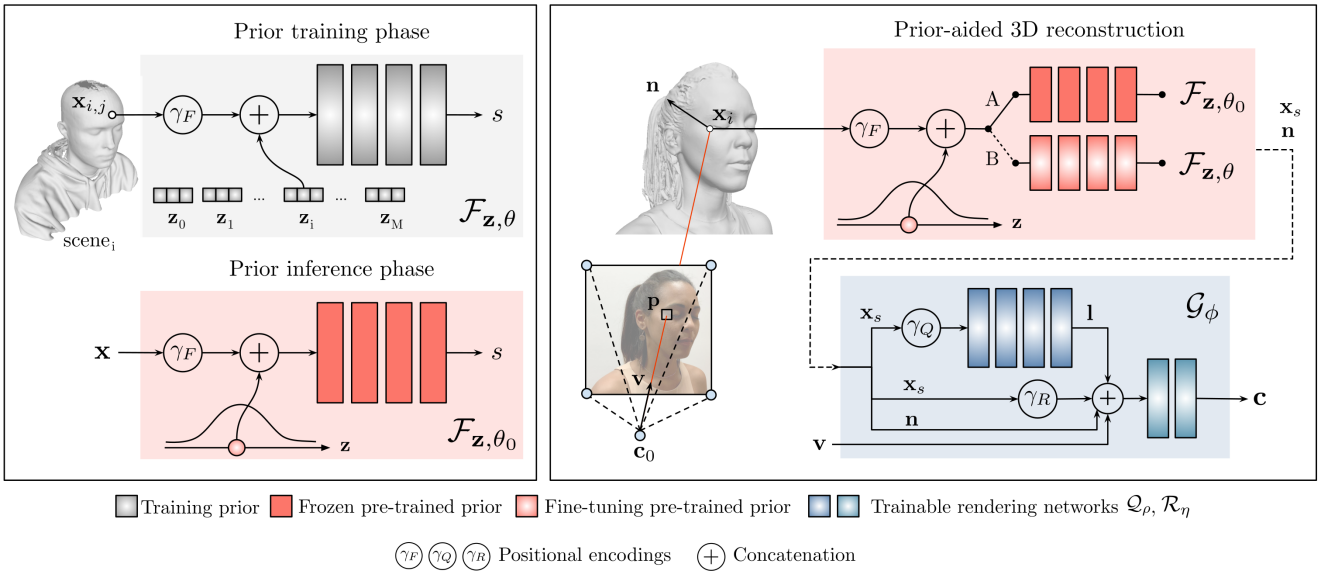


Figure 2. Overview of our method. **Left.** The two configurations of the prior model at training and inference phases. **Right.** Integration of the pre-trained prior model with the implicit differentiable renderer. During the prior-aided 3D reconstruction process, the geometry network starts off with frozen weights (commuter at position A), constraining the predicted shape to lie within its pre-learnt latent space, and is eventually unfrozen (commuter at position B) to allow fine-tuning of the fine details.

supervision, and it is still an open problem how to use these priors when the supervision signal is generated from 2D images.

As done with morphable models, implicit shape models can be used to constrain image-based 3D reconstruction systems to make them more reliable. Drawing inspiration from this idea, in this work we leverage implicit shape models [27] to guide the optimization-based implicit 3D reconstruction method [54] towards more accurate and robust solutions, even under few-shot in-the-wild scenarios.

## 3. Method

Given a small set of $N \geq 3$ input images $\mathbf{I}_v$, $v = 1, \ldots, N$, with associated head masks $\mathbf{M}_v$ and camera parameters $\mathbf{C}_v$, our goal is to recover the 3D head surface $\mathcal{S}$ using only visual cues as supervision. Formally, we aim to approximate the signed distance function (SDF) $\mathcal{F} : \mathbf{x} \rightarrow s$ such that $\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^3 | \mathcal{F}(\mathbf{x}) = 0\}$.

In order to approximate $\mathcal{F}$, we propose to optimize a previously learnt probabilistic model $\mathcal{F}_{\mathbf{z},\theta_0}$, that represents a prior distribution over 3D head SDFs. $\mathbf{z}$ and $\theta_0$ are a latent vector encoding specific shapes and the learnt parameters of an auto-decoder [42], respectively. Building on DeepSDF [27], we learn these parameters from thousands of incomplete scans. We describe this process in Section 3.1.

At test time, the reconstruction process is reduced to finding the optimal parameters $\{\mathbf{z}^*, \theta^*\}$ such that $\mathcal{F}_{\mathbf{z}^*,\theta^*} \sim \mathcal{F}$. To that end, we compose the prior model $\mathcal{F}_{\mathbf{z},\theta_0}$, which we also refer to as geometry network, with a rendering network $\mathcal{G}_\phi : (\mathbf{x}, \mathbf{n}, \mathbf{v}) \rightarrow \mathbf{c}$ that models the RGB radiance emitted from a surface point $\mathbf{x}$ with normal $\mathbf{n}$ in a viewing direction $\mathbf{v}$, and minimize a photometric error w.r.t. the input images $\mathbf{I}_v$, as in [54]. Moreover, we propose a two-step optimization schedule that prevents the reconstruction process from getting trapped into local minima and, as we shall see in the results section, leads to much more accurate, robust and realistic reconstructions. We describe the reconstruction step in Section 3.2.

