Implicit Shape and Appearance Priors for Few-Shot Full Head Reconstruction

Pol Caselles[1,2], Eduard Ramon[2,3*], Jaime García[2], Gil Triginer[2†], Francesc Moreno-Noguer[1,3*]

[1] Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain
[2] Crisalix SA [3] Amazon

***Abstract*—Recent advancements in learning techniques that employ coordinate-based neural representations have yielded remarkable results in multi-view 3D reconstruction tasks. However, these approaches often require a substantial number of input views (typically several tens) and computationally intensive optimization procedures to achieve their effectiveness. In this paper, we address these limitations specifically for the problem of few-shot full 3D head reconstruction. We accomplish this by incorporating a probabilistic shape and appearance prior into coordinate-based representations, enabling faster convergence and improved generalization when working with only a few input images (even as low as a single image). During testing, we leverage this prior to guiding the fitting process of a signed distance function using a differentiable renderer. By incorporating the statistical prior alongside parallelizable ray tracing and dynamic caching strategies, we achieve an efficient and accurate approach to few-shot full 3D head reconstruction.**

Moreover, we extend the H3DS dataset, which now comprises 60 high-resolution 3D full-head scans and their corresponding posed images and masks, which we use for evaluation purposes. By leveraging this dataset, we demonstrate the remarkable capabilities of our approach in achieving state-of-the-art results in geometry reconstruction while being an order of magnitude faster than previous approaches.

***Index Terms*—Neural Radiance Field, Signed Distance Function, Few-shot 3D Reconstruction**

I. Introduction

In recent years, the digitalization of humans has emerged as an important area of research. The ability to create accurate digital representations of individuals holds immense value across a wide range of applications, including Virtual Reality (VR), Augmented Reality (AR), healthcare, entertainment, and security. To achieve this, the process often entails capturing photographs of a scene using conventional cameras or mobile devices. However, generating 3D reconstructions under non-controlled conditions or with a limited number of input images can be particularly challenging [1]–[4]. Among these challenges, the single-view setup stands out as the most ill-posed scenario. It involves a highly under-constrained problem that cannot be effectively solved without prior knowledge or additional information [5]–[10].

Statistical priors based on 3D Morphable Models [1]–[8], [10], [11] have become the standard approach for few-shot 3D face reconstruction. By adopting 3DMMs as a representation, the problem of 3D reconstruction can be simplified to estimating a small set of parameters that effectively capture a target 3D shape. This enables the generation of 3D reconstructions from small sets of images [1]–[4], even when only a single view is available [5]–[8], [10]. However, one significant drawback of morphable models is their limited expressiveness, particularly for capturing high-frequency details. To address this issue, researchers have explored post-processing techniques that transfer fine details from the image domain to enhance the 3D geometry [8], [10], [12]. Another limitation of 3DMMs is their inability to represent complex shapes and diverse topologies. Consequently, they are not well-suited for reconstructing complete heads with features such as hair, beard, facial accessories, and upper body clothing.

Model-free approaches based on discrete representations such as voxels [13], meshes or point clouds offer greater flexibility in representing a wide range of shapes. However, they come with computational limitations, as they face scalability challenges as resolution increases or are limited to fixed topologies. Despite significant advancements in computational resources in recent years, there remains a trade-off between resolution and memory. Overcoming these challenges, neural fields [14]–[21] have emerged as a solution that encodes both geometry and appearance as a continuous coordinate-based function within the weights of a neural network. Notably, recent work by [22], [23] has demonstrated the success of such representations in learning detailed 3D geometry directly from images, even in the absence of 3D ground truth supervision. Unfortunately, these methods currently rely on a significant number of input views, leading to time-consuming inference and limiting their applications.

Optimization-based techniques iteratively refine model parameters to minimize a cost function, resulting in high accuracy and fine detail capture [24], [25]. On the contrary, feed-forward methods highly rely on the quality and variability of the training data which limits their ability to reconstruct fine details in out-of-distribution samples at test time [26]. Nonetheless, they offer the advantage of speed and computational efficiency, making them suitable for real-time applications [27]–[31].

Recent advancements in this field, highlighted by H3D-Net [32] leverage large 3D scan datasets to integrate prior geometric knowledge into neural field models, allowing for significantly improved accuracy of full-head 3D reconstruc-

* Work done prior joining Amazon.
†Work done prior joining Astrazeneca.

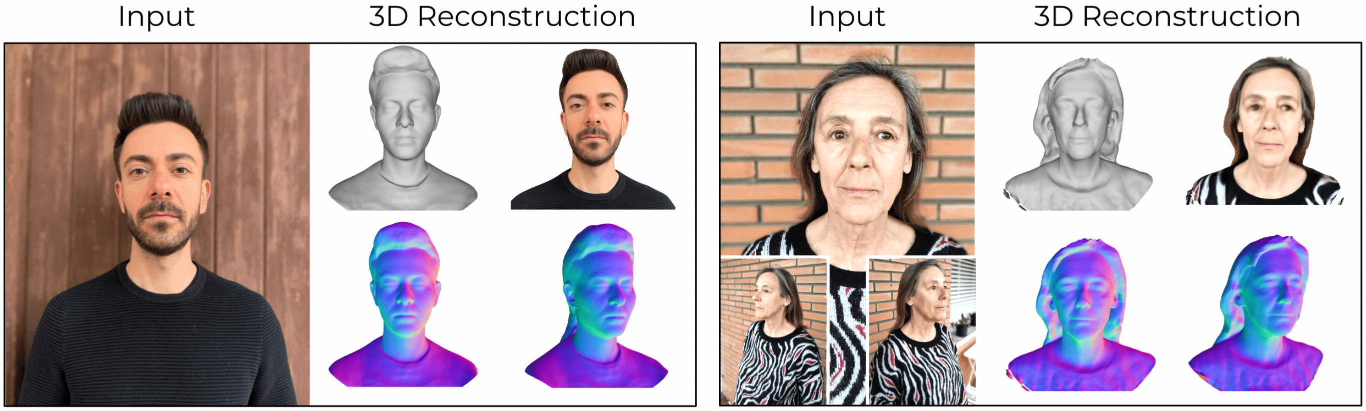

Fig. 1. **Few-shot full-head reconstruction using SIRA++**. Our approach enables high-fidelity 3D head reconstruction using only a few images. The figure showcases two examples, one obtained from a single input image (in 92 seconds) and the other from three input images (in 191 seconds). For each example, we present the input image/s on the left and the corresponding reconstruction on the right, including the 3D mesh, rendered mesh, and normal maps. The results demonstrate the effectiveness and efficiency of our method in generating detailed 3D head avatars with minimal input data.

tions, even when only a limited number of input images are available. Nonetheless, it is worth noting that [32] still faces challenges in terms of computational cost, demanding several hours of optimization per scene at inference time, and it is not suitable for scenarios where only a single input image is available.

To address the limitations of H3D-Net, we propose a novel prior that combines shape and appearance information. Our approach leverages an extensive dataset of 10.000 textured 3D head scans to pretrain an architecture consisting of two primary neural field decoders. The first decoder focused on modeling the complete geometry of the head using a signed distance function (SDF) as a 3D representation. The second decoder aims to capture the head appearance, including the face region, hair, and clothing of the upper torso. During inference, we utilize this pre-trained appearance and shape prior to initialize and guide the optimization of an Implicit Differentiable Renderer (IDR) [23] that, given a reduced number of input images, estimates the full head geometry. The learned prior facilitates faster convergence during optimization (∼191 seconds per 3 input images) and prevents the model from getting trapped in local minima. As a result, our method produces 3D shape estimates that capture fine details of the face, head, and hair from just a single input image.

The core of our approach, dubbed SIRA, has been previously presented in [33]. In this journal submission, we expand upon [33] through innovative strategies to efficiently eliminate non-intersecting rays. These strategies, combined with a parallelizable ray tracing algorithm and dynamic caching, result in a remarkable acceleration of over 10× compared to the previous implementation. We refer to this enhanced version of the approach as SIRA++.

Moreover, in this submission, we provide an in-depth analysis of SIRA++ for multi-view setup, including the joint shape and appearance priors, camera noise robustness, extended comparison with new 8 SOTA methods, as well as extensive evaluation in-the-wild CelebA-HQ dataset. In addition, we have made significant enhancements to the H3DS dataset, which consists of high-resolution 3D full head scans, images, masks, and camera poses, originally introduced in [32]. Our expansion has considerably increased the dataset size from 10 to 60 samples, resulting in a more extensive and diverse collection of data. This enhanced H3DS dataset serves as the foundation for a thorough evaluation of the performance of SIRA++. We compare our method with H3D-Net [32], SIRA [33], and other mesh and field based state-of-the-art approaches.

The experimental results demonstrate that SIRA++ surpasses H3D-Net by a substantial margin in terms of reconstruction error of the full head. Moreover, SIRA++ achieves comparable performance to SIRA but with significantly lower computational cost. Additionally, we compare SIRA++ with recent approaches that are based on parametric models, solely providing a 3D reconstruction of the face region rather than the entire head. SIRA++ consistently demonstrates improved results compared to these methods.

In summary, this work builds upon our earlier versions of H3D-Net [32] and SIRA [33], which were pioneering approaches in utilizing implicit functions for the reconstruction of full 3D human heads from a limited number of images. In this submission, we enhance the reconstruction accuracy of H3D-Net by introducing a novel data-driven prior that combines shape and appearance. Additionally, we significantly improve the computational efficiency of SIRA by implementing parallelizable ray tracing and dynamic caching strategies. These advancements result in an algorithm that is highly efficient and accurate for full 3D head reconstruction, even when only a few images are available (including the most challenging case of a single image). Fig. 1 shows two examples of the reconstructions we obtain from one (in 92 seconds) or five (in 191 seconds) input images.

Furthermore, this submission includes an extended version of the H3DS dataset, which we will make publicly available for evaluation purposes. This expanded dataset will provide researchers and practitioners with a valuable resource for assessing and benchmarking their algorithms and methodologies in the field of 3D head reconstruction.

