

SIRA: Relightable Avatars from a Single Image

Pol Caselles^{1,2,3}

Eduard Ramon^{1,2,*}

Jaime Garcia¹

Xavier Giro-i-Nieto^{2,3*}

Francesc Moreno-Noguer³

Gil Triginer¹

¹Crisalix SA

²Universitat Politècnica de Catalunya

³Institut de Robòtica i Informàtica Industrial, CSIC-UPC

Abstract

Recovering the geometry of a human head from a single image, while factorizing the materials and illumination, is a severely ill-posed problem that requires prior information to be solved. Methods based on 3D Morphable Models (3DMM), and their combination with differentiable renderers, have shown promising results. However, the expressiveness of 3DMMs is limited, and they typically yield over-smoothed and identity-agnostic 3D shapes limited to the face region. Highly accurate full head reconstructions have recently been obtained with neural fields that parameterize the geometry using multilayer perceptrons. The versatility of these representations has also proved effective for disentangling geometry, materials and lighting. However, these methods require several tens of input images. In this paper, we introduce SIRA, a method which, from a single image, reconstructs human head avatars with high fidelity geometry and factorized lights and surface materials. Our key ingredients are two data-driven statistical models based on neural fields that resolve the ambiguities of single-view 3D surface reconstruction and appearance factorization. Experiments show that SIRA obtains state of the art results in 3D head reconstruction while at the same time it successfully disentangles the global illumination, and the diffuse and specular albedos. Furthermore, our reconstructions are amenable to physically-based appearance editing and head model relighting.

1. Introduction

Digitalizing humans into 3D relightable avatars is key for a wide range of applications in e.g. augmented/virtual reality, 3D content production or the movie industry. In order to realistically render the captured models under new lighting conditions, it is not enough to just recover the 3D geometry, but also the rest of intrinsic properties of the scene need to be estimated, namely surface materials and scene illumina-

nation. It is specially challenging when the input data is acquired in non-controlled conditions, and when the number of input images is small [45, 29, 2, 7]. The single view setup is the most difficult scenario, and it poses a highly under-constrained problem that cannot be solved without a priori knowledge [39, 42, 31, 32, 33, 40].

In order to recover the intrinsic properties of a scene from a single image, also known as inverse rendering, state of the art methods [39, 6, 35, 15] introduce prior knowledge by means of 3D Morphable Models (3DMM) [12, 25, 17, 41, 3, 27, 18], which are combined with deep neural networks. These models estimate the 3D geometry, spatially varying surface properties like the diffuse and specular albedos, as well as global illumination properties in the form of spherical harmonics or spherical gaussians. The regressed components are supervised using explicit ground truth data or in a self-supervised fashion in the image domain using differentiable renderers. However, the expressiveness of 3DMMs is limited, as they are biased towards low frequencies for both the geometry and albedo, and they are typically restricted to the facial area.

Recently, scene representation methods based on neural fields [46] parameterized using multilayer perceptrons (MLP), have shown impressive results for the tasks of novel view synthesis and 3D reconstruction. One of the main advantages of neural fields in front of other representations like meshes or voxel grids is their great compromise between representational power and memory requirements. This, in combination with differentiable surface rendering [38], enables highly accurate 3D reconstructions and also to disentangle the scene into their intrinsic properties [50], leading to excellent models that can be rendered under novel lighting conditions. Their main limitation is that to supervise the learning process they require multiple views.

To overcome these limitations we introduce SIRA (Fig. 1), a neural field representation of the head, including hair and shoulders, that can be learned *from one single human portrait*. We approach the inverse rendering problem in two steps. First, we recover the 3D geometry by optimizing a

*This work was done prior to joining Amazon.