### 3.1. Learning a prior for human head SDFs

Given a set of $M$ scenes with associated raw 3D point clouds, we use the DeepSDF framework to learn a prior distribution of signed distance functions representing 3D heads, $\mathcal{F}_{\mathbf{z},\theta_0}$. While the original DeepSDF formulation requires watertight meshes as training data to use signed distances as supervision, we use the Eikonal loss [13] to learn directly from raw, and potentially incomplete, surface point clouds. In addition, Fourier features are used to overcome the spectral bias of MLPs towards low frequencies in low dimensional tasks [43]. We illustrate the training and inference process of the prior model in Figure 2-left.

For each scene, indexed by $i = 1, \ldots, M$, we sample a subset of points $\mathcal{P}_{\mathrm{s}}^{(i)}$ on the surface, and another set $\mathcal{P}_{\mathrm{v}}^{(i)}$ uniformly taken from a volume containing the scene, and minimize the following objective:

$$\underset{\{\mathbf{z}_i\},\theta}{\arg\min} \sum_{i=1}^{M} \mathcal{L}_{\mathrm{Surf}}^{(i)} + \lambda_0 \mathcal{L}_{\mathrm{Emb}}^{(i)} + \lambda_1 \mathcal{L}_{\mathrm{Eik}}^{(i)}, \qquad (1)$$

where $\lambda_0$ and $\lambda_1$ are hyperparameters and $\mathcal{L}^{(i)}_{\text{Surf}}$ accounts for the SDF error at surface points:

$$\mathcal{L}^{(i)}_{\text{Surf}} = \sum_{\mathbf{x}_j \in \mathcal{P}^{(i)}_{\text{s}}} |\mathcal{F}_{\mathbf{z}_i,\theta}(\mathbf{x}_j)| \ . \tag{2}$$

$\mathcal{L}^{(i)}_{\text{Emb}}$ enforces a zero-mean multivariate-Gaussian distribution with spherical covariance $\sigma^2$ over the space of latent vectors:

$$\mathcal{L}^{(i)}_{\text{Emb}} = \frac{1}{\sigma^2}||\mathbf{z}_i||^2_2. \tag{3}$$

Finally, $\mathcal{L}^{(i)}_{\text{Eik}}$ regularizes $F_{\mathbf{z}_i,\theta}$ with the Eikonal loss to ensure that it approximates a signed distance function by keeping its gradients close to unit norm:

$$\mathcal{L}^{(i)}_{\text{Eik}} = \sum_{\mathbf{x}_k \in \mathcal{P}^{(i)}_{\text{v}}} \left( \|\nabla_{\mathbf{x}} \mathcal{F}_{\mathbf{z}_i,\theta}(\mathbf{x}_k)\| - 1 \right)^2. \tag{4}$$

This regularization across the whole volume is necessary given that our meshes are not watertight and only a subset of surface points is available as ground truth [13].

After training, we have obtained the parameters $\theta_0$ that represent a space of human head SDFs. We can now draw signed distance functions of heads from $F_{\mathbf{z},\theta_0}$ by sampling the latent space $\mathbf{z}$. We use this pre-trained model as the prior for the 3D reconstruction schedule described in the following section.

### 3.2. Prior-aided 3D Reconstruction

Given a new scene, for which no 3D information is provided at this point, we aim to approximate the SDF that implicitly encodes the surface of the head by only supervising in the image domain. To that end, we compose the previously learnt geometry probabilistic model $\mathcal{F}_{\mathbf{z},\theta_0}$ with the rendering network $\mathcal{G}_\phi$, and supervise on the photometric error to find the optimal parameters $\mathbf{z}^*$, $\theta^*$ and $\phi^*$. The reconstruction process is illustrated in Figure 2-right. Note that, in contrast to [54], the geometry and rendering modules are perfectly decoupled.

For every pixel coordinate $p$ of each input image $\mathbf{I}_v$, we march a ray $\mathbf{r} = \{\mathbf{c}_0 + t\mathbf{v} | t \geq 0\}$, where $\mathbf{c}_0$ is the position of the associated camera $\mathbf{C}_v$, and $\mathbf{v}$ the viewing direction. The intersection point $\mathbf{x}_i$ between the ray $\mathbf{r}$ and the surface $\mathcal{S}_{\mathbf{z},\theta} = \{\mathbf{x} | \mathcal{F}_{\mathbf{z},\theta}(\mathbf{x}) = 0\}$ can be efficiently found using sphere tracing [15]. This intersection point can be made differentiable w.r.t $\mathbf{z}$ and $\theta$ without having to store the gradients corresponding to all the forward passes of the geometry network, as shown in [26] and generalized by [54]. The following expression is exact in value and first derivatives:

$$\mathbf{x}_s = \mathbf{x}_i - \frac{\mathbf{v}}{\nabla_{\mathbf{x}} \mathcal{F}_{\mathbf{z}_k,\theta_k}(\mathbf{x}_i) \cdot \mathbf{v}} \mathcal{F}_{\mathbf{z},\theta}(\mathbf{x}_i) \ . \tag{5}$$

Here $\mathbf{z}_k$ and $\theta_k$ denote the parameters of $\mathcal{F}_{\mathbf{z},\theta}$ at iteration $k$, and $\mathbf{x}_s$ represents the intersection point made differentiable w.r.t. the geometry network parameters.

