InstantAvatar: Efficient 3D Head Reconstruction via Surface Rendering

Antonio Canela [1,2,3] Pol Caselles [1,2,3] Ibrar Malik [1,2,3] Eduard Ramon [1,2*]

Jaime García [1] Jordi Sánchez-Riera [3] Gil Triginer [1] Francesc Moreno-Noguer [3]

[1] *Crisalix Labs* [2] *Universitat Politècnica de Catalunya* [3] *Institut de Robòtica i Informàtica Industrial, CSIC-UPC*

Abstract

Recent advances in full-head reconstruction have been obtained by optimizing a neural field through differentiable surface or volume rendering to represent a single scene. While these techniques achieve an unprecedented accuracy, they take several minutes, or even hours, due to the expensive optimization process required. In this work, we introduce InstantAvatar, a method that recovers full-head avatars from few images (down to just one) in a few seconds on commodity hardware. In order to speed up the reconstruction process, we propose a system that combines, for the first time, a voxel-grid neural field representation with a surface renderer. Notably, a naive combination of these two techniques leads to unstable optimizations that do not converge to valid solutions. In order to overcome this limitation, we present a novel statistical model that learns a prior distribution over 3D head signed distance functions using a voxel-grid based architecture. The use of this prior model, in combination with other design choices, results into a system that achieves 3D head reconstructions with comparable accuracy as the state-of-the-art with a 100× speed-up.

1. Introduction

Obtaining 3D avatars from images or videos is an essential building block for many mixed reality applications. Traditional methods based on 3D morphable models (3DMM) [19, 34] have proved to be a robust solution, but they struggle to capture details outside of the training distribution, and cannot handle topological changes caused by hair, clothes and other accessories. Voxel grids provide more flexibility than 3DMM to represent arbitrary shapes, but they are limited by their memory requirements. In order to avoid these limitations, neural fields have been proposed as a highly expressive 3D representation at a fixed memory footprint.

In combination with neural rendering, neural fields have achieved impressive results for the tasks of novel view synthesis [28] and 3D reconstruction [55], and have been successfully applied to represent full heads with unprecedented accuracy and photorealism [7, 16, 38]. More concretely, methods that represent the geometry implicitly using a Signed Distance Function (SDF) have led to very detailed surface reconstructions, either when optimized through surface [55] or volume rendering [32, 52, 54]. These methods have been applied to recover full-head avatars from videos [12, 57] or still photos [7, 38], achieving more detailed and visually appealing reconstructions than their 3DMM-based alternatives.

However, techniques based on differentiable rendering take several minutes, or even hours, to reconstruct a single

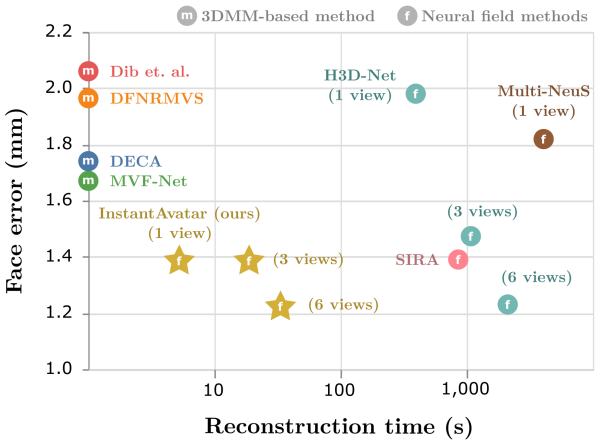

Figure 1. **Reconstruction time comparison**. InstantAvatar is a method that obtains full head 3D avatars from one or few images in a matter of seconds. This figure reports the time vs reconstruction error for ours and state-of-the-art methods, when considering only the face region where all methods are applicable (full head error metrics are reported in the experimental section). InstantAvatar speed is only surpassed by 3DMM methods, which, however are significantly less accurate. Compared against other neural field approaches, InstantAvatar obtains a 100× speed up at similar reconstruction error values.

*This work was done prior to joining Amazon.

scene due to the costly optimization process involved. To alleviate this, some works have leveraged hybrid architectures that encode neural fields as a composition of a feature grid and a shallow multi-layer perceptron (MLP) [29, 43, 44]. These hybrid architectures have led to much faster optimizations of neural fields through volume rendering. Still, the convergence time for these approaches remains in the tens of minutes [18], which is prohibitively high for many applications.

A promising direction towards further speeding up the optimization of neural fields through differentiable rendering is optimizing a hybrid neural architecture through surface rendering instead of volumetric rendering. The potential speed gains arise because volume rendering is slowed down by the computation of both a forward and a backward pass for a large number of points along rendered rays. Unfortunately, combining hybrid neural architectures and surface rendering is far from trivial. We observe that the lack of inductive biases for smoothness in a feature grid, together with the highly sparse sampling from the surface rendering leads to an unstable optimization process that does not converge to valid solutions. This problem is accentuated when the number of input images is reduced.

In this paper, we propose a method for successfully optimizing hybrid neural architectures through surface rendering, and apply it to the challenging task of full head 3D reconstruction from few input images, down to a single image. In this case, our main objective is to achieve a significant speed-up while obtaining competitive results on geometry reconstruction. Thus, we keep the rendering setup as presented in previous papers [38, 55]. We begin by training a multi-resolution grid-based neural field to represent a statistical prior of head SDFs, and then use it during the 3D reconstruction process. We avoid dual surface/volume rendering techniques [32, 52, 54], and use direct surface rendering [55] instead. We show that, while a naive optimization would fail to converge, using a statistical prior allows us to reliably arrive to a coarse solution, which we later refine. To further accelerate and stabilise the reconstruction convergence, we include normal cues as supervision, leveraging the predictions of a monocular normal estimator[56]. In summary, our main contributions are:

- We introduce, for the first time, a framework that combines a grid-based architecture with a surface rendering method that yields to fast and accurate 3D reconstructions from one or few input images.
- We leverage on a statistical prior, obtained with thousands of 3D head models, to guide network convergence and achieve a reconstruction accuracy on a par with state of the art methods, but with $\sim$100$\times$ speed-up.
- We provide an optimal training scheme for grid-based structures combined with surface rendering methods exhibited through a variety of datasets evaluated.