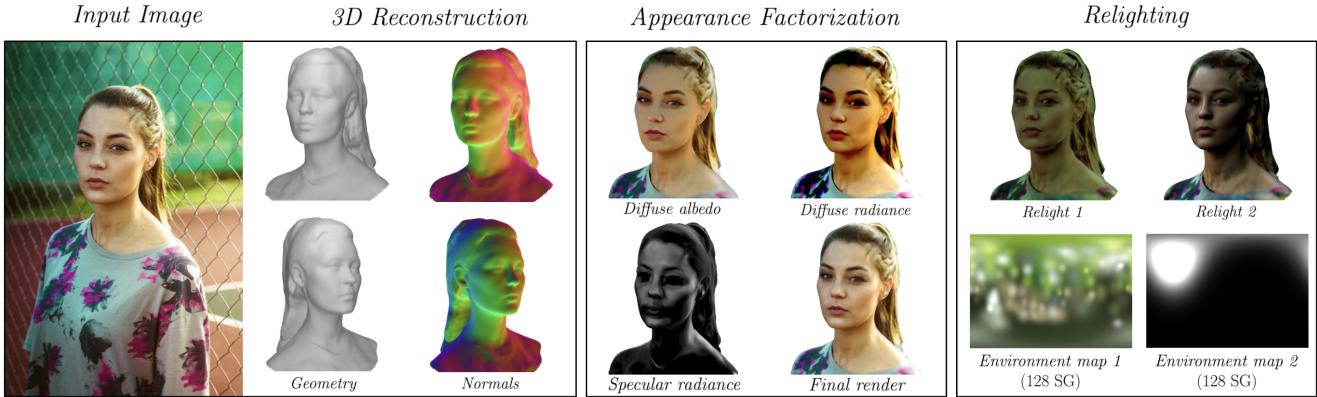


Figure 1: We introduce SIRA, a method which, from a single image, reconstructs human head avatars with high fidelity geometry and factorized lights and surface materials. This information can then be used for relighting purposes.

signed distance function in combination with surface rendering, in the style of [48]. To resolve the inherent ambiguities of this problem, we introduce a novel shape and appearance statistical model (SA-SM) that is used throughout the reconstruction process. Then, in a second step where the 3D geometry is already estimated, we factorize the appearance into diffuse and specular albedos, and global illumination. For this, we leverage an appearance factorization statistical model (AF-SM), which is trained in a self-supervised fashion via physically-based rendering.

A thorough evaluation demonstrates that the reconstructed geometry compares favorably to recent state-of-the-art methods, while in addition we also provide material and lighting parameters. The outcomes of SIRA enable, for the first time, the digitalization of human heads into relightable avatars from a single image. In summary, our contributions are threefold:

- A novel shape and appearance statistical model (SA-SM) that allows to recover the 3D geometry of portrait scenes from a single image.
- A novel appearance factorization statistical model (AF-SM) that splits surface radiance fields into diffuse and specular albedos, and global illumination.
- A methodology for training the aforementioned statistical models and using them to create relightable avatars of 3D heads, including hair and shoulders.

2. Related work

Inverse rendering from a single portrait image. Recovering the 3D head geometry from a single image is an ill-posed problem. Typically, 3DMMs in combination with deep neural networks are used to learn a mapping from an input image to a 3D geometry [42, 31, 32, 40]. In addition, some methods decompose the scene into diffuse albedo, and global illumination in the form of spherical harmonics [39, 6, 35, 15]. Recently, [35] introduced a morphable

model that disentangles the albedo into diffuse and specular components, enabling their factorization at test time [6]. In [15], the diffuse and specular reflectance is modelled using image-to-image translation networks in the texture space, which are learnt using ground truth data. While these methods enable the inverse rendering of portrait scenes, they present a number of limitations [8]. First, 3DMMs are usually restricted to the facial region [12, 3]. Second, these methods only model low frequency geometry and further post-processing is usually required to obtain fine details [32]. And third, the topology of 3DMMs is fixed, which limits the shapes these methods can represent [27]. These are important limitations for reconstructing realistic avatars.