Next, we evaluate the mapping $\mathcal{G}_\phi$ at $\mathbf{x}_s$, $\mathbf{n} = \nabla_{\mathbf{x}} \mathcal{F}_{\mathbf{z},\theta}(\mathbf{x}_s)$ and $\mathbf{v}$ to estimate the color $\mathbf{c}$ for the pixel $p$ in the image $\mathbf{I}_v$:

$$\mathbf{c} = \mathcal{G}_\phi(\mathbf{x}_s, \mathbf{n}, \mathbf{v}) \ . \tag{6}$$

Finally, in order to optimize the surface parameters $\mathbf{z}$ and $\theta$, and the rendering parameters $\phi$, we minimize the following loss [54]:

$$\mathcal{L} = \sum_{v=1}^{N} \mathcal{L}^{(v)}_{\text{RGB}} + \beta_0 \mathcal{L}^{(v)}_{\text{Mask}} + \beta_1 \mathcal{L}^{(v)}_{\text{Eik}}, \tag{7}$$

where $\beta_0$ and $\beta_1$ are hyperparameters. We next describe each component of this loss. Let $\mathcal{P}$ be a mini-batch of pixels from view $v$, $\mathcal{P}_{\text{RGB}}$ the subset of pixels whose associated ray intersects $\mathcal{S}_{\mathbf{z},\theta}$ and which have a nonzero mask value, and $\mathcal{P}_{\text{Mask}} = \mathcal{P} \setminus \mathcal{P}_{\text{RGB}}$. The $\mathcal{L}^{(v)}_{\text{RGB}}$ is the photometric error, computed as:

$$\mathcal{L}^{(v)}_{\text{RGB}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}_{\text{RGB}}} |\mathbf{I}_v(p) - \mathbf{c}_v(p)| \ . \tag{8}$$

$\mathcal{L}^{(v)}_{\text{Mask}}$ accounts for silhouette errors:

$$\mathcal{L}^{(v)}_{\text{Mask}} = \frac{1}{\alpha |\mathcal{P}|} \sum_{p \in \mathcal{P}_{\text{Mask}}} \text{CE}(\mathbf{M}_v(p), s_{v,\alpha}(p)) \ , \tag{9}$$

where $s_{v,\alpha} = \text{sigmoid}(-\alpha \min_{t \geq 0} \mathcal{F}_{\mathbf{z},\theta}(\mathbf{r}_t))$ is the estimated silhouette, CE is the binary cross-entropy and $\alpha$ is a hyperparameter. Lastly, $\mathcal{L}_{\text{Eik}}$ encourages $\mathcal{F}_{\mathbf{z},\theta}$ to approximate a signed distance function as in Equation 4.

Instead of jointly optimizing all the parameters $\{\mathbf{z}, \theta, \phi\}$ to minimize $\mathcal{L}$ we introduce a two-step optimization schedule which is more appropriate for auto-decoders like DeepSDF. We begin by initializing the geometry network $\mathcal{F}_{\mathbf{z},\theta}$ with the previously learnt prior for human head SDFs, $\mathcal{F}_{\mathbf{z},\theta_0}$, and a randomly sampled $\mathbf{z}_0$ such that $\|\mathbf{z}_0\| < \epsilon$ to stay near the mean of the latent space. In a first phase, we only optimize $\mathbf{z}$ and $\phi$ as $\arg\min_{\mathbf{z},\phi} \mathcal{L}$, which is equivalent to the standard auto-decoder inference. By doing so, the resulting surface $\mathcal{S}_{z^*,\theta}$ is forced to stay within the learnt distribution of 3D heads. Once the geometry and the radiance mappings have reached an equilibrium, *i.e.* the optimization has converged, we unfreeze the decoder parameters $\theta$ to fine-tune the whole model as $\arg\min_{\mathbf{z},\theta,\phi} \mathcal{L}$.

In Section 5, we empirically prove that by using this optimization schedule instead of optimizing all the parameters at once, the obtained 3D reconstructions are more accurate and less prone to artifacts, specially in few-shot setups.

## 4. Implementation details

Our implementation of the prior model closely follows the one proposed in [13], with the addition that we apply

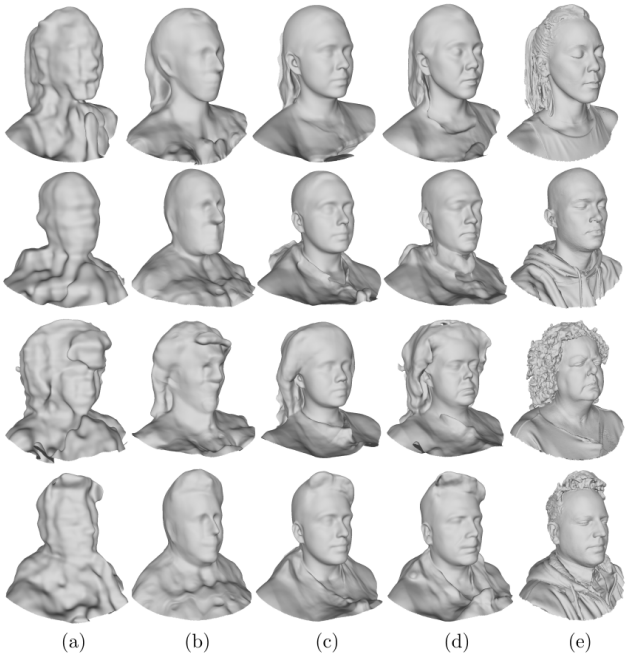


Figure 3. **Ablation study** of our method in the few-shot setup (3 views). From left to right: (a) Ours with geometric initialization [2] and no schedule, (b) Ours with small prior initialization (500 subjects) and no schedule, (c) Ours with large prior initialization (10,000 subjects) and no schedule, (d) Ours with large prior initialization and schedule and (e) ground truth.