2. Related work

Reconstructing objects in 3D from very few in-the-wild images, or even a single one, is a highly unconstrained problem that requires prior knowledge in order to be solved. This typically comes in the from statistical models [4, 20, 26, 34, 42, 51].

3D Morphable Models (3DMM) are the most common representation to encode statistical properties of an object category, and have been extensively used for multi-view 3D face reconstruction [2, 10, 37, 53] and single view 3D reconstruction [39, 40, 48–50]. However, while they are fast and robust, they lack flexibility. In order to capture more details, they are combined with differentiable post-processing techniques like normal maps, which capture high-frequency information [22, 40, 49].

Recently, neural fields have been proposed as a more flexible representation for 3D shape modelling [25, 27, 33]. In combination with differentiable rendering, neural fields have been successfully applied for the task of 3D object reconstruction from several images [31, 55]. In order to reduce the amount of required input information, [7, 38, 56] introduce priors that guide the optimization process towards plausible solutions. While reducing the number of input images also reduces the optimization time, these methods still require several minutes or even hours to reconstruct a single scene. Apart from presenting good results in terms of geometry reconstruction accuracy, some methods also cover face expressions as in [6], however this method is out of our scope as it uses RGB-D data and its runtime is in a higher order of magnitude.

Several efforts have been devoted to speed up the optimization process of neural fields. [41, 46] propose using meta-learning to reduce the number of steps to optimize a single model. [23] learns the integral of volumetric rendering. Other methods propose more efficient architectures that can be queried faster. [24] proposes an architecture that can use an arbitrary number of sequential layers to query the value of a neural field at expenses of constraining the frequency domain. This enables to use coarse-to-fine optimization schemes, avoiding unnecessary computations at the coarser levels. Similarly, [30, 43, 56] combine multi-resolution feature grids with shallow MLPs in order to reduce the inference time too, and to adapt resources to the optimization stage. Methods based on multi-resolution feature grids have been successfully used in combination with differentiable volume rendering to obtain 3D reconstructions [18, 56]. However, volumetric rendering requires sampling several points per ray, making it computationally intense.

While surface rendering is more efficient than volumetric rendering, since it only requires a single point per ray, it has not yet been applied to reconstruct 3D scenes in combination with multi-resolution feature grids. We believe this is

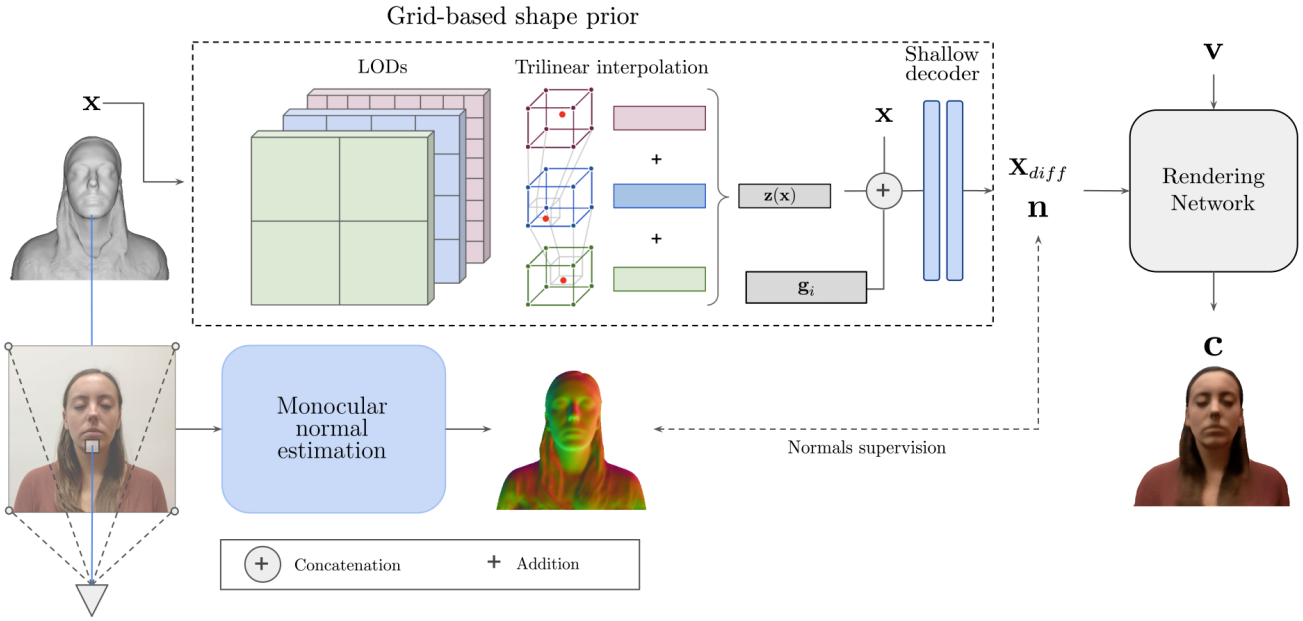

Figure 2. **Overview of our method.** For each query point $\mathbf{x}$ we obtain the feature $\mathbf{z}(\mathbf{x})$ from the multi-resolution feature grid at different levels of detail. Afterwards, we concatenate the positional encoding applied to $\mathbf{x}$, the global feature $\mathbf{g}_i$, and the grid feature $\mathbf{z}(\mathbf{x})$ to query the SDF parameterized by a shallow MLP. We supervise the gradient of the SDF with the predicted normal at the pixel location where the ray intersects. Finally, we use a rendering network to predict the radiance emitted from the surface point $\mathbf{x}_{diff}$, with normal $\mathbf{n}$, in a viewing direction $\mathbf{v}$.

due to the optimization problems derived from using a neural architecture without inductive biases for continuity plus the highly sparse sampling from surface rendering. In this paper we build and use an statistical model to enable, for the first time, optimizing a hybrid neural architecture with surface rendering.