II. Related Work

In this section, we review the related work on face and full head reconstruction, with a primary focus on statistical

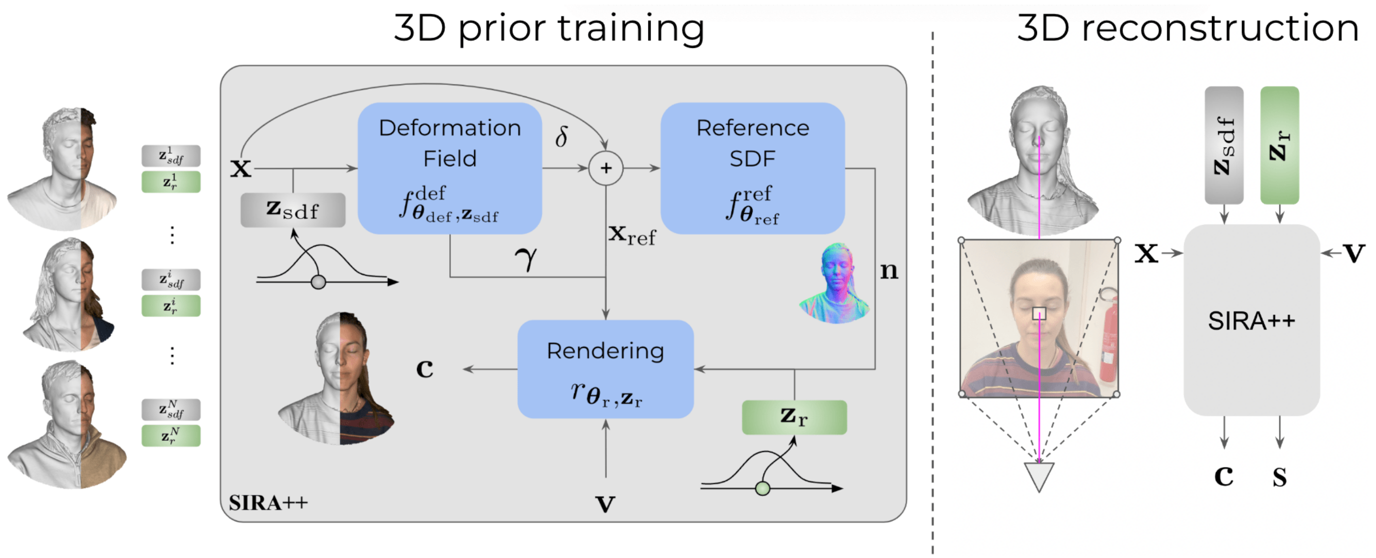

Fig. 2. **Overview of SIRA++. Left:** We construct a surface appearance statistical model using a dataset of raw head scans paired with multiview posed images. This involves learning a codebook of shapes $\mathbf{z}_{\text{sdf}}$ and appearances $\mathbf{z}_{\text{rend}}$, alongside two decoders that approximate a signed distance function and a renderer. The prior is trained using an autodecoder approach. **Right:** The pre-trained prior model is integrated with the implicit differentiable renderer. To begin the optimization process with a plausible human head, we sample from the manifold of shape and appearance latents. During the initial iterations, our focus is on training the latents to approximate the closest human head within our statistical model. Subsequently, we unfreeze the deformation and rendering networks, enabling fine-tuning of the fine details. Throughout the entire optimization phase, the reference network remains frozen, ensuring consistent results.

morphable models and recent advancements in neural fields for 3D reconstruction. Note that while there exists extensive literature on 3D body reconstructions, our paper specifically narrows its scope to those dealing with the full head region.

Statistical morphable models. The utilization of 3D Morphable Models (3DMMs) has become the predominant paradigm for face reconstruction from images, particularly in single-view or few-shot scenarios. These statistical models [34]–[36] are widely adopted and primarily focus on the face region. In the context of single-shot methods, remarkable capabilities in face reconstruction have been demonstrated by [5], [27]–[30], even in challenging in-the-wild scenes. In an alternative approach, [37], [38] employ a generative adversarial network for 3D face fitting. Recent works have also employed 3DMMs to estimate 3D geometry and spatially varying surface properties, such as diffuse and specular albedos, along with global illumination properties [28], [39], [39].

Achieving higher fidelity face reconstructions can be accomplished by leveraging multiple input images [1], [3], [40]. Previous approaches have predominantly focused on optimizing model parameters through a multi-view analysis-by-synthesis strategy [28], [39]. To enhance robustness in specific face areas, [41] utilizes occlusion cues. Furthermore, [27] proposes learning the statistical model within a metric space to effectively capture variations in scale. Other works [42] introduce emotional information to improve expression capture. Nevertheless, While these models have advanced 3D face reconstruction, they struggle to capture fine anatomical details and they are limited to the face region.

Neural fields for 3D reconstruction. In recent years, neural fields have emerged as the leading approach for scene representation [43], yielding remarkable results in novel view synthesis [16], [20], [24], [44], [45] and 3D reconstruction [22], [23], [32], [46]. These techniques have found practical applications in modeling full-head avatars [44], [47]–[49]. By leveraging surface priors [17] and surface rendering [23], neural fields enable highly accurate 3D reconstructions of the full head, encompassing intricate details such as hair and shoulders [32], [48]. [50] employs a similar approach to construct implicit morphable faces with consistent texture parameterization and introduces single-shot inversion to obtain reconstructions from input images. This method, however, requires substantial computational time, with a single scene taking approximately 3 hours to process.

In an effort to enhance the geometric detail of morphable models, several approaches have combined them with implicit representations. For example, [51] enhances a morphable model representation through the use of an implicitly learned displacement field. However, this approach is limited to the face region. On the other hand, full head reconstruction is achieved in [52] through a feed-forward network that learns vertex displacements throughout the entire head. Another hybrid representation is proposed in [53], which combines the fine-grained control mechanism of 3DMMs with the high-quality representation of implicit functions parametrized by neural networks. This approach enables the generation of animatable full-head avatars from videos. In an effort to expedite the training and rendering process, [54] introduces a deformable point-based representation. Unfortunately, all of these approaches still rely on a significant number of input frames, which diverges from the few-shot or single-shot scenarios considered in this paper.

Recently, model-free approaches in combination with pixel-aligned features [19], [55]–[59] have emerged as an approach to obtain fast reconstructions as they don't require test-time optimization. PIFU [19] introduced the concept of pixel-aligned features to condition a learned occupancy field from single to multiple input images and Phorum [57] extended it by using a signed distance field as a shape representation.

JIFF [60] combines features from a face morphable model to enhance high-frequency details and KeypointNeRF [61] aggregates pixel-aligned features with a relative spatial encoder using volumetric rendering. However, pixel-aligned single feed-forward methods are still behind optimization-based approaches in terms of quality reconstruction. As we will show in the experimental section, these methods do not always guarantee realistic and accurate full-head reconstructions and are sensitive to camera pose estimation as they don't optimize cameras at test time.

III. Method

A. Problem Formulation

Our objective is to recover the 3D head surface, denoted as $\mathcal{S}$, from a small set of $N \geq 1$ input images $\mathbf{I}_v$, where $v = 1, \ldots, N$. Each input image is accompanied by its respective head mask $\mathbf{M}_v$ and camera parameters $\mathbf{T}_v$. We represent the surface $\mathcal{S}$ as the zero-level set of a signed distance function $f^{\mathrm{sdf}} : \mathbf{x} \to s$, such that $\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^3 \mid f^{\mathrm{sdf}}(\mathbf{x}) = 0\}$. After estimating f^{sdf} from the visual cues $\mathbf{I}_v$, $\mathbf{M}_v$ and $\mathbf{T}_v$, the surface $\mathcal{S}$ can be obtained as a post-processing step using Marching Cubes [62]

To tackle the inherent underconstrained nature of recovering 3D geometry from image data, the incorporation of regularizations is crucial in resolving existing ambiguities. The problem becomes increasingly challenging as the number of available images decreases. We specifically address this challenge within a range of 1 to 32 views, where the one-shot regime poses the greatest difficulty due to the absence of multi-view cues to disentangle geometric and color information. To tackle this challenge, we propose a novel architecture that capitalizes on both geometric and appearance priors. By harnessing these priors, our approach achieves precise 3D reconstructions even in scenarios lacking multi-view consistency.

We recover the 3D geometry from the input images through analysis-by-synthesis with differentiable surface rendering as in [23], [32]. Our proposed architecture addresses the challenge of limited multi-view information by leveraging two key inductive biases. Firstly, we decompose the signed distance function f^{sdf} into a reference SDF and a deformation field [63]. This parameterization serves as an implicit bias, ensuring that the composed SDF remains close to the reference. Secondly, we create a statistical prior that models the variations of shape and appearance of 3D head surfaces. This model is used during inference to achieve a reliable initialization and faster convergence. To enhance the robustness of the analysis-by-synthesis process during inference, we optimize the parameters of this statistical model, drawing inspiration from [32], to achieve a reliable initialization. These inductive biases significantly enhance the performance of SIRA++, surpassing that of [32].

B. Surface Appearance Statistical Model (SA-SM)

In order to learn a statistical model capturing head shapes and appearances (Fig. 2), we curate a dataset comprising scenes that contain raw head scans along with corresponding multiview posed images. For each scene, denoted by an index $i = 1 \ldots M$, we extract a collection of surface points $\mathbf{x} \in \mathcal{P}_{\mathrm{s}}^{(i)}$ along with their respective normal vectors $\mathbf{n}$. We project each surface point onto the images where it is visible, obtaining a set denoted as $\mathcal{C}_{\mathbf{x}}^{(i)} = (\mathbf{c}, \mathbf{v})$, which consists of pairs comprising the associated RGB color $\mathbf{c}$ and the corresponding viewing direction $\mathbf{v}$.

SA-SM Architecture: Our architecture consists of two primary neural field decoders: an SDF decoder denoted as $f^{\mathrm{sdf}}_{\boldsymbol{\theta}_{\mathrm{sdf}},\mathbf{z}_{\mathrm{sdf}}}$, and a rendering function decoder represented as $r_{\boldsymbol{\theta}_{\mathrm{r}},\mathbf{z}_{\mathrm{r}}}$. Here, $\mathbf{z}_{\mathrm{sa}} = \{\mathbf{z}_{\mathrm{sdf}}, \mathbf{z}_{\mathrm{r}}\}$ refers to the latent vectors encompassing the shape and appearance spaces of the SA-SM. Additionally, $\boldsymbol{\theta}_{\mathrm{sa}} = \{\boldsymbol{\theta}_{\mathrm{sdf}}, \boldsymbol{\theta}_{\mathrm{r}}\}$ denotes the parameters associated with their respective decoders.

The SDF decoder consists of two sub-functions: a deformation field and a reference SDF. Our experimental results (Section VI-B, Figure 6 and Table II) demonstrate that this separation introduces an implicit bias, effectively limiting significant deviations from the reference SDF. As a result, it enhances the stability of few-shot 3D reconstructions. The deformation field, is mathematically defined as:

$$f^{\mathrm{def}}_{\boldsymbol{\theta}_{\mathrm{def}},\mathbf{z}_{\mathrm{sdf}}} : \mathbb{R}^3 \to \mathbb{R}^{3+N_\gamma} \ , \ \mathbf{x} \mapsto (\boldsymbol{\delta}, \boldsymbol{\gamma}), \tag{1}$$

is parameterised by internal parameters $\boldsymbol{\theta}_{\mathrm{def}}$ and the latent vector $\mathbf{z}_{\mathrm{sdf}}$. It maps input coordinates, $\mathbf{x}$, to a deformation 3-vector, $\boldsymbol{\delta}$. Additionally, it generates an auxiliary feature vector $\boldsymbol{\gamma}$ of dimension N_γ, which encodes higher-level geometric information utilized by the differentiable renderer [23]. Note, however, that the rendering network does not include global lightning effects such us secondary lightning and self-shadows, as it is only conditioned on position, viewing direction, normals, and the appearance latent.