| | (a) | (b) | (c) | (d) |
|---|---|---|---|---|
| Face mean distance [mm] | 4.04 | 2.68 | 1.90 | **1.49** |
| Full-head mean distance [mm] | 16.68 | 17.08 | 14.59 | **12.76** |

Table 1. **Ablation study** of our method in the few-shot setup (3 views). The face and full-head mean distances are the averages over all the subjects in the H3DS dataset. The configurations a,b,c,d are the same as those described in Figure 3.

a positional encoding $\gamma_{\mathcal{F}}$ to the input coordinates $\mathbf{x}$ with 6 log-linear spaced frequencies. The encoded 3D coordinates are concatenated with the $\mathbf{z}$ latent vector of size 256 and set as the input to the decoder. The decoder is a MLP of 8 layers with 512 neurons in each layer and single skip connection from the input of the decoder to the output of the 4th layer. We use Softplus as activation function in every layer except the last, where no activation is used. The prior model is trained for 100 epochs using Adam [19] with standard parameters, learning rate of $10^{-4}$ and learning rate step decay of 0.5 every 15 epochs. The training takes approximately 50 minutes for a small dataset (500 scenes) and 10 hours for a large one (10,000 scenes).

The 3D reconstruction network is composed by the prior model described above and a mapping $\mathcal{G}_\phi$ that is split into two sub-networks $\mathcal{Q}_\rho$ and $\mathcal{R}_\eta$ as shown in Figure 2. $\mathcal{Q}_\rho$ is a MLP implemented exactly as the decoder of the prior

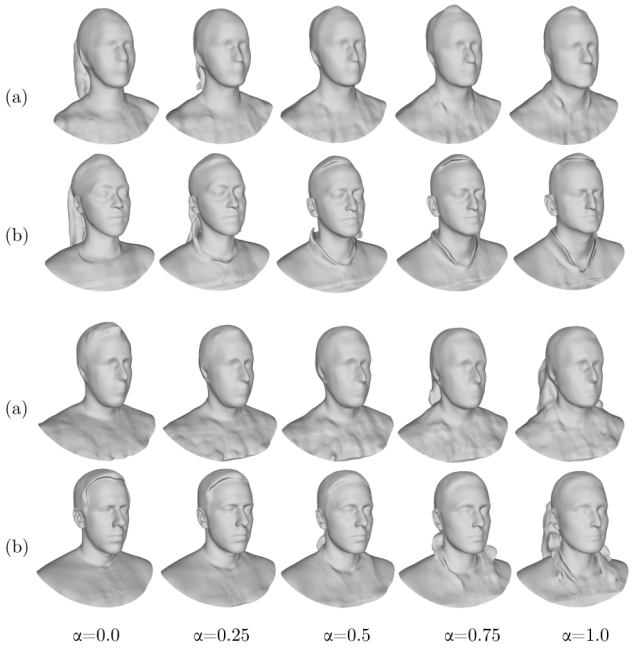


Figure 4. **Latent interpolation** between different subjects, being $\alpha$ a linear interpolation factor in $\mathbf{z}$ space. (a) uses the small prior model (500 subjects), and (b) the large prior (10,000 subjects).

model, except for the input layer, which takes in a 3 dimensional vector, and the output layer, which outputs a 256-dimensional vector $\mathbf{l}$. As in [54], $\mathcal{R}_\eta$ is a smaller MLP composed by 4 layers, each 512 neurons wide, no skip connections, and ReLU activations except in the output layer which is tanh. We also apply the positional encodings $\gamma_{\mathcal{Q}}$ and $\gamma_{\mathcal{R}}$ to $\mathbf{x}_s$ with 6 and 4 log-linear spaced frequencies respectively. Each scene is trained for 2000 epochs using Adam with fixed learning rate of $10^{-4}$ and learning rate step decay of 0.5 at epochs 1000 and 1500. The scene reconstruction process takes approximately 25 minutes for scenes of 3 views and 4 hours and 15 minutes for scenes of 32.

All the experiments for both prior and reconstruction models have been performed using a single Nvidia RTX 2080Ti.

## 5. Experiments

In this section, we evaluate quantitatively and qualitatively our multi-view 3D reconstruction method. We empirically demonstrate that the proposed solution surpasses the state of the art in the few-shot [3, 52] and many-shot [54] scenarios for 3D face and head reconstruction in-the-wild.