3. Method

Our goal is to reconstruct the 3D head of a subject from a small set of images, with associated foreground masks and camera parameters. We parametrize the surface to reconstruct $\mathcal{S}$ as the zero level iso-surface of a continuous signed distance function (SDF).

To guide the reconstruction process, we build a grid-based shape statistical model that represents a prior distribution over 3D head SDFs. We extend the auto-decoder architecture proposed in [33] with a multi-resolution 3D feature grid that allows reducing the size of the MLP decoder and speeding up SDF queries. To encourage feature grids of increasing resolution to encode progressively finer geometric details we propose a curated scheduling for the prior training.

During the 3D reconstruction process, our architecture builds on top of H3D-Net [38], using differentiable surface rendering and minimising photoconsistency and silhouette errors. Crucially, at the beginning of the optimization, the SDF is constrained to be within the learned shape prior space. This allows reaching a coarse solution first, avoiding irrecoverable incorrect local minima, after which we can lift the constraint to obtain finer details.

To better guide the optimization, we supervise the gradient of the SDF with the predictions of a monocular normal map estimation model, similar as in [56]. Inspired by ray sampling in volumetric rendering approaches [28], we use an algorithm better suited for parallel computing to find the zero level set iso-surface for a given SDF that is up to 30% faster than the sphere tracing algorithm provided in [55].

We next detail each of the modules of our method. An overview of the method can be found in Figure 2.

3.1. Grid-based shape prior

Architecture. Parametrizing a space of signed distance functions with a monolithic MLP decoder is computationally expensive, since any evaluation involves the entire network. To trade compute time for memory, we extend the H3D-Net [38] framework with a multi-level dense feature grid, which allows reducing the size of the MLP decoder. To obtain the signed distance d for a query point $\mathbf{x} \in R^3$ we compute:

$$d = \mathcal{F}_\theta([\gamma_\mathcal{F}(\mathbf{x}), \mathbf{z}, \mathbf{g}]) \ , \tag{1}$$

where $\mathcal{F}_\theta$ is a shallow MLP decoder, $[\cdot, \cdot]$ denotes concatenation, $\gamma(\cdot)$ denotes sinusoidal positional encoding [47], and $\mathbf{g}$ is a latent vector encoding specific shapes in the shape space.

The vector $\mathbf{z}$ is a feature interpolated from a multi-resolution feature grid. For each level $l \in [4, 5, 6]$ we define

a dense voxel grid V_l with 2^{l*3} voxels in a bounding volume of $[-1, 1]^3$. We define an associated feature grid Z_l with $(2^l + 1)^3$ features located at the corners of each voxel in V_l. Any query point in the bounding volume can be assigned to a voxel for every resolution level, and a feature can be computed by trilinearly interpolating the corner features. To aggregate the features from different levels we sum them:

$$\mathbf{z}(\mathbf{x}) = \sum_{l \in L} \text{interp}(\mathbf{x}, Z_l) \ . \tag{2}$$

Our shape prior has been trained on a collection of coarse head scans. It consists of non-watertight meshes, indexed by $i = 1, \ldots, M$. To learn an SDF from a mesh in this set we will minimize the signed distance on the surface points, and ensure the correctness of the function with an eikonal loss term [13]. The space shape can be learned from the following objective function [38]:

$$\underset{\{\mathbf{z}, \mathbf{g}_i\}, \theta}{\arg\min} \sum_{i=1}^{M} \mathcal{L}_{\text{Surf}}^{(i)} + \lambda_0 \mathcal{L}_{\text{Emb}}^{(i)} + \lambda_1 \mathcal{L}_{\text{Eik}}^{(i)} \ , \tag{3}$$

where λ_0 and λ_1 are hyperparameters and $\mathcal{L}_{\text{Surf}}$, $\mathcal{L}_{\text{Eik}}$ and $\mathcal{L}_{\text{Emb}}$ account respectively for the SDF error at surface points, the eikonal term as in [7, 14, 38, 55] and a regularisation applied to latents $\mathbf{z}(\mathbf{x})$ and $\mathbf{g}_i$. We enforce a zero-mean multivariate Gaussian distribution with spherical covariance σ^2 over the spaces of shapes: $\mathcal{L}_{\text{Emb}}^{(i)} = \frac{1}{\sigma^2}\big(\|\mathbf{z}(\mathbf{x})\|_2 + \|\mathbf{g}_i\|_2\big)$. More details about these losses can be found at the supplementary material.

Training schedule. We aim to train the shape prior so that feature grids of increasing resolution encode progressively finer geometric details. To this end, we propose the following scheduling. We initialise with zeros the feature grids and the decoder using geometric initialization [1]. During a first stage, taking N epochs, we train the coarse grid Z_4 along with the decoder $\mathcal{F}_\theta$ and the global latent $\mathbf{g}$, leaving the finer feature grids frozen. Once a coarse fit of the scenes has been obtained, we freeze all the parameters except for the next feature grid (Z_5). After $2 * N$ epochs, we repeat the same procedure, training only the last level (Z_6) during $4 * N$ epochs. The progressively longer stages are due to the higher resolution grid, and therefore larger number of parameters, introduced in each stage.