The predicted deformation is utilized to map an input coordinate $\mathbf{x}$ to a coordinate $\mathbf{x}_{\mathrm{ref}}$ within a reference space. In this reference space, we evaluate a reference signed distance function (SDF) $f^{\mathrm{ref}}_{\boldsymbol{\theta}_{\mathrm{ref}}}$, which is parameterized by internal parameters $\boldsymbol{\theta}_{\mathrm{ref}}$. This mapping process is expressed as follows:

$$\begin{aligned} \mathbf{x}_{\mathrm{ref}} \quad &= \quad \mathbf{x} + \boldsymbol{\delta}, && (2) \\ f^{\mathrm{ref}}_{\boldsymbol{\theta}_{\mathrm{ref}}} \quad &: \quad \mathbb{R}^3 \to \mathbb{R} \ , \ \mathbf{x}_{\mathrm{ref}} \mapsto s. && (3) \end{aligned}$$

Combining the components described above, we obtain the composed SDF decoder:

$$f^{\mathrm{sdf}}_{\boldsymbol{\theta}_{\mathrm{sdf}},\mathbf{z}_{\mathrm{sdf}}} : \mathbf{x} \mapsto f^{\mathrm{ref}}_{\boldsymbol{\theta}_{\mathrm{ref}}}(\mathbf{x}^{\mathrm{ref}}), \tag{4}$$

where the decoder internal parameters are $\boldsymbol{\theta}_{\mathrm{sdf}} = (\boldsymbol{\theta}_{\mathrm{def}}, \boldsymbol{\theta}_{\mathrm{ref}})$. The second main component of our architecture is the rendering function:

$$r_{\boldsymbol{\theta}_{\mathrm{r}},\mathbf{z}_{\mathrm{r}}} : (\mathbf{x}_{\mathrm{ref}}, \mathbf{n}, \mathbf{v}, \boldsymbol{\gamma}) \mapsto \mathbf{c} \ , \tag{5}$$

which is parameterised by internal parameters $\boldsymbol{\theta}_{\mathrm{r}}$ and a latent vector $\mathbf{z}_{\mathrm{r}}$. This function assigns an RGB color $\mathbf{c}$, to each combination of 3D coordinates in the reference space $\mathbf{x}_{\mathrm{ref}}$, unit normal vector $\mathbf{n}$, and unit viewing direction vector $\mathbf{v}$ in the real space.

SA-SM Training: To train our SA-SM we adopt an auto-decoder framework in which each scene is associated with a set of latents $\mathbf{z}_{\mathrm{sa}}^{(i)} = \{\mathbf{z}_{\mathrm{sdf}}^{(i)}, \mathbf{z}_{\mathrm{r}}^{(i)}\}$. These latents are optimized alongside the statistical model parameters $\boldsymbol{\theta}_{\mathrm{sa}}$. Upon

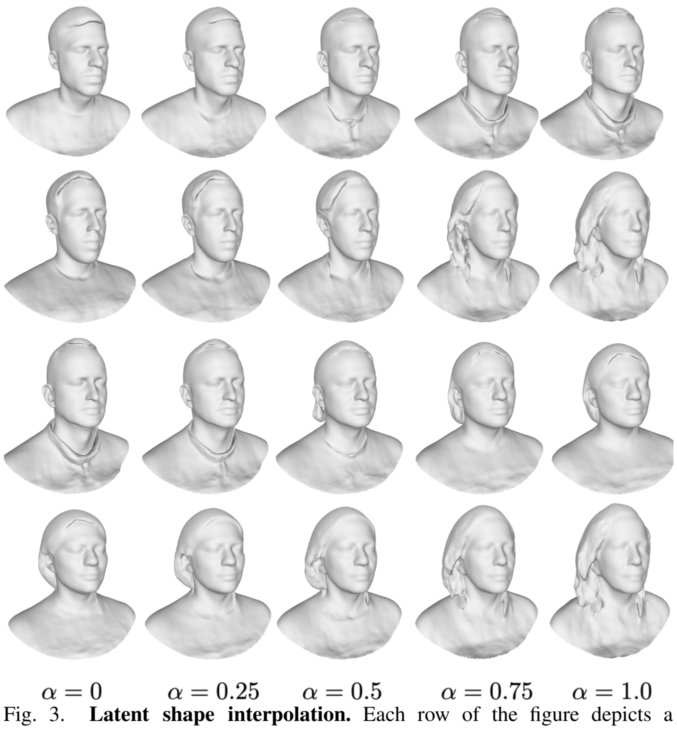

Fig. 3. **Latent shape interpolation.** Each row of the figure depicts a latent interpolation between different subjects, controlled by a weight α. This interpolation process showcases the smooth and gradual transformation of shapes, reflecting the continuous variation in human head representations along the latent space.

completion of training, we obtain the optimized parameters $\boldsymbol{\theta}_{\mathrm{sa},0} = \{\boldsymbol{\theta}_{\mathrm{sdf},0}, \boldsymbol{\theta}_{\mathrm{r},0}\}$ which establish that any combination of latents $(\mathbf{z}_{\mathrm{sdf}}, \mathbf{z}_{\mathrm{r}})$ within the latent space corresponds to a well-behaved SDF, $f^{\mathrm{sdf}}_{\boldsymbol{\theta}_{\mathrm{sdf},0},\mathbf{z}_{\mathrm{sdf}}}$ and appearance $f^{\mathrm{rend}}_{\boldsymbol{\theta}_{\mathrm{r},0},\mathbf{z}_{\mathrm{r}}}$ of a human head. To simplify notation and reduce complexity, we subsequently omit the dependency on the decoder's internal parameters.

In order to capture the space of head shapes, we sample a set of points on the surface of each training scan denoted as $\mathcal{P}_{\mathrm{s}}^{(i)}$. Subsequently, we compute the surface error loss given by:

$$\mathcal{L}_{\mathrm{Surf}}^{(i)} = \sum_{\mathbf{x}_j \in \mathcal{P}_{\mathrm{s}}^{(i)}} |f^{\mathrm{sdf}}_{\mathbf{z}_{\mathrm{sdf}}^{(i)}}(\mathbf{x}_j)|. \tag{6}$$

In addition, we uniformly sample another set of points from the scene volume, $\mathcal{P}_{\mathrm{v}}^{(i)}$, and compute the Eikonal loss [64]:

$$\mathcal{L}_{\mathrm{Eik}}^{(i)} = \sum_{\mathbf{x}_k \in \mathcal{P}_{\mathrm{v}}^{(i)}} (\|\nabla_{\mathbf{x}} f^{\mathrm{sdf}}_{\mathbf{z}_{\mathrm{sdf}}^{(i)}}(\mathbf{x}_k)\| - 1)^2. \tag{7}$$

To encourage small-magnitude and zero-mean deformations, we incorporate a regularization term that prevents solutions where the deformations unnecessarily compensate for offset or scaled reference SDFs. The regularization term is defined as:

$$\mathcal{L}_{\mathrm{Def}}^{(i)} = \frac{1}{|\mathcal{P}_{\mathrm{s}}^{(i)}|}\left(\sum_{\mathbf{x}_j \in \mathcal{P}_{\mathrm{s}}^{(i)}} \|\boldsymbol{\delta}_j^{(i)}\|_2 + \left\| \sum_{\mathbf{x}_j \in \mathcal{P}_{\mathrm{s}}^{(i)}} \boldsymbol{\delta}_j^{(i)} \right\|_2 \right), \tag{8}$$

where $\boldsymbol{\delta}_j^{(i)}$ represents the deformation vector applied to the 3D point $\mathbf{x}_j$ within the scene indexed by i.

Similar to the approach in [63], we employ a landmark consistency loss to ensure consistency among the 3D face landmarks. For each scene i, we automatically annotate a set of 3D face landmarks denoted as $\{\mathbf{x}_l^{(i)}\}$ where $l = 1 \ldots L$. We then define the following loss that measures their deformed coordinate mismatch between pairs of scenes:

$$\mathcal{L}_{\mathrm{Lm}}^{(i)} = \sum_{j \neq i} \sum_{l}^{L} \|\mathbf{x}_{\mathrm{ref,l}}^{(i)} - \mathbf{x}_{\mathrm{ref,l}}^{(j)}\|^2 \ , \tag{9}$$

where $\mathbf{x}_{\mathrm{ref,l}}^{(i)}$ represents the position of landmark l of scene i in the reference space.

The SA-SM model learns a distribution of head appearances from the posed images associated with each training scene. To evaluate the rendering function (Eq. 5), we calculate the coordinates in the reference space $\mathbf{x}_{\mathrm{ref}}$ corresponding to the surface point (Eq. 3), along with the high-level descriptor γ (Eq. 1). Additionally, we extract the surface normals, $\mathbf{n}$, by normalizing the gradient of the SDF [23]. Using these components, we define the color loss as follows:

$$\mathcal{L}_{\mathrm{Col}}^{(i)} = \sum_{\mathbf{x} \in \mathcal{P}_{\mathrm{s}}^{(i)}} \sum_{(\mathbf{c},\mathbf{v}) \in \mathcal{C}_{\mathbf{x}}^{(i)}} \|r_{\mathbf{z}_{\mathrm{r}}^{(i)}}(\mathbf{x}_{\mathrm{ref}}, \mathbf{n}, \mathbf{v}, \gamma) - \mathbf{c}\| \ . \tag{10}$$

Finally, the $\mathcal{L}_{\mathrm{Emb}}^{(i)}$ term enforces a zero-mean multivariate-Gaussian distribution with spherical covariance of σ^2 over the spaces of shape and appearance latent vectors: $\mathcal{L}_{\mathrm{Emb}}^{(i)} = \frac{1}{\sigma^2}\left(\|\mathbf{z}_{\mathrm{sdf}}^{(i)}\|_2 + \|\mathbf{z}_{\mathrm{r}}^{(i)}\|_2\right)$. Combining all the loss terms, we minimize the following objective:

$$\begin{aligned} \arg\min_{\{\mathbf{z}_{\mathrm{sa}}^{(i)}\},\boldsymbol{\theta}_{\mathrm{sa}}} \sum_i \mathcal{L}_{\mathrm{Surf}}^{(i)} + \lambda_1 \mathcal{L}_{\mathrm{Eik}}^{(i)} + \lambda_2 \mathcal{L}_{\mathrm{Def}}^{(i)} + \lambda_3 \mathcal{L}_{\mathrm{Lm}}^{(i)} + \\ \lambda_4 \mathcal{L}_{\mathrm{Col}}^{(i)} + \lambda_5 \mathcal{L}_{\mathrm{Emb}}^{(i)} \end{aligned} \tag{11}$$

where λ_{1-5} are scalar hyperparameters.

Expressivity of the SA-SM Prior: In order to assess the representation power of the shape and appearance prior we have learned, we conduct simple experiments by fitting our model to unseen subjects and interpolating their latent codes. In Fig. 3, we illustrate this process specifically for the shape prior. The ability to represent diverse fitting subjects indicates a rich and expressive manifold that captures diverse human head variations. Notably, the latent interpolation between different subjects (even across different genders) results in a remarkably smooth transition of feasible human heads. This observation highlights the structure and continuity of the learned latent codes. In Fig. 4, we present a similar experiment where we simultaneously interpolate in both the shape space (vertical direction of the figure, governed by the weight α) and the appearance space (horizontal direction, controlled by the weight β). Once again, we observe that our learned shape-and-appearance prior gracefully transitions between the two subjects represented at the top-left and bottom-right corners of the figure. Interestingly, we observe that while the appearance latent effectively captures most of the color variance, such as the t-shirt, certain details like the mustache and beard are better represented by the geometry latent. This phenomenon arises because SIRA++ jointly models a distribution of 3D shapes and appearances, enabling the geometry latent to explain color variations that are statistically correlated with the underlying geometry.