### 5.1. Datasets

**Prior training.** In order to train the geometry prior, we use an internal dataset made of 3D head scans from 10,000 individuals. The dataset is perfectly balanced in gender and

| | 3DFAW | H3DS | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 3 views | 3 views | | 4 views | | 8 views | | 16 views | | 32 views | |
| | face | face | head | face | head | face | head | face | head | face | head |
| MVFNet [52] | 1.54 | 1.66 | - | - | - | - | - | - | - | - | - |
| DFNRMVS [3] | 1.53 | 1.83 | - | - | - | - | - | - | - | - | - |
| IDR [54] | 3.92 | 3.52 | 17.04 | 2.14 | 8.04 | 1.95 | 8.71 | 1.43 | 5.94 | 1.39 | 5.86 |
| H3D-Net (Ours) | **1.37** | **1.49** | **12.76** | **1.65** | **7.95** | **1.38** | **5.47** | **1.24** | **4.80** | **1.21** | **4.90** |

Table 2. **3D reconstruction method comparison.** Average surface error in millimeters computed over all the subjects in the 3DFAW and H3DS datasets. Find the precise definition of the face/head metrics, as well as a description of the distribution of the views, in section 5.2.

diverse in age and ethnicity. The raw data is automatically processed to remove internal mesh faces and non-human parts such as background walls. Finally, all the scenes are aligned by registering a template 3D model with non-rigid Iterative Closest Point (ICP) [1].

**3DFAW [29].** We evaluate our method in the 3DFAW dataset. This dataset provides videos recorded in front, and around, the head of a person in static position as well as mid-resolution 3D ground truth of the facial region. We select 5 male and 5 female scenes and use them to evaluate only the facial region.

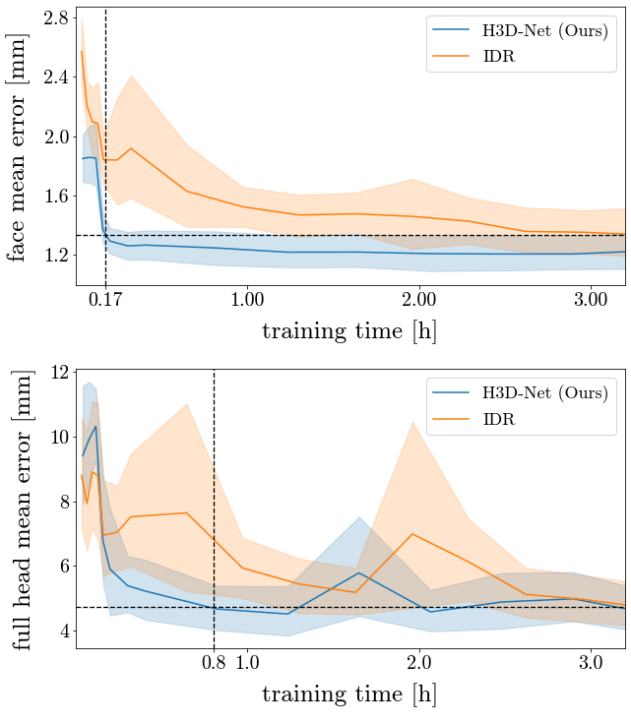


Figure 5. **3D reconstruction convergence** comparison between H3D-Net and IDR [54] using 32 views. Metrics are computed over all the samples in the H3DS dataset. The dotted lines indicate the time when our method first surpasses the best mean error attained by IDR over the entire optimization. **Top.** Mean surface error in the face. **Bottom.** Mean surface error in the full head.

**H3DS.** We introduce and release a new dataset called *H3DS*, the first dataset containing high resolution full head 3D textured scans and 360º images with associated ground truth camera poses and ground truth masks. The 3D geometry has been captured using a structured light scanner, which leads to more precise ground truth geometries than the ones from 3DFAW [29], which were generated using Multi-View Stereo (MVS). The dataset consists of 10 individuals, 50% man and 50% woman. We use this dataset to evaluate the accuracy of the different methods in both the full head and the facial regions.

## 5.2. Experiments setup

We use the 3DMM-based methods MVFNet [52] and DFNRMVS [3], and the model-free method IDR [54] as baselines to compare against H3D-Net.

In the few-shot scenario (3 views), all the methods are evaluated on the 3DFAW and H3DS datasets. To benchmark our method when more than 3 views are available, we compare it against IDR on the H3DS dataset.

The evaluation criteria have been the same for all methods and in all the experiments. The predicted 3D reconstruction is roughly aligned with the ground truth mesh using manually annotated landmarks, and then refined with rigid ICP [4]. Then, we compute the unidirectional Chamfer distance from the predicted reconstruction to the ground truth. All the distances are computed in millimeters.

We report metrics in two different regions, the face and the full head. For the finer evaluation in the face region, we cut both the reconstructions and the ground truth using a sphere of 95 mm radius and with center at the tip of the nose of the ground truth mesh, and refine the alignment with ICP as in [29, 49]. Then, we compute the Chamfer distance in this sub-region. For the full head evaluation, the ICP alignment is performed using an annotated region that includes the face, the ears, and the neck, since it is a region visible in all view configurations (3, 4, 8, 16 and 32). These configurations are defined by their yaw angles as follow: $\mathcal{V}_3 = \{0, \pm 45\}$, $\mathcal{V}_4 = \{\pm 45, \pm 90\}$ and $\mathcal{V}_N = \{\frac{360}{N} i\}_{i=1}^{N}$ for $N = 8, 16, 32$. In this case, the Chamfer distance is computed for all the vertices of the reconstruction.