In order to make the training process more stable and promote low frequency shapes to be encoded by coarse levels, we use a progressive masking of the positional encoding $\gamma(\cdot)$ [21]. At the beginning of the optimization process all frequencies are masked and it is being unmasked progressively between epochs 0 and $N/2$.

3.2. Surface reconstruction

To obtain the 3D reconstruction of a scene given posed images and foreground masks, we follow a surface rendering approach to optimize our pre-trained statistical model, similar as in [38]. For every pixel, we march a ray and find the ray-surface intersection point, $\mathbf{x}$, which we make differentiable with respect to the network parameters through implicit differentiation [55], which we denote by $\mathbf{x}_{diff}$. We then render the object at this point, with normal $\mathbf{n}$, seen from the ray viewing direction $\mathbf{v}$, using a differentiable rendering function $\mathcal{G}_\phi(\mathbf{x}, \mathbf{n}, \mathbf{v})$, with learnable parameters ϕ. In addition to the usual losses penalising photoconsistency error, and silhouette error, we add a loss term supervising the surface normal vectors with the predictions of a monocular normal map regressor. We use a cosine similarity loss

$$\mathcal{L}_{\text{norm}} = 1 - \nabla_{\mathbf{x}} d \cdot \hat{\mathbf{n}} \ , \tag{4}$$

where d is the SDF value as defined in eq. 1, so its gradient at the surface corresponds to the normal vector, and $\hat{\mathbf{n}}$ is the prediction of our monocular normal map estimation model.

The scheduling used to optimize InstantAvatar for a specific scene closely follows the one proposed in [38]. First, we freeze the weight of the decoder and we train the global feature vector $\mathbf{g}_i$ and all the feature grids Z_l during 30 epochs. Afterwards, we unfreeze the decoder $\mathcal{F}_\theta$ until epoch 100.

It is worth mentioning that, taking into account the feature grid architecture being used, we are then not able to make full use of the eikonal equation due to the intervoxel non-continuity of the SDF derivatives with respect to $\mathbf{x}$. In addition, the eikonal loss has minima which are not SDFs as explained in [36], for this reason, the eikonal regularization is disabled during the surface reconstruction process.

Monocular normals estimation. Our architecture consists of a U-Net++ [58] with an EfficientNet-B1 encoder [45] pre-trained on ImageNet classification. We train the model to predict a normal map from a single input image, expressing normals in the camera frame of reference. The training data is generated from a collection of raw head scans paired with posed images, which we use to render ground truth normal maps. We generate 50k pairs of images and normal maps, covering $\pm 90^{\text{o}}$ from the frontal view. See more details in the supplementary material.

Ray-surface intersection finding algorithm. We take inspiration from volumetric rendering approaches [28] to accelerate the ray tracing algorithm introduced in [55]. Instead of using an iterative procedure to find the roots of an SDF, we propose a parallelizable approach more suited for high-end GPUs. First, we sample N_p points along each ray from the pixel towards the bounding volume in a viewing direction $\mathbf{v}$. Secondly, we select the first interval between points where the field has a sign change and we sample N_f points in this segment. We return as the intersection coordinate, the first point where the field has a zero crossing. See more details in the supplementary material.

3.3. Implementation details

We now describe the implementation details of our architecture and how we train it. Before passing the input coordinates $\mathbf{x}$ to the decoder network, we apply a positional encoding $\gamma_{\mathcal{F}}(\mathbf{x})$ with 6 log-linear spaced frequencies. The encoded 3D coordinates are concatenated with the $\mathbf{g}_i$ global latent vector of size 256, and the feature interpolated from a multi-resolution feature grid $\mathbf{z}$ of length 8 and set as the input to the decoder. The decoder $\mathcal{F}_\theta$ is an MLP of 3 layers with 512 neurons in each layer. We use Softplus as activation function in every layer except for the last one, where no activation is used. As in [38], the rendering function $\mathcal{G}_\phi(\mathbf{x}, \mathbf{n}, \mathbf{v})$ is implemented as a composition of two subnetworks $\mathcal{Q}_\rho$ and $\mathcal{R}_\eta$. $\mathcal{Q}_\rho(\gamma_{\mathcal{Q}}(\mathbf{x}))$ is an MLP of 8 layers with 512 neurons in each layer and a single skip connection from the input of the network to the output of the 4th layer. We use Softplus as activation function at every layer except for the last, where no activation is used. The $\mathcal{Q}_\rho$ output is a 256-dimensional vector $\mathbf{l}$ that is concatenated to $\gamma_{\mathcal{R}}(\mathbf{x})$ and $\mathbf{n}$ and provided to $\mathcal{R}_\eta$ as input. As in [55], $\mathcal{R}_\eta(\gamma_{\mathcal{R}}(\mathbf{x}), \mathbf{n}, \mathbf{l}, \mathbf{v})$ is an MLP composed by 4 layers, each 512 neurons wide, no skip connections, ReLU activations for every layer except for the output layer which has tanh and outputs the 3-dimensional RGB color. We also apply the positional encodings $\gamma_{\mathcal{Q}}$ and $\gamma_{\mathcal{R}}$ to $\mathbf{x}$ with 6 and 4 log-linear spaced frequencies respectively. The prior is trained for 700 epochs in total (MLP, global latent and Z_4 during the first 100 epochs, then Z_5 during 200 epochs, and lastly Z_6 for 400 epochs), using Adam [17] with standard parameters, learning rate of 10^{-4} and learning rate step decay of 0.5 every 50 epochs. The value parameterizing the length of the different stages of the prior training is $N = 100$, so the entire training takes $7N = 700$ epochs. The shape prior training takes approximately 24 hours for a dataset of 10k scenes.

4. Experiments

In this section we evaluate InstantAvatar in single and multi-view settings, and report both qualitative and quantitative results. We compare against the 3DMM-based methods: DECA [11], Dib et al. [9], DFNRMVS [2], MVF-Net [53] and against neural fields methods: Multi-NeuS [5], H3D-Net [38] and SIRA [7]. We conduct these experiments on the following public datasets: H3DS [38], 3DFAW [35] and CelebA-HQ [15].