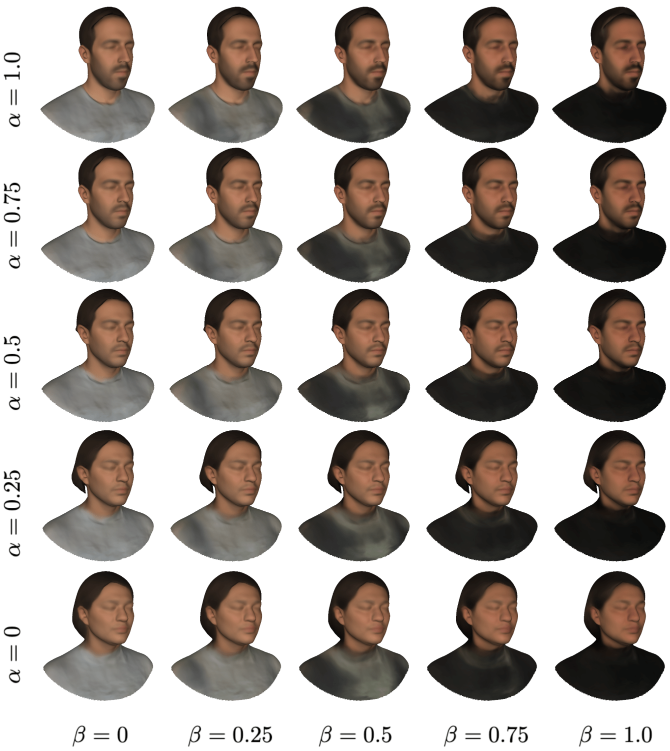

Fig. 4. **Latent shape and appearance interpolation.** This figure shows the joint shape and appearance latent interpolation between two subjects (top-left and bottom-right). Shape interpolation of the latent $\mathbf{z}_{sdf}$ is controlled by means of the weight α. Appearance interpolation of $\mathbf{z}_r$ is controlled by β.

C. Geometry Reconstruction

With our pre-trained statistical model at hand, we can tackle the task of obtaining a 3D reconstruction and an appearance from portrait images $\mathbf{I}_v$ with associated camera parameters $\mathbf{T}_v$ and foreground mask $\mathbf{M}_v$.

To obtain precise 3D reconstructions of new scenes, our approach involves rendering the geometry described by f^{sdf} using the differentiable rendering function r from Eq. 5, while minimizing a photoconsistency error. Here's a step-by-step breakdown of our process:

First, for a given pixel coordinate p in the input image $\mathbf{I}_v$, we sample a ray $\mathbf{r} = \mathbf{t} + k\mathbf{v}|k \geq 0$, where $\mathbf{t}$ represents the position of the associated camera $\mathbf{T}_v$, and $\mathbf{v}$ is the viewing direction. We then find the intersection coordinates of this ray with the composed SDF (Eq. 4).

Next, we make this intersection point differentiable with respect to $\mathbf{z}_{\text{sdf}}$ and $\boldsymbol{\theta}_{\text{sdf}}$ through implicit differentiation [22], [23]. The resulting differentiable intersection coordinates $\mathbf{x}_{\text{s}}$ enable us to obtain their associated 3D displacement $\boldsymbol{\delta}$ and feature vector $\boldsymbol{\gamma}$ (Eq. 1), as well as their corresponding coordinates in the reference space $\mathbf{x}_{\text{ref}}$ (Eq. 3), along with the normal vector $\mathbf{n} = \nabla_{\mathbf{x}} f^{\text{sdf}}$.

Finally, we compute the color associated with the ray as $\mathbf{c} = r(\mathbf{x}_{\text{ref}}, \mathbf{n}, \mathbf{v}, \boldsymbol{\gamma})$ using the differentiable rendering function r.

In order to optimize $\mathbf{z}_{\text{sa}}$ and $\boldsymbol{\theta}_{\text{sa}}$, we minimize the following loss [23]:

$$\mathcal{L} = \mathcal{L}_{\text{RGB}} + \lambda_6 \mathcal{L}_{\text{Mask}} + \lambda_7 \mathcal{L}_{\text{Eik}}, \tag{12}$$

where λ_6 and λ_7 are hyperparameters.

TABLE I
TESTING TIME COMPARISON ON H3DS DATASET WITH 1 AND 3 INPUT VIEWS. SS STANDS FOR SELECTIVE SAMPLING.

	Test time 1 view ↓	Test time 3 views ↓	Time reduction ↑
MVFNet [1]	-	$< 10s$	-
DFNRMVS [3]	-	$< 10s$	-
DECA [28]	$< 10s$	-	-
MICA [27]	$< 10s$	-	-
FaceScape [29]	$< 10s$	-	-
FaceVerse [30]	720s	-	-
HRN [31]	$< 5s$	–	-
PIFU [19]	$< 10s$	$< 10s$	-
JIFF [60]	$< 10s$	$< 10s$	-
IDR [23]	-	$\sim 1000s$	-
NeuS2 [25]	-	$\sim 300s$	-
H3D-Net [32]	-	$1050s$	-
SIRA [33]	-	$3663s$	-
SIRA++ (w/o SS, w/o Cache)	-	$379s$	0%
SIRA++ (w SS, w/o Cache)	-	$275s$	27%
SIRA++ (w/o SS, w Cache)	-	$232s$	39%
SIRA++ (w SS, w Cache)	-	**191s**	**50%**

We will now elaborate on each component of this loss. Let $\mathcal{P}$ be a mini-batch of pixels from the image $\mathbf{I}_v$. We define $\mathcal{P}_{\text{RGB}}$ as the subset of pixels whose associated ray intersects the surface defined by f^{sdf} and have a nonzero foreground mask value, while $\mathcal{P}_{\text{Mask}} = \mathcal{P} \setminus \mathcal{P}_{\text{RGB}}$. Let's delve into the specifics:

The first component, $\mathcal{L}_{\text{RGB}}$, addresses photometric error, which is computed as follows:

$$\mathcal{L}_{\text{RGB}} = |\mathcal{P}|^{-1} \sum_{p \in \mathcal{P}_{\text{RGB}}} |\mathbf{I}_v(p) - \mathbf{c}_v(p)|. \tag{13}$$

The second component, $\mathcal{L}_{\text{Mask}}$, accounts for silhouette errors. It is defined as:

$$\mathcal{L}_{\text{Mask}} = \frac{1}{\lambda_8 |\mathcal{P}|} \sum_{p \in \mathcal{P}_{\text{Mask}}} \text{CE}(\mathbf{M}_v(p), s_{\lambda_8}(p)) \;, \tag{14}$$

where $s(p) = \text{sigmoid}(-\lambda_8 \min_{t \geq 0} f^{\text{sdf}}(\mathbf{r}_t))$ is the estimated silhouette. We use the binary cross-entropy CE to measure the difference between the foreground mask $\mathbf{M}_v(p)$ and the estimated silhouette $s(p)$ for each pixel p in $\mathcal{P}_{\text{Mask}}$.

Lastly, $\mathcal{L}_{\text{Eik}}$ encourages f^{sdf} to approximate a signed distance function.

Instead of optimizing all the parameters $\boldsymbol{\theta}_{\text{sdf}}, \boldsymbol{\theta}_{\text{r}}, \mathbf{z}_{\text{sdf}}, \mathbf{z}_{\text{r}}$ simultaneously, we propose a two-step schedule. First, we initialize the geometry and rendering functions with the parameters obtained from the pretraining described in the last section, denoted as $\boldsymbol{\theta}_{\text{sdf},0}, \boldsymbol{\theta}_{\text{r},0}$. The initial shape and appearance latents, $\mathbf{z}_{\text{sdf}}$ and $\mathbf{z}_{\text{r}}$, are sampled from a multivariate normal distribution with zero mean and a small variance, ensuring that they start near the mean of the latent spaces. In the first optimization phase, we exclusively optimize the shape and appearance latents. This yields an initial approximation within the previously learned shape and appearance latent spaces. Subsequently, in the second phase, we unfreeze the parameters of the deformation and rendering networks, denoted as $\boldsymbol{\theta}_{\text{def}}, \boldsymbol{\theta}_{\text{r}}$ (Eqs. 1 and 5), while keeping the parameters of the reference Signed Distance Function (SDF), $\boldsymbol{\theta}_{\text{ref}}$, frozen.

This two-step scheduling plays a pivotal role in achieving accurate results, particularly in the one-shot regime. By unfreezing the deformation and rendering networks, we can attain highly detailed solutions that lie outside of the pre-learned latent spaces. However, a crucial aspect of our approach is expressing the shape as a deformed reference Signed Distance Function (SDF), which acts as a regularization mechanism, ensuring proper training convergence.

The fine-tuned shape parameters resulting from this two-step process are denoted as $\boldsymbol{\theta}_{\mathrm{def,ft}}$ and $\mathbf{z}_{\mathrm{sdf,ft}}$.

IV. Accelerating Reconstruction Geometry

Reconstruction methods relying on surface rendering often require considerable time to find the intersection between cast rays and the reconstructed surface, leading to run times ranging from 15 minutes to several hours per scene [23], [32], [33]. To address this challenge, we have introduced several enhancements in SIRA++, including code optimization, adoption of a more parallelizable ray tracing algorithm [65], implementation of a new scheduler to discard non-intersecting rays, and dynamic caching of the Signed Distance Function (SDF) during training. These improvements, as depicted in Table I, yield a substantial reduction in overall computation time, making our proposed method significantly faster and more efficient in comparison to existing techniques. In the following sections, we provide a detailed explanation of each of these enhancements.

Code optimitzation. SIRA++ has been developed within the PyTorch framework, with accelerated training on a single GPU. To optimize performance, we employ mixed precision during the ray tracing step, group query points into a single batch to minimize GPU calls, and minimize the use of memory copy instructions (e.g., reshape, concatenations, clones). Additionally, we strategically fuse operations, such as expressing sinus and cosines as a single call in the positional encoder. These measures collectively contribute to enhancing the efficiency and speed of the SIRA++ framework during the training process.

Selective sampling. In traditional 3D reconstruction methods based on surface or volumetric rendering, rays are often uniformly sampled on the image, leading to many rays inefficiently sampling empty space far from the surface. To enhance the efficiency of our reconstruction process, we implement a progressive strategy where we gradually stop sampling rays from the background during the optimization. This approach significantly reduces computational time while preserving the same level of reconstruction accuracy. By intelligently focusing on rays closer to the surface of interest, we achieve a faster and more efficient reconstruction process without compromising the quality of the final results.

Dynamic SDF caching. To efficiently find a ray-surface intersection, we search for the first sign flip of the Signed Distance Function (SDF) among a set of N_c equally-spaced points along the ray. However, to reduce the number of MLP queries, we implement a caching mechanism using a voxel grid. When we sample a point belonging to a voxel with a cached SDF value s, we compare it with a threshold ϵ. If $s \geq \epsilon$, we consider the point to be far from the surface and safely reuse the cached value. Otherwise, we re-evaluate the SDF network and update the cache accordingly.

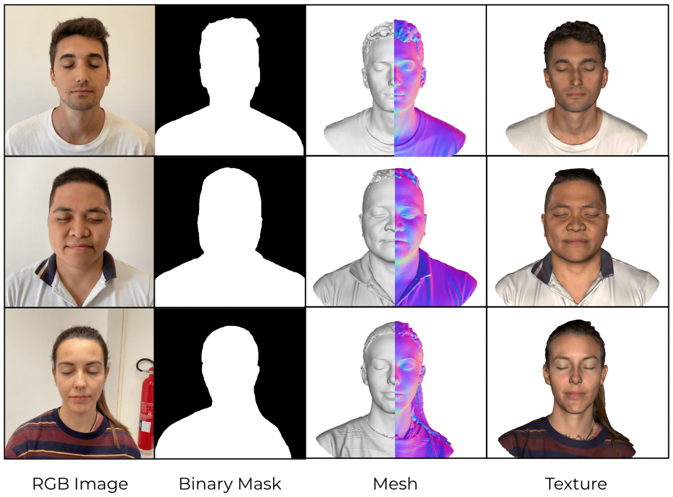

Fig. 5. **H3DS Dataset.** Three samples from the dataset, each scene composed of 60-100 RGB images, foreground masks, camera parameters, and high-resolution textured 3D meshes capturing the full head, including hair and upper body clothing .