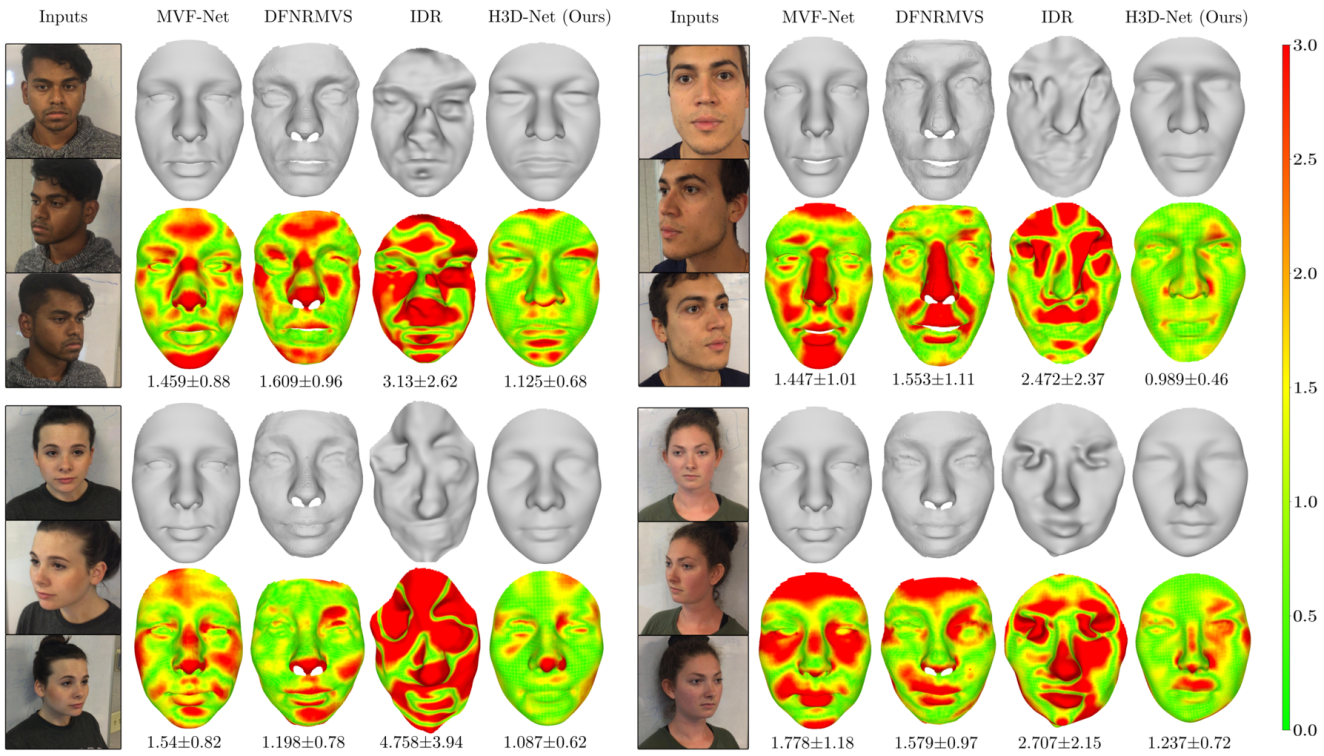


Figure 6. **Qualitative results** obtained for 4 subjects from the 3DFAW dataset [29] with only three input views. First and third rows show the reconstructed geometry and second and fourth rows show the surface error with the color code being in millimeters.

### 5.3. Ablation study

We conduct an ablation study on the H3DS dataset in the few-shot scenario (3 views) and show the numerical results in Table 1, and the qualitative results in Figure 3. First, we reconstruct the scenes without prior and without schedule (a). In this case, the geometry network is initialized using geometric initialization [2], representing a sphere of radius one at the beginning of the optimization. Then, we initialize the geometry network with two different priors, a small one trained on 500 subjects (b), and a large one trained on 10,000 subjects (c), and perform the reconstructions without schedule. As it can be observed, initializing the geometry network with a previously learnt prior leads to smoother and more plausible surfaces, specially when more subjects have been used to train it. It is important to note that the benefits of the initialization are not only due to a better initial shape, but also to the ability of the initial weights to generalize to unseen shapes, which is greater in the large prior model. Finally, we initialize the geometry network with the large prior and use the proposed optimization schedule during the reconstruction process. It can be observed how the resulting 3D heads resemble much more to the ground truth in terms of finer facial details.

Given the notable effect that the number of samples has in the learnt prior representations and in the resulting 3D reconstructions as well, we visualize latent space interpolations in Figure 4. To that end, we optimize the latent vector for two ground truth 3D scans as shown in Figure 2-left-bottom in order minimize Equation 1. Then, we interpolate between the two optimal latent vectors. As it can be observed, the 3D reconstructions resulting from the interpolation in $\mathbf{z}$ space of the large prior model are more detailed and plausible than the ones from the small prior model, suggesting that the later achieves poorer generalization.