4.1. Datasets

Training. We build the probabilistic shape prior, and train the monocular normals prediction model, with the same dataset used in [38] and [7]. It is made of 10k scenes composed of 3D head scans paired with multi-view posed images. The dataset is balanced in gender and diverse in age and ethnicity.

3DFAW. This dataset contains videos of human heads paired with 3D reconstructions of the facial area. It also contains high resolution photos taken with a professional camera. We select the same 10 cases from the low-resolution set as in [38], and 17 subjects from the high-resolution set provided by [9], for comparison purposes.

H3DS. It consists of 23 human head scenes with multi-view posed images, masks, and full-head 3D textured scans. The dataset consists of 13 men and 10 women.

CelebA-HQ. High-quality version of CelebA that consists of 30k in-the-wild frontal images at 1024×1024 resolution. At Figure 6 we have selected a subset of 8 scenes that are diverse in gender and ethnicity.

4.2. Experimental setup

For every evaluation dataset, we align the ground truth and the reconstructions using manually annotated facial landmarks. We refine this alignment by performing ICP [3] of the ground truth face region to the reconstruction. The face region is defined by all the vertices falling under a sphere of radius 95mm centered at the nose.

After aligning the meshes, we over-sample points from the reconstructions, rejecting points that are too close to each other. We use the unidirectional Chamfer distance from the defined regions on the ground truth to the predicted reconstructions.

4.3. Ablation study

We conduct an ablation study, evaluating variations of our method in the multi-view reconstruction setting (3 and 6 views) on all cases of the H3DS dataset. We show the qualitative results in Figure 3 and the quantitative results in Figure 4 and Table 1.

A key ingredient to optimize a grid-based SDF with differentiable rendering is to build a prior distribution over 3D head shapes. We observe that when using our architecture without shape prior (see Table 1), the per-scene reconstruction does not converge to plausible human head.

When training InstantAvatar without a grid, we observe that the decrease in model capacity leads to a decrease in the capacity to represent high-frequencies, which particularly affects the face region. As seen in Figure 3, this has a big impact in the ability to preserve the identity of the subject. To validate it, we have trained a 3-layer and 8-layer MLPs (each layer composed of 512 neurons) without grids and with predicted normals supervision. We have used these two MLP sizes in order to explore the model capacity-speed tradeoff and compare both to our final approach. As shown in Figure 4, the multi-grid based method (ours) is the one that converges faster.

As shown in Table 1, we obtain the smaller error in the face metric when we introduce normal cues as supervision for the gradient of the SDF during the optimization process.

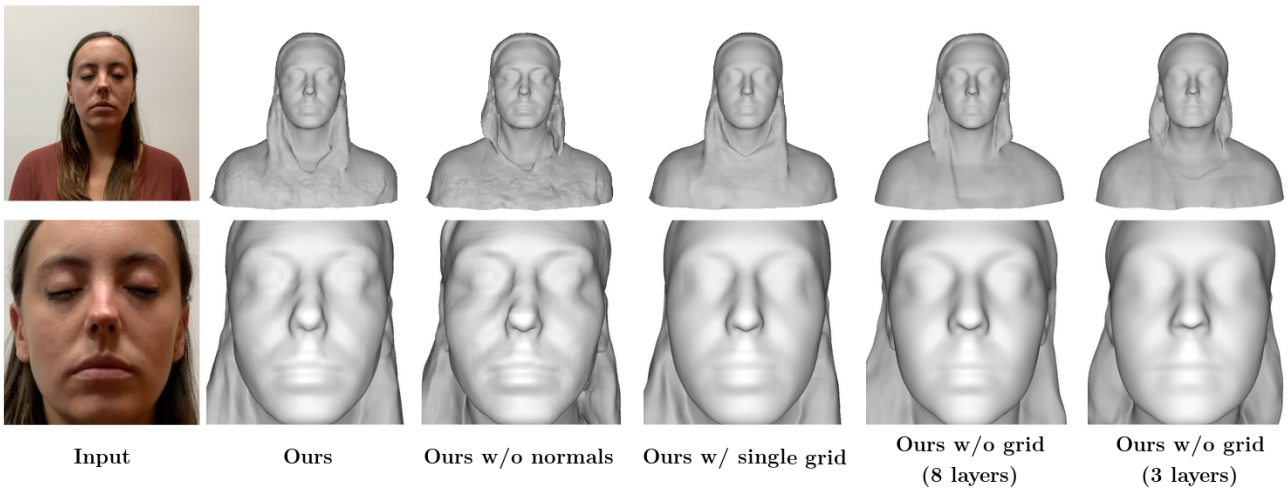

Figure 3. **Ablation: Qualitative comparison.** We conduct an ablation study to qualitatively compare variations of our model using a H3Ds dataset scene in the multi-view setting (6 views). The bottom row zooms into the face region to better appreciate the differences among configurations. Both our final approach and the one without normals supervision outperform the rest of alternatives. However, when normals supervision is not considered the resulting shape tends to be excessively sharp (e.g. the outermost part of the eyebrows) or erroneous (hair). The single grid and the 8-layer MLP (without grid) results are comparable, although they are both unable to capture the high-frequency details obtained with our final model .