To prevent deadlocks, where the cached SDF value remains unjustifiably greater than ϵ, we introduce a random forcing mechanism. With a probability p, we deliberately trigger the evaluation and re-caching of the SDF network. After locating the interval where the first SDF sign flip occurs, we further refine the intersection estimation by repeating this process with a set of N_f sub-sampled points.

V. Implementation details

Equations 1, 3 and 4 are implemented using Multi-Layer Perceptrons (MLPs) with one skip connection from the network's input to the input of a hidden layer, following the approach used in [66]. We incorporate a SoftPlus activation function in all the hidden layers of the network architecture. Additionally, positional encoding (PE) [67] is applied to some of the inputs of the networks, further enhancing their representation capabilities.

The SA-SM pretraining optimization is iterated for 100 epochs using the Adam optimizer [68] with standard parameters. We set the learning rate to 10^{-4} and apply a learning rate step decay of 0.5 every 15 epochs. To balance the different components of the loss function, we set the loss hyperparameters as follows: $\lambda_1 = 0.1$, $\lambda_2 = \lambda_3 = \lambda_5 = 10^{-3}$, and $\lambda_4 = 1$. Additionally, we automatically annotate six 3D facial landmarks for each scene, which are then utilized for the landmark consistency loss.

The weights of the reference SDF network (Eq. 3) are initialized using the geometric initialization method described in [69]. As for the deformation and rendering networks, their weights are initialized as multivariate Gaussians with zero mean and variance 10^{-4}. Furthermore, the latent vectors $\mathbf{z}_{\mathrm{sdf}}$ and $\mathbf{z}_{\mathrm{r}}$ are initialized as zero vectors.

We adopt a progressive masking strategy for the positional encoding (PE) of the input to the reference SDF [70]–[72] to minimize artifacts on the reference shape and improve

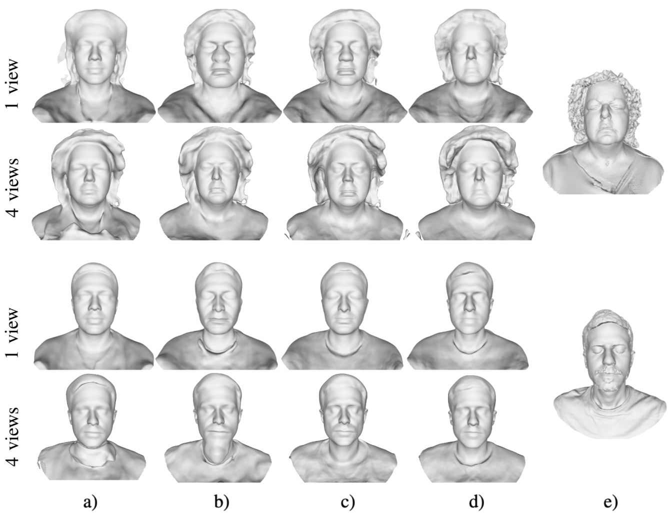

Fig. 6. **Ablation study:** 3D reconstruction for two subjects from the H3DS dataset. (a) H3D-Net [32]; (b) is H3D-Net + progressive masking; (c) H3D-Net + Deformation Field + Reference SDF; (d) SIRA++ (Ours); (e) Ground Truth. See text for details.

training stability. This technique involves initially masking the higher frequency bands, effectively acting as a dynamic low-pass filter. By allowing the model to focus on reaching robust coarse solutions first before incorporating high-frequency content, we achieve better convergence behavior. To implement this approach, we introduce a parameter $\zeta \in [0, L]$ that is proportional to the progress of the training, where L represents the total number of frequencies used in the PE. The Fourier embedding of frequency k is then multiplied by a scalar $w_k(\zeta)$, defined as:

$$w_k(\zeta) = \begin{cases} 0 & \zeta \leq k \\ (1 - \cos{(\zeta - k)\pi})/2 & 0 \leq \zeta - k \leq 1 \ . \\ 1 & \zeta - k \geq 1 \end{cases} \tag{15}$$

We start with masking all frequencies in the PE and gradually unmask them between epochs 5 and 10 by linearly increasing the parameter ζ from 0 to L.

Reconstructing geometry from images:

At test time, the 3D reconstruction of a scene is performed over 2000 epochs using the Adam optimizer with an initial learning rate of 10^{-4}. We apply a learning rate step decay of 0.5 at epochs 1000 and 1500 to adaptively adjust the learning rate during the optimization process.

For the mask loss $\mathcal{L}_{\text{Mask}}$, we schedule the parameter λ_8 following the approach in [23]. We use a two-step scheduling strategy, where the weights of the deformation and rendering networks are unfrozen at epoch 100, enabling more focused fine-tuning during the latter stage of the optimization.

During optimization, we drop 12% of the background rays every 250 epochs during the initial 1000 epochs. For caching, we implement a voxel grid with a size of 64^3. We set ϵ to 0.1 and we randomly sample points along the ray with a probability of $p = 0.2$ to ensure exploration and avoid deadlocks. The ray sampling is implemented using $N_c = 75$ and $N_f = 25$ steps.

TABLE II
Ablation study in the one-shot (1 view) and 4 views setup. The face and full-head mean distances are the averages over the 23 subjects in the H3DS dataset in mm. The configurations a, b, c, and d are the same as those described in Figure 6

	(a)	(b)	(c)	(d)
Face mean distance (1 view)	1.97	2.17	1.86	**1.46**
Full-head mean distance (1 view)	15.40	14.00	**13.40**	14.16
Face mean distance (4 views)	1.49	1.54	1.38	**1.29**
Full-head mean distance (4 views)	11.38	10.61	**9.00**	10.51

TABLE III
Camera noise comparison. Evaluation on H3DS dataset with 3 input views.

	Noise	Noise Error (σ)	face (mm) ↓	head (mm) ↓
H3D-Net [32] ours	✗	-	1.44 **1.30**	11.6 **10.59**
H3D-Net [32] ours	✓	0.002	1.51 **1.30**	12.35 **10.62**
H3D-Net [32] ours	✓	0.01	1.60 **1.37**	12.36 **10.74**
H3D-Net [32] ours	✓	0.02	1.71 **1.48**	12.96 **10.83**
H3D-Net [32] ours	✓	0.04	1.87 **1.71**	13.36 **11.84**

VI. Experiments

A. Datasets

Prior training. To train the geometry prior, we utilize an internal dataset comprising 3D head scans from 10,000 individuals. This dataset is intentionally designed to be well-balanced in terms of gender representation and diverse in terms of age and ethnicity. Before training, the raw data undergoes an automatic processing step to remove internal mesh faces and non-human parts, such as background walls. To ensure consistency and alignment across the dataset, all the scenes are registered by using a non-rigid Iterative Closest Point (ICP) approach to align each head scan with a template 3D model.

H3DS. There exist several 3D face datasets [29], [30], [73]–[76] that can be used for various tasks, however, large-scale datasets containing high-quality 3D data of full head scans, including hair and shoulders, paired with casual posed RGB images are currently scarce. To address this limitation, we have significantly expanded the H3Ds dataset [32] by tripling the number of scenes, resulting in a total of 60 subjects. Each subject in the dataset is represented by approximately 100 RGB photos with a resolution of 512x512 pixels, capturing a full 360-degree view around the head. These RGB images are accompanied by foreground masks and camera parameters. Moreover, to enable accurate and reliable ground truth evaluation, the dataset includes high-quality 3D textured scans for each subject. These 3D scans are composed of approximately 150,000 vertices and 400,000 faces, complemented by a texture map with a resolution of 2048x2048 pixels (see fig. 5). This dataset can be used either for optimization-based methods or for validation purposes.

TABLE IV

3D reconstruction comparison. Average surface error (in mm) computed over all subjects in 3DFAW and H3DS datasets. We place ”-” for not applicable configurations and ”$*$” for experiments that raised out of memory error.

	3DFAW		H3DS 2.0														
	1 view	3 view	1 views		3 views		4 views		6 views		8 views		16 views		32 views	
	face	face	face	head	face	head	face	head	face	head	face	head	face	head	face	head
MVFNet [1]	-	1.56	-	-	1.73	-	-	-	-	-	-	-	-	-	-	-
DFNRMVS [3]	-	1.69	-	-	1.83	-	-	-	-	-	-	-	-	-	-	-
DECA [28]	1.71	-	1.99	-	-	-	-	-	-	-	-	-	-	-	-	-
MICA [27]	1.83	-	2.08	-	-	-	-	-	-	-	-	-	-	-	-	-
FaceVerse [30]	1.88	-	2.57	-	-	-	-	-	-	-	-	-	-	-	-	-
FaceScape [29]	1.61	-	1.78	-	-	-	-	-	-	-	-	-	-	-	-	-
HRN [40]	1.60	-	1.73	-	-	-	-	-	-	-	-	-	-	-	-	-
PIFU [19]	2.19	1.99	1.98	12.6	1.70	11.3	1.85	11.8	2.03	10.9	$*$	$*$	$*$	$*$	$*$	$*$
JIFF [60]	1.48	1.47	1.85	11.5	1.80	11.7	1.79	11.2	1.79	10.9	$*$	$*$	$*$	$*$	$*$	$*$
IDR [23]	-	3.92	-	-	3.51	28.8	3.33	14.3	3.12	16.4	2.97	12.6	2.88	12.4	2.04	10.6
NeuS2 [25]	-	-	-	-	-	-	3.96	8.29	2.18	6.60	2.70	5.79	2.11	4.55	1.85	4.14
H3D-Net [32]	1.70	1.37	-	-	1.44	11.6	1.41	8.32	1.21	5.75	1.35	7.88	1.17	6.91	1.04	5.67
SIRA++ (Ours)	**1.35**	**1.32**	**1.57**	**10.79**	**1.18**	**8.63**	**1.23**	**5.31**	**1.04**	**4.81**	**1.18**	**4.63**	**1.07**	**4.26**	**1.02**	**4.12**

The data acquisition process for each scene in the dataset involves several steps. Initially, the camera of an iPad Pro is calibrated using an attached structured light sensor, specifically the Occipital Structure Sensor Pro. This calibration process allows us to obtain paired RGB images and camera parameters, along with a low-resolution mesh scan. Simultaneously, a high-end Artec Eva scanner is utilized to capture high-quality 3D scans. Subsequently, we align the low and high-resolution meshes, along with the paired cameras, by employing six manually annotated 3D landmarks and utilizing the iterative closest point (ICP) refinement technique. Furthermore, for each image within the dataset, we manually annotate a foreground mask, providing additional information for foreground-background separation.

3DFAW. [76] This dataset provides videos recorded 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.

B. Ablation Study

For the ablation study, we focus on a subset of 23 scenes, randomly chosen, from the H3DS dataset.