Neural fields for 3D reconstruction and scene factorization. Recently, neural fields have been proposed as scene representations [46], obtaining impressive results on novel view synthesis [20, 19, 23, 28, 34] and 3D reconstruction [21, 30, 48, 50], with application to modelling full head avatars [23, 24, 52, 10]. By combining surface priors [22] and surface rendering [48], neural fields enable very accurate 3D reconstructions of the full head, including hair and shoulders [30, 52]. However, these methods require multi-view information and they do not recover the diffuse and specular albedos, and the illumination. Recently, novel priors that jointly model surface and appearance have been proposed [49], but they have not been applied to 3D reconstruction from images. Neural fields have also been proposed for de-rendering scenes from several multi-view posed images [51, 4, 36, 5, 14, 37, 50]. These methods combine neural fields with physically-based renderers, disentangling the intrinsic properties of the scene through direct supervision on images. They also introduce priors on the materials to disambiguate the appearance factorization process [51]. However, it still remains a challenge to solve the inverse rendering problem from a single image using neural fields.

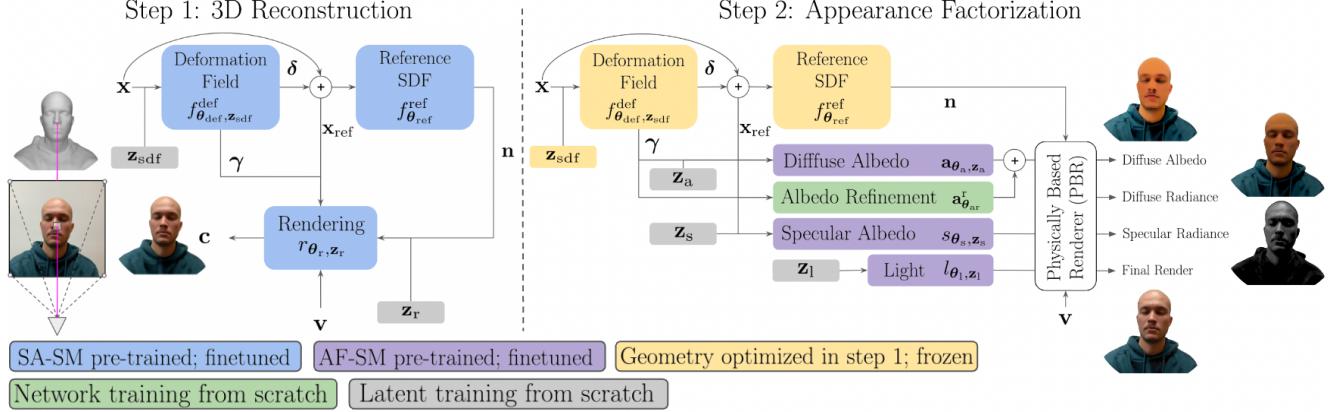


Figure 2: Inverse rendering using SIRA.

3. Method

We follow an analysis-by-synthesis approach to retrieve all the components of a relightable avatar from a single posed image \mathbf{I} with associated foreground mask \mathbf{M} , and camera parameters \mathbf{C} . The 3D geometry is represented as a signed distance function (SDF) $f^{\text{sdf}} : \mathbf{x} \rightarrow s$, such that the surface \mathcal{S} is implicitly defined as $\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^3 | f^{\text{sdf}}(\mathbf{x}) = 0\}$. To capture the complex appearance of human faces, we factor the surface radiance into a global illumination and spatially-varying diffuse and specular albedos, which we also implement as neural fields. All our neural fields are implemented as multilayer perceptrons. Find implementation details in the supplementary material.

Instead of using a single architecture to simultaneously optimize the geometry and appearance from scratch, we split the problem into a 3D reconstruction and an appearance factorization part. This two-step approach, similar to [51], allows us to adapt the training scheme to each of the two problems independently, and to introduce appropriate inductive biases. These are key to resolve the ambiguities that exist in single-view 3D reconstruction and appearance factorization (Fig. 2).