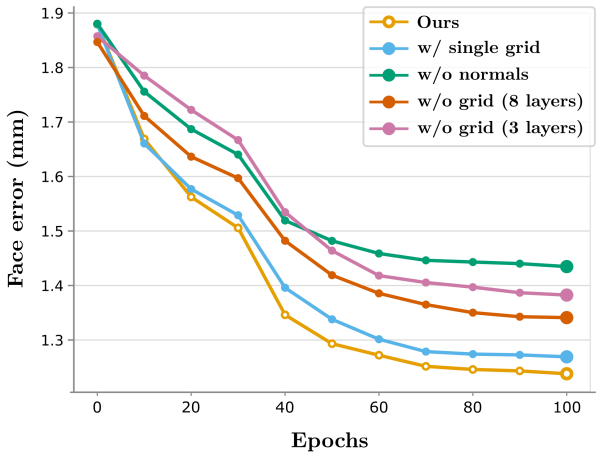

Figure 4. **Ablation: Convergence speed.** Convergence speed across epochs averaged over all cases of the H3Ds dataset in the multi-view setting (6 views). Grid-base methods show a faster convergence in contrast to MLPs approaches. The supervision of normals helps convergence in the fine-tuning process at the reconstruction stage.

We have done an extensive hyperparameter space exploration, and confirmed that the enhanced robustness provided by the normals ensures consistent accuracy. Using a single coarse grid provides robust reconstructions but lacks detail, in contrast, the multi-grid approach is able to represent better the high-frequency details. However, this improvement in the high-frequency domain may not be fully reflected at the shown metrics (Table 1) and would be part of potential future work, as it is an ongoing research subject [8].

In Table 2, we compare the performance of different shape prior training schedules, where our final approach with both progressive levels of detail schedule and progressive positional encoding masking outperforms the ablated setups.

	3 views	6 views
	face ↓	face ↓
Ours w/o shape prior	6.98	20.72
Ours w/o normals	1.59	1.44
Ours w/o grid (3 layers)	1.55	1.38
Ours w/o grid (8 layers)	1.46	1.34
Ours w/ single grid	**1.39**	1.27
Ours	**1.39**	**1.22**

Table 1. **Ablation: Quantitative comparison.** Each layer of the MLPs are composed of 512 neurons. Error in millimeters. Average over all cases of the H3Ds dataset in the multi-view setting (3 and 6 views).

	1 view	3 views	6 views
	face ↓	face ↓	face ↓
(A) Ours w/o progressive PE masking	1.77	1.48	1.33
(B) Ours w/o progressive LODs schedule	1.70	1.54	1.35
(C) **Ours**	**1.50**	**1.39**	**1.22**

Table 2. **Shape prior training schedules ablation.** Reconstruction metrics (unidirectional Chamfer distance from ground-truth to predicted reconstructions) using different prior schedules. We compare our final shape prior schedule (C), which progressively optimizes the levels of detail from coarse to fine, with another schedule that optimizes every level at the same time (B). We also provide an ablation where we have removed the progressive positional encoding masking (A), therefore is fully operative from the start.

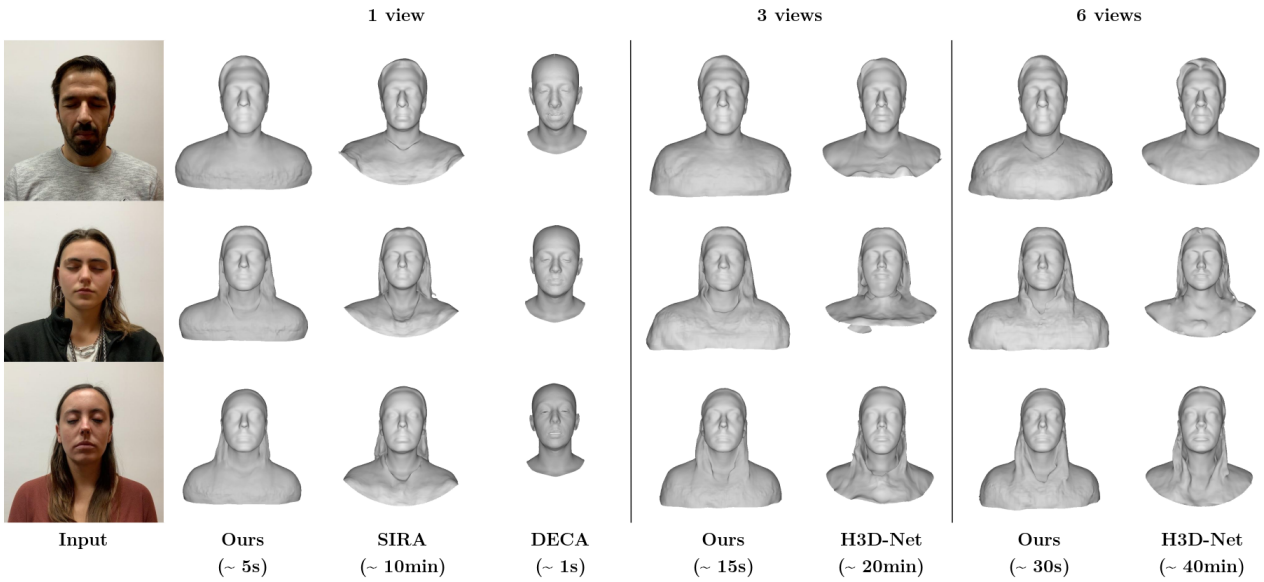

Figure 5. **Qualitative results.** We compare qualitatively InstantAvatar with other state of the art methods on H3DS dataset for 1 view, 3 views and 6 views. **1 view:** SIRA is able to better capture the identity of the subject, however, it takes 10 min of training. DECA on the other hand, can only predict the face region. **3 views:** H3D-Net achieves good bias but at a high variance where we can clearly see artefacts on the chin and the hair. **6 views:** H3D-Net is able to recover the hair and face regions with similar quality as InstantAvatar.

4.4. Quantitative results

Single-view reconstruction. We perform reconstructions from front facing images, and report the surface error in Table 3. We see that our method achieves competitive error metrics, while being more than two orders of magnitude faster than state-of-the-art methods based on differentiable rendering. More specifically, InstantAvatar is the only method capable of performing full-head reconstruction, including hair and garments, in seconds.