Architecture analysis. SIRA++ architecture (see sec. III) introduces two significant differences compared to H3D-Net [32]: it represents the geometry as a deformed reference SDF and incorporates pretraining for rendering human head appearances. We have found both of these strategies to be crucial for achieving high-quality results, especially when the number of views decreases. To thoroughly investigate their impact, we conduct the ablation study on the H3DS dataset for both the one-shot and three-shot scenarios. The qualitative results are presented in Figure 6.

We utilize H3D-Net [32] as our baseline. As evident from Fig. 6, this architecture underfits the scene when only the latent vector is optimized (a), and in (b) it becomes unstable when its unique geometry decoder is fine-tuned following the progressive masking of Eq. 15. To address these issues, we introduce a significant enhancement by splitting the geometry into a deformation field and a reference SDF (c). This modification leads to more plausible and stable solutions. Furthermore, SIRA++ (d) leverages joint modeling of the distribution of 3D shapes and appearances with the SA-SM, enabling better disambiguation of geometric and visual information. Consequently, the 3D models generated by SIRA++ highly resemble the input images in (e). The quantitative results of this study, over the 23 scenes are reported in Table II.

Robustness to camera noise. In real-world scenarios, a certain level of inaccuracy in camera poses is inevitable, leading to multi-view inconsistencies. To evaluate the robustness of SIRA++ in such situations, we conducted an ablation study by introducing varying levels of noise into the camera poses. We applied different levels of Gaussian noise, where the standard deviation σ controlled the amount of noise injected into the rotation matrix and position of the camera. Remarkably, our method demonstrates strong resilience against the injected noise, and consistently betten than [32]. The results of this study are presented in Table III.

C. Quantitative results

We conducted a comprehensive comparison of our method with several 3DMM-based reconstruction works, including MVFNet [3], DFNRMVS [1], DECA [28], MICA [27], FaceScape [29], FaceVerse [30] and HRN [40]. Additionally, we compared our approach to the model-free methods IDR [23], NeuS2 [25], PIFU [19], JIFF [60] and H3D-Net [32]. For the quantitative evaluation, we used the unidirectional Chamfer distance, measuring the surface error from the ground truth to the predictions. The results of this comparison are summarized in Table IV.

Both model-free methods (H3D-Net and SIRA++) outperform the 3DMM-based methods for all the evaluated view configurations. Notably, the enhancement due to the prior in SIRA++ becomes more significant as the number of views decreases. However, the prior does not hinder the model from also becoming more accurate when more views are available, which is a limitation in 3DMM-based approaches.

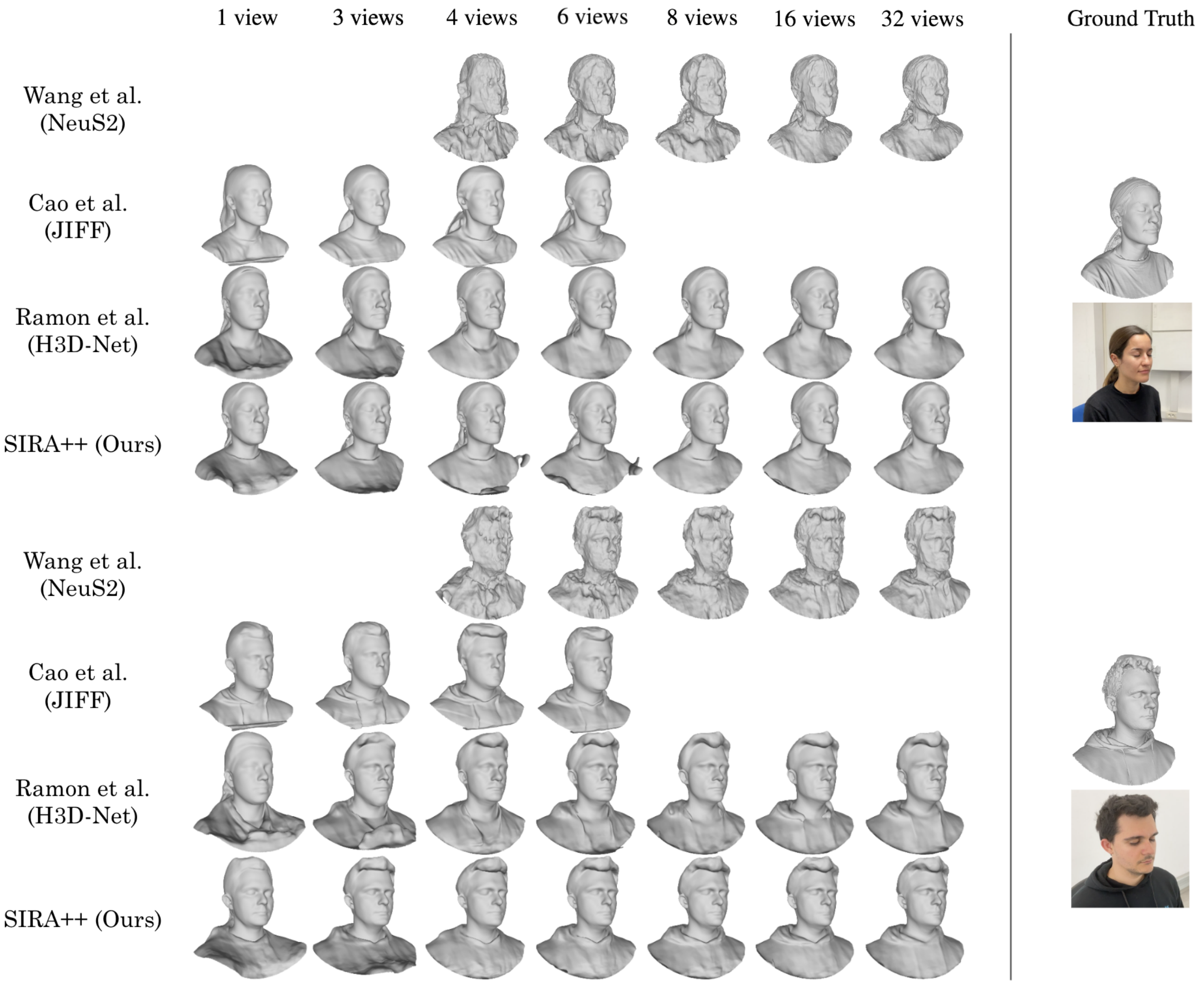

Fig. 7. **Qualitative results on two subjects of the H3DS dataset**, for NeuS2 [25], JIFF [60], H3D-Net [32] and SIRA++, with an increasing number of input views.

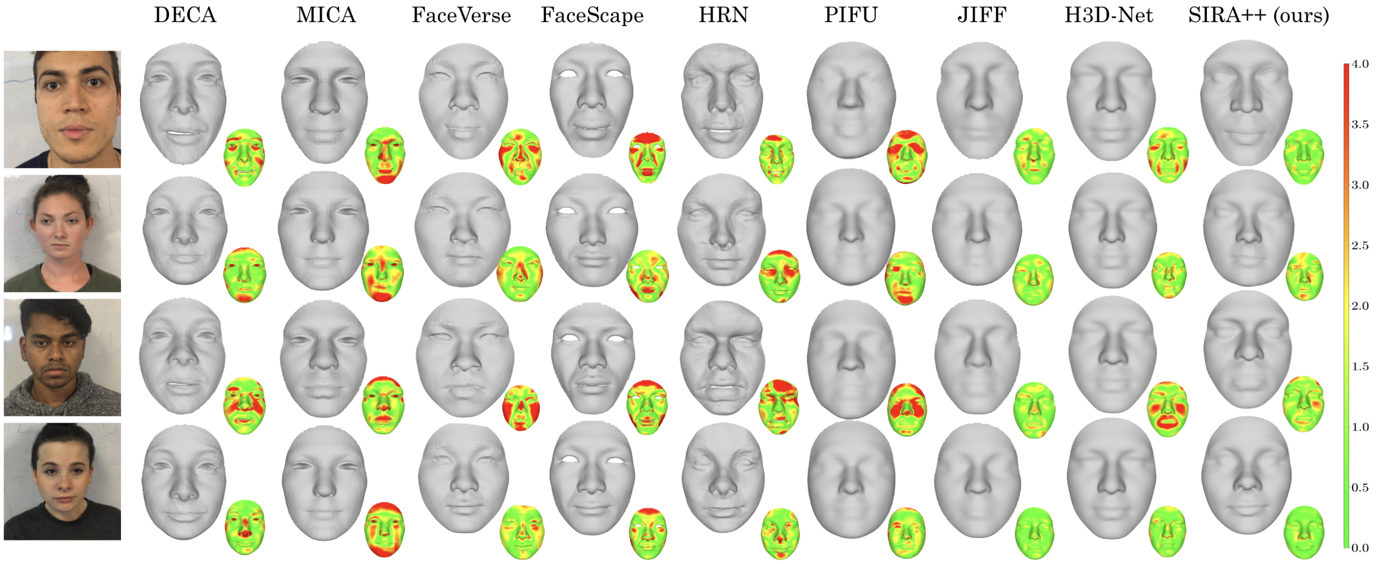

Fig. 8. **Qualitative results on the 3DFAW dataset for a single input image.** Each 3D reconstructed face is accompanied by a heatmap, where reddish areas indicate larger errors in mm.

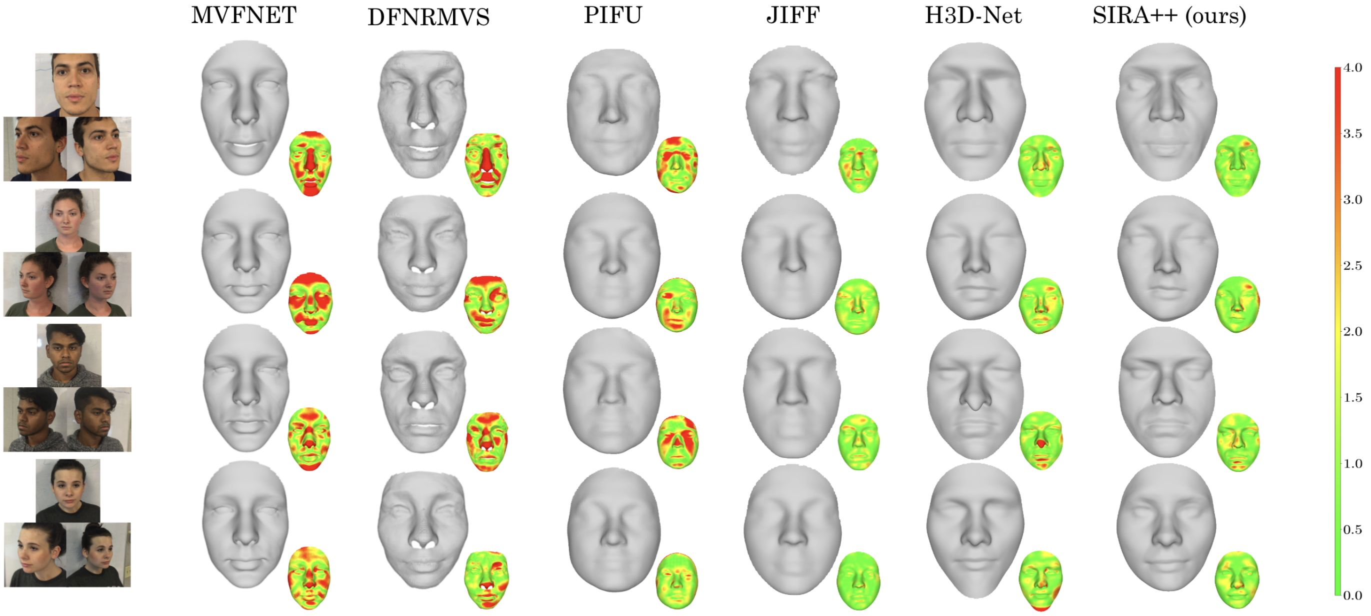

Fig. 9. **Qualitative results on the 3DFAW dataset for three input images.** Each 3D reconstructed face is accompanied by a heatmap, where reddish areas indicate larger errors in mm.