### 5.4. Quantitative results

Quantitative results in terms of surface error are reported in Table 2. Remarkably, H3D-Net outperforms both 3DMM-based methods in the few-shot regime, and the model-free method IDR when the largest number of views (32) are available. It is worth noting how the enhancement due to the prior is more significant as the number of views decreases. Nevertheless, the prior does not prevent the model from becoming more accurate when more views are available, which is a current limitation of model-based approaches.

We also analyze the trade-off between the optimization time and the accuracy in IDR and H3D-Net for the case of 32 views, which we illustrate in Figure 5. It can be observed that, despite reaching similar errors asymptotically, in average our method achieves the best performance attained by IDR much faster. In particular, we report convergence gains

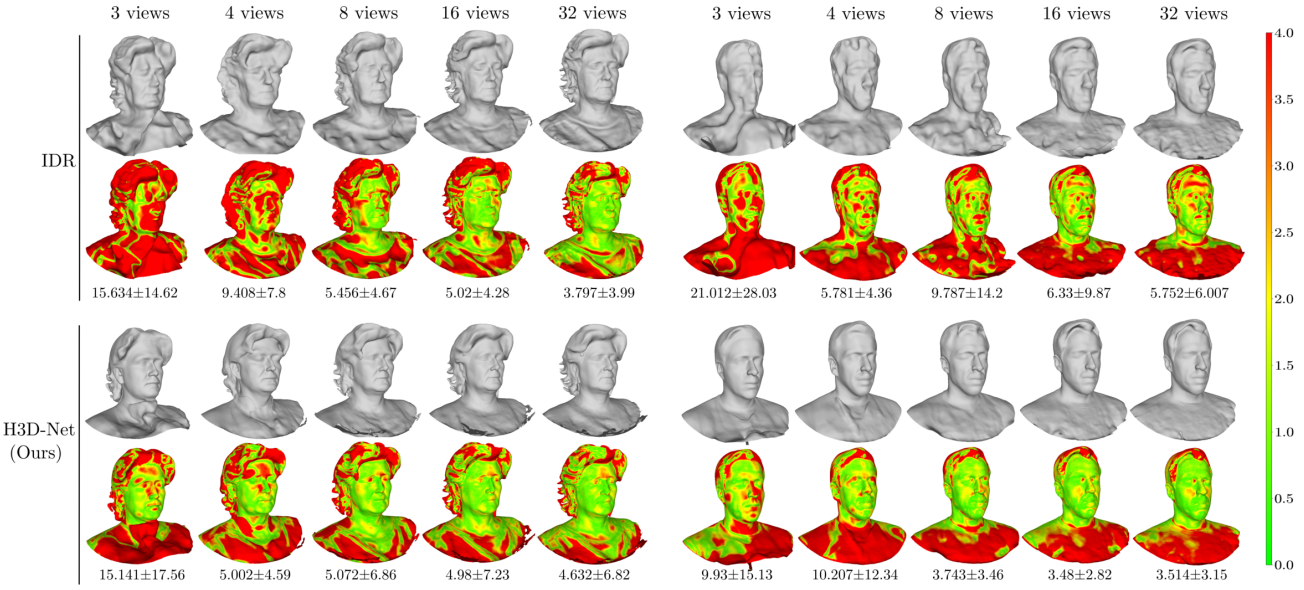


Figure 7. **Qualitative results** obtained for 2 subjects from the H3DS dataset when varying the number of views. The first and second rows correspond with the results of IDR and the third and the forth with H3D-Net results (ours). The surface error is represented with the color code in millimeters.

of 20× for the facial region error and 4× for the full head. Moreover, the smaller variance observed in H3D-Net (blue) indicates that it is a more stable method.

### 5.5. Qualitative results

Quantitative results show improvements over the baselines in both few-shot and many-shot setups. Here, we study how this is translated into the reconstructed 3D shape.

In Figure 6, we qualitatively evaluate the three baselines and H3D-Net for the case of 3 input views. As expected, IDR [54] is the worst performing model in this scenario, generating reconstructions with artifacts and with no resemblance to human faces. On the other hand, 3DMM-based models [3, 52] achieve more plausible shapes, but they struggle to capture fine anatomical details, specially in difficult areas such as the nose, eyebrows, cheeks and chin. H3D-Net, in contrast, is able to capture much more detail and reduce significantly the errors over the whole face.

We also evaluate the impact that varying the number of available views has on the reconstructed surface, and compare our method to IDR [54]. As shown in Figure 7, H3D-Net is able to obtain surfaces with less error (greener) with far fewer views, which is consistent with the quantitative results reported in Table 2. Notably, it can also be observed that, even when errors are numerically similar (first and third columns), the reconstructions from H3D-Net are much more realistic. In addition, H3D-Net improvements are especially notable within the face region. We attribute this to the fact that training data used to build the prior model is more rich in this area, whereas training examples frequently present holes in other parts of the head.

## 6. Conclusions

In this work we have presented H3D-Net, a method for high-fidelity 3D head reconstruction from small sets of in-the-wild images with associated head masks and camera poses. Our method combines a pre-trained probabilistic model, which represents a distribution of head SDFs, with an implicit differentiable renderer that allows direct supervision in the image domain. By constraining the reconstruction process with the prior model, we are able to robustly recover detailed 3D human heads, including hair and shoulders, from only three input images. After a thorough quantitative and qualitative evaluation, our experiments show that our method outperforms both model-based methods in the few-shot setup and model-free methods when a large number of views are available. One limitation of our method is that it still requires several minutes to generate 3D reconstructions. An interesting direction for future work is to use more efficient representations for SDFs and color priors in order to speed up the optimization process.