	1 view				
	3DFAW	3DFAW HR	H3DS		
	face ↓	face ↓	face ↓	head ↓	time ↓
Dib et al.	1.99	2.16	2.06	-	∼ **1s**
DECA	1.57	**1.43**	1.74	22.22	∼ **1s**
Multi-NeuS	-	-	1.81	13.92	∼ 1h
H3D-Net	1.63	1.47	1.97	14.36	∼ 10min
SIRA	**1.36**	1.51	**1.39**	14.43	∼ 10min
Ours	1.58	1.69	1.50	**12.43**	∼ **5s**

Table 3. **3D reconstruction method comparison (single-view).** Error in millimeters. Time in seconds, approximate upper bound. Methods above the division line are 3DMM-based and methods below are model-free.

Multi-view reconstruction. In this comparison, the reconstructions are done with 3 and 6 views taken from a similar distance and height. For 3 views, we use a frontal image directly looking at the head and two views placed at ± 45 yaw angle. For 6 views, we additionally use two lateral views at ± 90 degrees, and a frontal image but with a lower pitch angle.

	3 views				6 views		
	3DFAW	H3DS			H3DS		
	face ↓	face ↓	head ↓	time ↓	face ↓	head ↓	time ↓
MVF-Net	1.61	1.67	-	∼ **1s**	-	-	-
DFNRMVS	1.74	1.96	-	∼ **1s**	-	-	-
H3D-Net	**1.32**	1.47	11.13	∼ 20min	1.24	**6.24**	∼ 40min
Ours	**1.32**	**1.39**	**9.71**	∼ **15s**	**1.22**	8.03	∼ **30s**

Table 4. **3D reconstruction method comparison (multi-view).** Error in millimeters. Time in seconds. Methods above the division line are 3DMM-based and methods below are model-free.

In Table 4 we can see the multi-view results. We achieve better results compared with the parametric methods using 3 views, and for both view configurations we get comparable results to the implicit reconstruction methods, while being two orders of magnitude faster. Note that the normals estimation process takes less than 1 second, therefore it does not affect significantly the final time results.

Grid-based architecture efficiency. We propose a grid-based approach in order to speed up the query time of the network that models the SDF. In surface rendering, to find the intersection point it is needed to query multiple times along each ray. Therefore, replacing the monolithic MLP architecture by a grid-based approach with multiple trainable local features per grid and a shallow MLP, can significantly reduce this query time, as the local features indexing

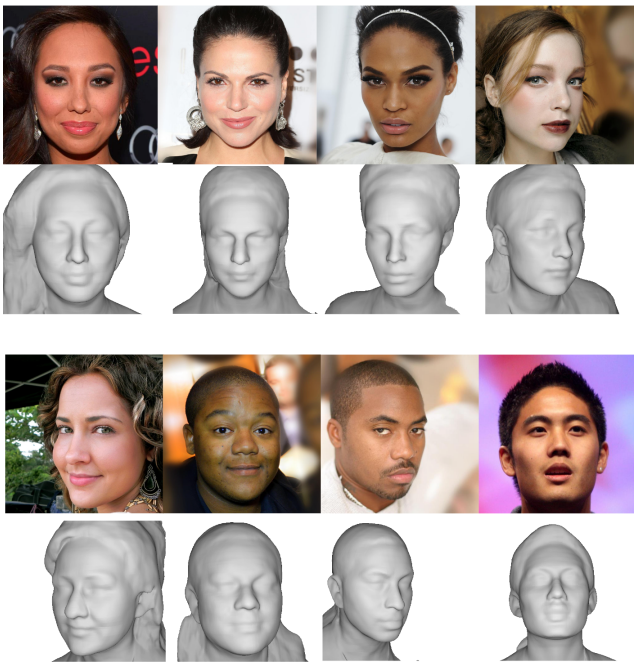

Figure 6. **Qualitative results.** InstantAvatar results on CelebHQ dataset for a single input image.

is significantly faster than computing a MLP forward with extra layers.

4.5. Qualitative results

Qualitative results show comparable quality reconstruction in all the datasets while being up to 100x times faster. Our model is competitive in the one shot scenario, as well as in a few-shot setup, as shown in Figure 5. Unlike 3DMM-based approaches, which only provide an accurate reconstruction of the face region, InstantAvatar is able to recover the identity by providing the geometry of the full head, including hair and shoulders.

Single-view reconstruction. As expected, the SIRA model [7] performs best in the single-image setting, as it is designed to achieve high accuracy in this configuration, but at the run-time cost of several minutes of reconstruction time. However, InstantAvatar is able to obtain competitive visual results in a matter of seconds. On the other hand, the 3DMM-based models [9, 11] struggle to capture fine anatomical details and they don't model hair or shoulders as stated in [38]. Furthermore, we can see that our model is able to obtain more robust and natural results in areas such as the shoulders, which are challenging to estimate from a single view.

It is essential that the method is not biased and does not discriminate against any group, religion, colour, gender, age, or disability status. Therefore, we provide in Figure 6 the results on the Celeb-HQ dataset composed of in-the-wild single images from the frontal view, to show that our model is diverse.

Multi-view reconstruction. We evaluate the performance when increasing the number of input views on the reconstruction surface, in both the face and full head regions. As InstantAvatar is a method that requires optimization per-scene, we consistently see improvement as the number of views increases. As we can see in Figure 5, we perform on par with the state-of-the-art implicit reconstruction method [38] in terms of Chamfer distance in both the face and full-head regions. Remarkably, InstantAvatar is able to better reconstruct challenging high frequency regions such as the hair, as it can be seen in the 3th row of Figure 5.