TABLE V
User study: We collect 375 responses from 25 participants to measure visual fidelity of the reconstructions.

	NeuS2	JIFF	H3D-Net	SIRA++ (ours)
VF↑	1.0	2.21	3.29	**3.5**

In the one-shot regime, SIRA++ stands out over 3DMM-based approaches and H3D-Net, all of which yield similar results. Both single feed-forward PIFU and JIFF methods generate convincing and robust results, although reconstructions are smooth and they are not able to capture high-frequency details, especially in the face region. SIRA++ consistently outperforms IDR, NeuS2 and H3D-Net for all configurations, demonstrating its improved ability to generalize under data scarcity. When more views become available and the task is more constrained, SIRA++ and H3D-Net converge towards comparable performance, as the prior knowledge becomes less critical for obtaining plausible results. These findings are further supported by the qualitative results obtained.

D. Qualitative results

Figure 7 illustrates the qualitative results of our approach, SIRA++, and state-of-the-art approaches NeuS2 [25], JIFF [60] and H3D-Net [32], for two subjects from the H3DS dataset under an increasing number of input images. Notably, our method, SIRA++, demonstrates superior performance in surface reconstruction, yielding surfaces with reduced errors and a more realistic appearance, particularly within the facial region. Even with a smaller number of input views, our approach excels at obtaining accurate and visually appealing results. To quantitatively asess these results, we perform a user study, with 25 human participants, to evaluate visual fidelity (see Table V). We present a photo of a subject and renders of reconstruction for each method. We ask the participants to rank them based on visual fidelity (how well the reconstruction captures the details of the person shown on the image). We assign a numeric value between 1 and 4 for each response based on the order. Results show that SIRA++ reconstructions are consistently better perceived as digital representation of full-heads.

Fig. 8 showcases results for single input images obtained from the 3DFAW dataset [76]. Our method, SIRA++, is compared against the one-shot methods DECA, MICA, FaceVerse, FaceScape, HRN, PIFU, JIFF, and H3D-Net. Next to each 3D reconstruction, we display a heatmap representing the reconstruction error. Note that SIRA++ outperforms the 3DMM-based methods, DECA, MICA, FaceVerse, FaceScape and HRN, and model-free methods, PIFU, JIFF and H3D-Net significantly. Specifically, it excels in critical regions like the nose and mouth, which are pivotal in defining the unique anatomical features of each individual.

Similarly, Fig. 9 provides an analysis for the case of three input images. Here, we compare SIRA++ against MVF-Net [1], DFNRMVS [3], PIFU, JIFF and H3D-Net. Again, the methods based on coordinate-based neural representation, H3D-Net, and especially our SIRA++, outperform those relying on 3D Morphable Models (MVF-Net and DFRMVS) and single feed-forward models (PIFU and JIFF).

This is further highlighted in our last experiment, summarized in Fig. 10, where we specifically focus on the most challenging scenario of utilizing just one single and in-the-wild input image, randomly taken from the celebA-HQ dataset [77]. In this case SIRA++ also consistently outperforms single feed-forward methods, PIFU and JIFF as well as the 3DMM methods, DECA, MICA FaceVerse, FaceScape and HRN. DECA and MICA struggle to capture fine anatomical details, often leading to similar-looking faces across different scenes. On the other hand, FaceVerse and FaceScape produce biased outputs toward Asian characteristics, as the training data is composed of Asian subjects. Additionally, these approaches

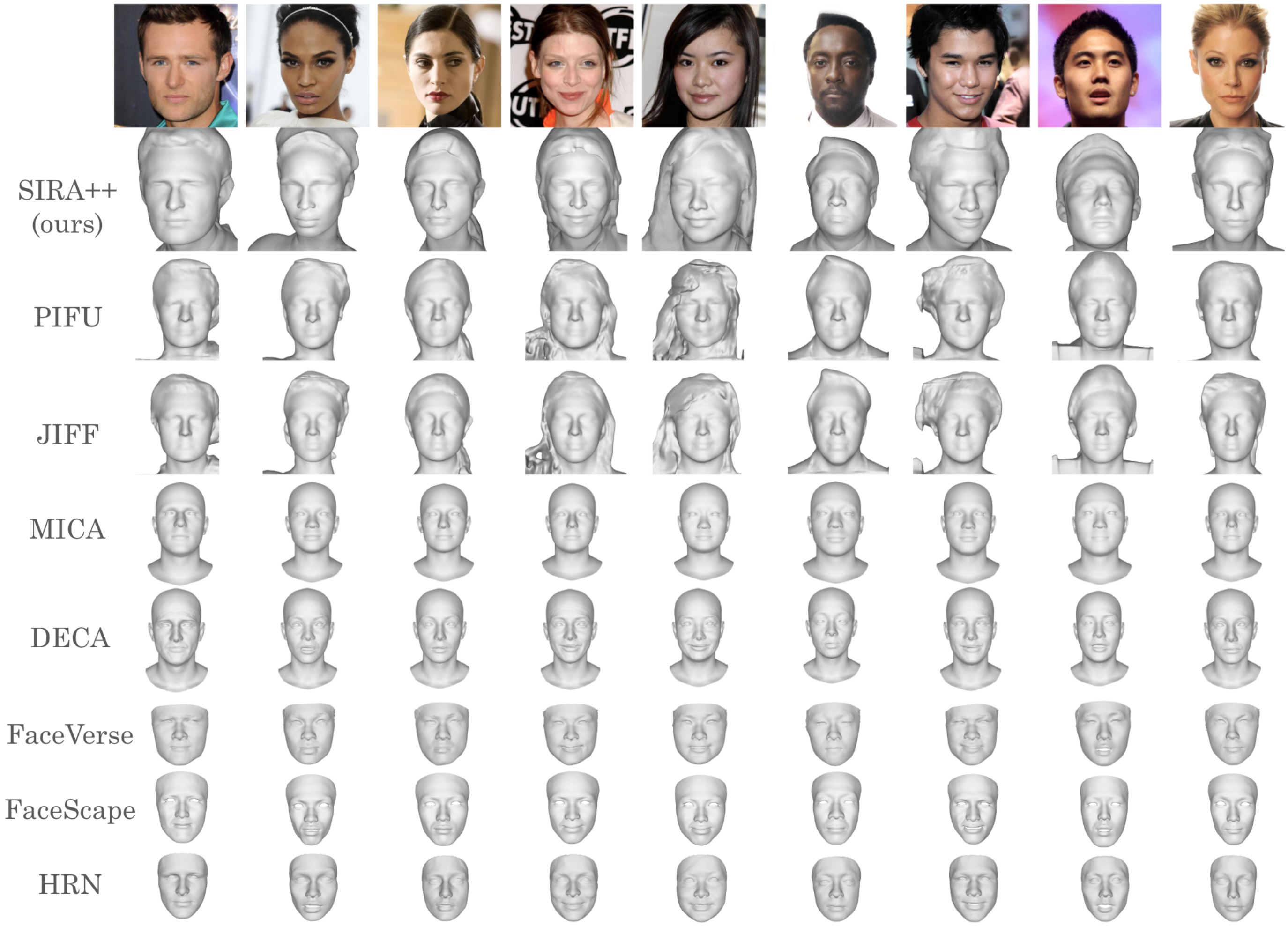

Fig. 10. **Qualitative results on the CelebA-HQ dataset for a single input image.**

are limited to reconstructing only the facial region and fail to recover the hair and shoulders, which significantly impact perception. In contrast, our method excels at capturing not only the facial features but also includes the hair, upper body clothing, and other high-frequency anatomical details, especially in the cheeks and mouth regions. This ability results in 3D shapes that retain the identity of the person, showcasing the unique characteristics of the individual.

VII. Limitations and Future Work

In our experiments, we show that SIRA++ demonstrates significant advancements in quick personalization of pretrained geometry and appearance priors from a few headshot images. However, we believe that there are still several limitations and opportunities for future work. Firstly, our rendering networks are based on a DeepSDF decoder, and future work could focus on combining them with Gaussian Splats for fast, high-quality rendering. Although our approach achieves state-of-the-art reconstruction accuracy within 200 seconds, this may still be prohibitive for real-time applications or scenarios requiring fast processing of multiple scenes. Future work could focus on optimizing the efficiency of the algorithm. Another limitation is the sensitivity of our method to the quality of the input images. Our approach is sensitive to the resolution, angles, occlusion, and information present in the input images. Variations in these factors can significantly affect the reconstruction quality.

VIII. Conclusions

In this paper, we have introduced SIRA++, a method for high-fidelity full 3D head reconstruction in few-shot and in-the-wild scenarios. To address the inherent ambiguity of the problem, we proposed a novel statistical model based on neural fields, which encoded shape and appearance into low-dimensional latent spaces. The thorough evaluation demonstrated that our approach achieved state-of-the-art results in full head geometry reconstruction. Moreover, through a detailed ablation study, we showcased the robustness of our method to camera pose misalignment. We also presented a set of improvements that led to an impressive 80% reduction in computation time compared to previous approaches, H3D-Net [32] and SIRA [33].

References

[1] F. Wu, L. Bao, Y. Chen, Y. Ling, Y. Song, S. Li, K. N. Ngan, and W. Liu, “Mvf-net: Multi-view 3d face morphable model regression,” in *CVPR*, 2019.

[2] E. Ramon, J. Escur, and X. Giro-i Nieto, “Multi-view 3d face reconstruction in the wild using siamese networks,” in *ICCV Workshops*, 2019.

[3] Z. Bai, Z. Cui, J. A. Rahim, X. Liu, and P. Tan, “Deep facial non-rigid multi-view stereo,” in *CVPR*, 2020.

[4] P. Dou and I. A. Kakadiaris, "Multi-view 3d face reconstruction with deep recurrent neural networks," *Image and Vision Computing*, vol. 80, pp. 80–91, 2018.

[5] A. Tewari, M. Zollhofer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt, "Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction," in *ICCV*, 2017.

[6] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni, "Regressing robust and discriminative 3d morphable models with a very deep neural network," in *CVPR*, 2017.

[7] E. Richardson, M. Sela, and R. Kimmel, "3d face reconstruction by learning from synthetic data," in *3DV*, 2016.

[8] E. Richardson, M. Sela, R. Or-El, and R. Kimmel, "Learning detailed face reconstruction from a single image," in *CVPR*, 2017.

[9] M. Sela, E. Richardson, and R. Kimmel, "Unrestricted facial geometry reconstruction using image-to-image translation," in *ICCV*, 2017.

[10] A. T. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. G. Medioni, "Extreme 3d face reconstruction: Seeing through occlusions." in *CVPR*, 2018.

[11] F. Moreno-Noguer and P. Fua, "Stochastic exploration of ambiguities for nonrigid shape recovery," vol. 35, no. 2, pp. 463–475, 2013.

[12] J. Lin, Y. Yuan, T. Shao, and K. Zhou, "Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks," in *CVPR*, 2020.