## Acknowledgments

This work has been partially funded by the Spanish government with the projects MoHuCo PID2020-120049RB-I00, DeeLight PID2020-117142GB-I00 and Maria de Maeztu Seal of Excellence MDM-2016-0656, and by the Government of Catalonia under 2017 DI 028.

## References


[1] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *CVPR*, 2007.

[2] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *CVPR*, 2020.

[3] Ziqian Bai, Zhaopeng Cui, Jamal Ahmed Rahim, Xiaoming Liu, and Ping Tan. Deep facial non-rigid multi-view stereo. In *CVPR*, 2020.

[4] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, 1992.

[5] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3d face morphable models" in-the-wild". In *CVPR*, 2017.

[6] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *CVPR*, 2016.

[7] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[8] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019.

[9] Shiyang Cheng, Georgios Tzimiropoulos, Jie Shen, and Maja Pantic. Faster, better and more detailed: 3d face reconstruction with graph convolutional networks. In *ACCV*, 2020.

[10] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016.

[11] Pengfei Dou and Ioannis A Kakadiaris. Multi-view 3d face reconstruction with deep recurrent neural networks. *Image and Vision Computing*, 80:80–91, 2018.

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

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

[14] Christian Häne, Shubham Tulsiani, and Jitendra Malik. Hierarchical surface prediction for 3d object reconstruction. In *3DV*, 2017.

[15] John C Hart. Sphere tracing: A geometric method for the antialiased ray tracing of implicit surfaces. *The Visual Computer*, 12(10):527–545, 1996.

[16] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *ICCV*, 2017.

[17] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018.

[18] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *NeurIPS*, 2017.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

[21] Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In *CVPR*, 2020.

[22] Gidi Littwin and Lior Wolf. Deep meta functionals for shape representation. In *ICCV*, 2019.

[23] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019.

[24] 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. *arXiv preprint arXiv:2003.08934*, 2020.

[25] Francesc Moreno-Noguer and Pascal Fua. Stochastic exploration of ambiguities for nonrigid shape recovery. 35(2):463–475, 2013.

[26] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020.

[27] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019.

[28] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *AVSS*, pages 296–301, 2009.

[29] Rohith Krishnan Pillai, László Attila Jeni, Huiyuan Yang, Zheng Zhang, Lijun Yin, and Jeffrey F Cohn. The 2nd 3d face alignment in the wild challenge (3dfaw-video): Dense reconstruction from video. In *ICCV Workshops*, 2019.

[30] Stylianos Ploumpis, Haoyang Wang, Nick Pears, William AP Smith, and Stefanos Zafeiriou. Combining 3d morphable models: A large scale face-and-head model. In *CVPR*, 2019.

[31] Albert Pumarola, Antonio Agudo, Lorenzo Porzi, Alberto Sanfeliu, Vincent Lepetit, and Francesc Moreno-Noguer. Geometry-aware network for non-rigid shape prediction from a single view. In *CVPR*, 2018.

[32] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, 2021.

[33] Albert Pumarola, Stefan Popov, Francesc Moreno-Noguer, and Vittorio Ferrari. C-flow: Conditional generative flow models for images and 3d point clouds. In *CVPR*, 2020.

[34] Albert Pumarola, Jordi Sanchez-Riera, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3dpeople: Modeling the geometry of dressed humans. In *ICCV*, 2019.

[35] Eduard Ramon, Janna Escur, and Xavier Giro-i Nieto. Multi-view 3d face reconstruction in the wild using siamese networks. In *ICCV Workshops*, 2019.

[36] Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data. In *3DV*, 2016.

[37] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, 2017.

[38] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019.

[39] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *ICCV*, 2017.

[40] Vincent Sitzmann, Eric R Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. Metasdf: Meta-learning signed distance functions. *arXiv preprint arXiv:2006.09662*, 2020.

[41] Vincent Sitzmann, Michael Zollhoefer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[42] Shufeng Tan and Michael L Mayrovouniotis. Reducing data dimensionality through optimizing neural network inputs. *AIChE Journal*, 41(6):1471–1480, 1995.

[43] Matthew Tancik, Pratul P Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *arXiv preprint arXiv:2006.10739*, 2020.

[44] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV*, 2017.

[45] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard G Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *CVPR*, 2018.

[46] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *CVPR*, 2019.

[47] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *CVPR*, 2018.

[48] Luan Tran and Xiaoming Liu. On learning 3d face morphable model from in-the-wild images. *TPAMI*, 43(1):157–171, 2019.

[49] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *CVPR*, 2017.

[50] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018.

[51] Huawei Wei, Shuang Liang, and Yichen Wei. 3d dense face alignment via graph convolution networks. *arXiv preprint arXiv:1904.05562*, 2019.

[52] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *CVPR*, 2019.

[53] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *NeurIPS*, 2016.

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

[55] 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. *arXiv preprint arXiv:2011.14143*, 2020.

[56] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021.