5. Conclusions

We have demonstrated that it is possible to push the speed of implicit reconstruction methods based on neural fields by using a hybrid grid-based architecture in combination with surface rendering. We have shown that, while a naive combination of these two ideas leads to poor results due to unstable optimisation, these difficulties can be overcome by introducing appropriate geometry priors. In particular, we have demonstrated the effectiveness of pre-training the geometry network as a statistical shape model, and of using monocular normal predictions as cues, to stabilise the optimisation. Using these techniques, we achieve comparable state-of-the-art 3D reconstructions of human avatars with a 100x speedup over neural-field-based alternatives. This reduced runtime can significantly help the widespread adoption of these 3D reconstruction methods.

6. Limitations and future work

First, an architecture based on dense grids may have some memory concerns, which have been discussed and handled already [44] and was not at our primary scope, which was mainly speed. Also, it is also important to mention that the final representation capacity of our architecture is somehow limited by the quality of the normals predicted.

Second, despite its numerous advantages, grid representations combined with surface rendering still being an under-determined problem. In particular, in absence of a stronger inductive bias, we have showed how beneficial it is to guide this field with the help of prior knowledge. However, the commented prior architecture is limited in such a way that only a half of it - its decoder side - is optimized into its latent space for each different scene. The feature grid side of the prior, conversely, does not depend on the scene so it simply becomes a good initialization, which eases convergence, but does not let us make full use of it. Accordingly, an appropriate prior architecture which would let the grid local features to be optimized throughout its prior latent space, could potentially unlock the full capacity of these grid representations.

References

[1] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[2] Ziqian Bai, Zhaopeng Cui, Jamal Ahmed Rahim, Xiaoming Liu, and Ping Tan. Deep facial non-rigid multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 5

[3] Paul J Besl and Neil D McKay. A method for registration of 3-d shapes. In *ACM Transactions on Graphics (TOG)*, 1992. 5

[4] Alan Brunton, Timo Bolkart, and Stefanie Wuhrer. Multilinear wavelets: A statistical shape space for human faces. In *Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV)*, 2014. 2

[5] Egor Burkov, Ruslan Rakhimov, Aleksandr Safin, Evgeny Burnaev, and Victor Lempitsky. Multi-neus: 3d head portraits from single image with neural implicit functions. *arXiv preprint arXiv:2209.04436*, 2022. 5

[6] Chen Cao and Tomas Simon. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (TOG)*, 2022. 2

[7] Pol Caselles, Eduard Ramon, Jaime Garcia, Xavier Giro-i Nieto, Francesc Moreno-Noguer, and Gil Triginer. Sira: Relightable avatars from a single image. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. 1, 2, 4, 5, 8

[8] Zenghao Chai, Haoxian Zhang, Jing Ren, Di Kang, Zhengzhuo Xu, Xuefei Zhe, Chun Yuan, and Linchao Bao. Realy: Rethinking the evaluation of 3d face reconstruction. In *Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV)*, 2022. 6

[9] Abdallah Dib, Cedric Thebault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 5, 8

[10] Pengfei Dou and Ioannis A Kakadiaris. Multi-view 3d face reconstruction with deep recurrent neural networks. *Image and Vision Computing*, 80:80–91, 2018. 2

[11] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 5, 8

[12] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[13] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of Machine Learning and Systems (MLSys)*. 2020. 4

[14] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 4

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 5

[16] Petr Kellnhofer, Lars C Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. Neural lumigraph rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[18] Hai Li, Xingrui Yang, Hongjia Zhai, Yuqian Liu, Hujun Bao, and Guofeng Zhang. Vox-surf: Voxel-based implicit surface representation. *arXiv preprint arXiv:2208.10925*, 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 (TOG)*, 36(6):194–1, 2017. 1

[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 Transactions on Graphics (TOG)*, 36(6):194:1–194:17, 2017. 2

[21] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 4

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

[23] David Lindell, Julien Martel, and Gordon Wetzstein. AutoInt: Automatic integration for fast neural volume rendering. *arXiv preprint arXiv:2012.01714*, 2020. 2

[24] David B Lindell, Dave Van Veen, Jeong Joon Park, and Gordon Wetzstein. Bacon: Band-limited coordinate networks for multiscale scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[25] Gidi Littwin and Lior Wolf. Deep meta functionals for shape representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015. 2

[27] 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 (CVPR)*, 2019. 2

[28] 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. In *Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV)*, 2020. 1, 3, 4

[29] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 2

[30] Thomas Müller, Alex Evans, Christoph Schied, and Alexan-

der Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41(4):102:1–102:15, 2022. 2

[31] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[32] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[33] 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 (CVPR)*, 2019. 2, 3

[34] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments*, 2009. 1, 2

[35] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCV Workshops)*, 2019. 5

[36] Albert Pumarola, Artsiom Sanakoyeu, Lior Yariv, Ali Thabet, and Yaron Lipman. Visco grids: Surface reconstruction with viscosity and coarea grids. In *Advances in Neural Information Processing Systems (NeurIPS)*. 4

[37] Eduard Ramon, Janna Escur, and Xavier Giro-i Nieto. Multi-view 3d face reconstruction in the wild using siamese networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCV Workshops)*, 2019. 2

[38] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giró-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 (ICCV)*, 2021. 1, 2, 3, 4, 5, 8

[39] Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data. In *International Conference on 3D Vision (3DV)*, 2016. 2

[40] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[41] Vincent Sitzmann, Eric R. Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. Metasdf: Meta-learning signed distance functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[42] William A. P. Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua Tenenbaum, and Bernhard Egger. A morphable face albedo model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[43] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[44] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 8

[45] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019. 4

[46] Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P. Srinivasan, Jonathan T. Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[47] 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. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[48] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017. 2

[49] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard G Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[50] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

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

[52] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1, 2

[53] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 5

[54] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Vol-

ume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2

[55] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2, 3, 4, 5

[56] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3

[57] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[58] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018. 4