[13] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large pose 3d face reconstruction from a single image via direct volumetric cnn regression," in *ICCV*, 2017.

[14] Z. Chen and H. Zhang, "Learning implicit fields for generative shape modeling," in *CVPR*, 2019.

[15] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *CVPR*, 2019.

[16] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European conference on computer vision*. Springer, 2020, pp. 405–421.

[17] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 165–174.

[18] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *CVPR*, 2021.

[19] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *ICCV*, 2019.

[20] V. Sitzmann, M. Zollhöfer, and G. Wetzstein, "Scene representation networks: Continuous 3d-structure-aware neural scene representations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[21] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "pixelnerf: Neural radiance fields from one or few images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4578–4587.

[22] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, "Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision," in *CVPR*, 2020.

[23] L. Yariv, Y. Kasten, D. Moran, M. Galun, M. Atzmon, B. Ronen, and Y. Lipman, "Multiview neural surface reconstruction by disentangling geometry and appearance," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2492–2502, 2020.

[24] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, "Nerf in the wild: Neural radiance fields for unconstrained photo collections," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7210–7219.

[25] Y. Wang, Q. Han, M. Habermann, K. Daniilidis, C. Theobalt, and L. Liu, "Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[26] A. Morales, G. Piella, and F. M. Sukno, "Survey on 3D face reconstruction from uncalibrated images," *Comput. Sci. Rev.*, vol. 40, no. 100400, p. 100400, May 2021.

[27] W. Zielonka, T. Bolkart, and J. Thies, "Towards metrical reconstruction of human faces," in *European Conference on Computer Vision (ECCV)*. Springer International Publishing, Oct. 2022.

[28] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3d face model from in-the-wild images," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–13, 2021.

[29] H. Zhu, H. Yang, L. Guo, Y. Zhang, Y. Wang, M. Huang, M. Wu, Q. Shen, R. Yang, and X. Cao, "Facescape: 3d facial dataset and benchmark for single-view 3d face reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.

[30] L. Wang, Z. Chen, T. Yu, C. Ma, L. Li, and Y. 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.

[31] C. Häne, S. Tulsiani, and J. Malik, "Hierarchical surface prediction for 3d object reconstruction," in *3DV*, 2017.

[32] E. Ramon, G. Triginer, J. Escur, A. Pumarola, J. Garcia, X. Giro-i Nieto, and F. Moreno-Noguer, "H3d-net: Few-shot high-fidelity 3d head reconstruction," in *ICCV*, 2021, pp. 5620–5629.

[33] P. Caselles, E. Ramon, J. Garcia, X. Giro-i Nieto, F. Moreno-Noguer, and G. Triginer, "Sira: Relightable avatars from a single image," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2023, pp. 775–784.

[34] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *AVSS*, 2009, pp. 296–301.

[35] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, vol. 36, no. 6, pp. 194:1–194:17, 2017. [Online]. Available: https://doi.org/10.1145/3130800.3130813

[36] L. Bao, X. Lin, Y. Chen, H. Zhang, S. Wang, X. Zhe, D. Kang, H. Huang, X. Jiang, J. Wang, D. Yu, and Z. Zhang, "High-fidelity 3d digital human head creation from rgb-d selfies," *ACM Transactions on Graphics*, 2021.

[37] B. Gecer, S. Ploumpis, I. Kotsia, and S. P. Zafeiriou, "Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[38] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou, "Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[39] A. Dib, C. Thebault, J. Ahn, P.-H. Gosselin, C. Theobalt, and L. 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*, 2021, pp. 12 819–12 829.

[40] B. Lei, J. Ren, M. Feng, M. Cui, and X. Xie, "A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images," 2023.

[41] C. Li, A. Morel-Forster, T. Vetter, B. Egger, and A. Kortylewski, "Robust model-based face reconstruction through weakly-supervised outlier segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 372–381.

[42] R. Danecek, M. J. Black, and T. Bolkart, "EMOCA: Emotion driven monocular face capture and animation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 20 311–20 322.

[43] Y. Xie, T. Takikawa, S. Saito, O. Litany, S. Yan, N. Khan, F. Tombari, J. Tompkin, V. Sitzmann, and S. Sridhar, "Neural fields in visual computing and beyond," *arXiv preprint arXiv:2111.11426*, 2021.

[44] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, "Nerfies: Deformable neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5865–5874.

[45] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 318–10 327.

[46] K. Zhang, F. Luan, Q. Wang, K. Bala, and N. Snavely, "Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5453–5462.

[47] K. Park, U. Sinha, P. Hedman, J. T. Barron, S. Bouaziz, D. B. Goldman, R. Martin-Brualla, and S. M. Seitz, "Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields," *arXiv preprint arXiv:2106.13228*, 2021.

[48] Y. Zheng, V. F. Abrevaya, X. Chen, M. C. Bühler, M. J. Black, and O. Hilliges, "Im avatar: Implicit morphable head avatars from videos," *arXiv preprint arXiv:2112.07471*, 2021.

[49] G. Gafni, J. Thies, M. Zollhofer, and M. Nießner, "Dynamic neural radiance fields for monocular 4d facial avatar reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8649–8658.

[50] C. Z. Lin, K. Nagano, J. Kautz, E. R. Chan, U. Iqbal, L. Guibas, G. Wetzstein, and S. Khamis, "Single-shot implicit morphable faces with

consistent texture parameterization," *arXiv preprint arXiv:2305.03043*, 2023.

[51] M. Zheng, H. Yang, D. Huang, and L. Chen, "Imface: A nonlinear 3d morphable face model with implicit neural representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 343–20 352.

[52] P.-W. Grassal, M. Prinzler, T. Leistner, C. Rother, M. Nießner, and J. Thies, "Neural head avatars from monocular rgb videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 653–18 664.

[53] Y. Zheng, V. F. Abrevaya, M. C. Bühler, X. Chen, M. J. Black, and O. Hilliges, "I M Avatar: Implicit morphable head avatars from videos," in *Computer Vision and Pattern Recognition (CVPR)*, 2022.

[54] Y. Zheng, W. Yifan, G. Wetzstein, M. J. Black, and O. Hilliges, "Pointavatar: Deformable point-based head avatars from videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[55] S. Saito, T. Simon, J. Saragih, and H. Joo, "Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 84–93.

[56] T. He, J. Collomosse, H. Jin, and S. Soatto, "Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9276–9287, 2020.

[57] T. Alldieck, M. Zanfir, and C. Sminchisescu, "Photorealistic monocular 3d reconstruction of humans wearing clothing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1506–1515.

[58] R. Shao, H. Zhang, H. Zhang, M. Chen, Y.-P. Cao, T. Yu, and Y. Liu, "Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 872–15 882.

[59] E. Corona, M. Zanfir, T. Alldieck, E. G. Bazavan, A. Zanfir, and C. Sminchisescu, "Structured 3d features for reconstructing controllable avatars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 954–16 964.

[60] Y. Cao, G. Chen, K. Han, W. Yang, and K.-Y. K. Wong, "Jiff: Jointly-aligned implicit face function for high quality single view clothed human reconstruction," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[61] M. Mihajlovic, A. Bansal, M. Zollhoefer, S. Tang, and S. Saito, "KeypointNeRF: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints," in *European conference on computer vision*, 2022.

[62] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," in *Seminal graphics: pioneering efforts that shaped the field*, 1998, pp. 347–353.

[63] T. Yenamandra, A. Tewari, F. Bernard, H.-P. Seidel, M. Elgharib, D. Cremers, and C. Theobalt, "i3dmm: Deep implicit 3d morphable model of human heads," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 803–12 813.

[64] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, "Implicit geometric regularization for learning shapes," *arXiv preprint arXiv:2002.10099*, 2020.

[65] A. Canela, P. Caselles, I. Malik, E. Ramon, J. García, J. Sánchez-Riera, G. Triginer, and F. Moreno-Noguer, "Instantavatar: Efficient 3d head reconstruction via surface rendering," 2023.

[66] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[67] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *arXiv preprint arXiv:2006.10739*, 2020.

[68] D. P. Kingma and J. Ba, *arXiv preprint arXiv:1412.6980*, 2014.

[69] M. Atzmon and Y. Lipman, "Sal: Sign agnostic learning of shapes from raw data," in *CVPR*, 2020.

[70] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, "Barf: Bundle-adjusting neural radiance fields," *arXiv preprint arXiv:2104.06405*, 2021.

[71] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, "Nerfies: Deformable neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 5865–5874.

[72] A. Hertz, O. Perel, R. Giryes, O. Sorkine-Hornung, and D. Cohen-Or, "Sape: Spatially-adaptive progressive encoding for neural optimization," in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[73] H. Dai, N. Pears, W. Smith, and C. Duncan, "Statistical modeling of craniofacial shape and texture," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 547–571, Feb. 2020.

[74] S. Sanyal, T. Bolkart, H. Feng, and M. Black, "Learning to regress 3D face shape and expression from an image without 3D supervision," in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 7763–7772.

[75] C. Chen, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "FaceWarehouse: a 3D facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2014.

[76] R. K. Pillai, L. A. Jeni, H. Yang, Z. Zhang, L. Yin, and J. F. Cohn, "The 2nd 3d face alignment in the wild challenge (3dfaw-video): Dense reconstruction from video." in *ICCV Workshops*, 2019.

[77] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

Pol Caselles Rico is a third-year industrial PhD candidate at Universitat Politècnica de Catalunya (UPC). He is currently working as an applied scientist at Crisalix Labs in Barcelona. His primary research interests are in Computer Vision and Machine Learning, with a specific emphasis on 3D shape reconstruction from 2D images, using implicit functions and model-free approaches.

Eduard Ramon received his PhD in computer science from the Universitat Politècnica de Catalunya in 2022. During his PhD and previously, he worked at Crisalix as a computer vision scientist and (co)authored several publications on topics related to statistical models, and 3D reconstruction of human bodies and faces. Currently, he is working as an applied scientist at Amazon.

Jaime Garcia Giraldez is the CEO and founder of Crisalix, Switzerland. He received his PhD in Biomedical Engineering from the Medical University of Bern. He has co-authored more than 10 publications in refereed journals and conferences. His research interests are in computer graphics, 3D and Augmented Reality, Computer Assisted Surgery (CAS), computer vision and AI.

Gil Triginer received his PhD in physics from the University of Oxford in 2019. Subsequently, he joined Crisalix Labs as a research scientist, with a focus on 3D reconstruction of human faces and bodies using deep learning techniques. He has coauthored publications in leading computer vision conferences including ICCV, WACV, and ACCV workshops.

Francesc Moreno-Noguer is a Principal Applied Scientist at Amazon Science, specializing in Computer Vision and Machine Learning. His research focuses on human shape and motion estimation, 3D reconstruction of both rigid and nonrigid objects, and camera calibration. He received the Polytechnic University of Catalonia's Doctoral Dissertation Extraordinary Award, multiple best paper awards (e.g. ECCV 2018 Honorable mention, ICCV 2017 workshop in Fashion, Intl. Conf. on Machine Vision applications 2016), outstanding reviewer awards at ECCV 2012 and CVPR 2014, and Google and Amazon Faculty Research Awards in 2017 and 2019, respectively. He has (co)authored over 200 publications in refereed journals and conferences (including 13 IEEE Transactions on PAMI, 5 Intl. Journal of Computer Vision, 30 CVPR, 13 ECCV and 10 ICCV).