In a first step, we recover the 3D geometry of the scene from a single image. [48, 30] have shown that f^{sdf} can be reconstructed from a collection of multiview posed images by using a differentiable rendering function $r : (\mathbf{x}, \mathbf{n}, \mathbf{v}) \rightarrow \mathbf{c}$ that models the radiance \mathbf{c} emitted from a surface point \mathbf{x} with normal \mathbf{n} in a viewing direction \mathbf{v} , and minimizing a photometric error w.r.t. the input images. However, achieving similar reconstructions from a single view remains a challenge due to the lack of multi-view cues, which are important to disambiguate geometric and color information. We propose an architecture that yields accurate 3D reconstructions without requiring multi-view information by leveraging two main inductive biases: first, we decompose f^{sdf} into a reference SDF and a deformation field [49]. We use this parameterisation as an implicit bias to constrain the

composed SDF to be close to the reference. Second, we pre-train f^{sdf} and r to represent a shape and appearance statistical model (SA-SM). Inspired by [30], at inference time we optimise the parameters of this statistical model to obtain a robust initialisation for the analysis-by-synthesis process. These inductive biases greatly improve the performance of SIRA over [30] for the single view setup.

In a second step, we factor the appearance of the reconstructed surface into a global illumination and spatially-varying diffuse and specular albedos using the physically-based renderer of [50]. To prevent shadows and specularities from being baked into the albedo, we first learn a statistical model of illumination and diffuse and specular albedos, or appearance factorization statistical model (AF-SM), which we later use to constrain the search of appearance parameters to a suitable subspace. At inference time, we fit the parameters of this statistical model to a new scene, obtaining a coarse initialisation. To recover personalised details outside of the statistical model, we fine-tune the neural fields that encode the appearance. We find that at this point it is important to use regularisation losses and scheduling to prevent the illumination from leaking into the albedo.

3.1. Statistical models

We use a collection of scenes, composed of raw head scans paired with multiview posed images, to learn the statistical models for shape and appearance (SA-SM), and appearance factorization (AF-SM). For every scene, indexed by $i = 1 \dots M$, we have a set of surface points $\mathbf{x} \in \mathcal{P}_s^{(i)}$ with associated normal vector \mathbf{n} . We project each surface point to the images where it is visible, obtaining a set $\mathcal{C}_{\mathbf{x}}^{(i)} = \{(\mathbf{c}, \mathbf{v})\}$ composed of pairs of associated RGB color \mathbf{c} , and viewing direction \mathbf{v} .

3.1.1 SA-SM architecture

We build a statistical model of shape and appearance, designed to enable the downstream task of single-view 3D



Figure 3: Inverse rendering of a scene under extreme illumination conditions using SIRA. Predictions from left to right: surface, normals, albedo, diffuse albedo, specular albedo and final rendering. Right-most: Input image.

reconstruction (Fig. 2-left). Our architecture is composed of two main neural field decoders: an SDF decoder, $f_{\theta_{\text{sdf}}, \mathbf{z}_{\text{sdf}}}^{\text{sdf}}$, and a non-physically-based rendering function decoder, $r_{\theta_r, \mathbf{z}_r}$. Here, $\mathbf{z}_{\text{sa}} = \{\mathbf{z}_{\text{sdf}}, \mathbf{z}_r\}$ are the latent vectors of the shape and appearance spaces of the SA-SM, and $\theta_{\text{sa}} = \{\theta_{\text{sdf}}, \theta_r\}$ are the parameters of their respective decoders.

The SDF decoder is structured in two sub-functions: a deformation function and a reference SDF. As we show in the results, this separation acts as an implicit bias against staying too far from the reference SDF, which stabilizes the single-view 3D reconstructions. The deformation function,

$$f_{\theta_{\text{def}}, \mathbf{z}_{\text{sdf}}}^{\text{def}} : \mathbb{R}^3 \rightarrow \mathbb{R}^{3+N_\gamma}, \quad \mathbf{x} \mapsto (\delta, \gamma), \quad (1)$$

parameterised by internal parameters θ_{def} together with the latent vector \mathbf{z}_{sdf} , maps input coordinates, \mathbf{x} , to a deformation 3-vector, δ . It also outputs an auxiliary feature vector γ of dimension N_γ , which encodes higher level information required by the differentiable renderer [48]. The predicted deformation is used to map an input coordinate \mathbf{x} to a coordinate \mathbf{x}_{ref} in a reference space where we evaluate a reference SDF $f_{\theta_{\text{ref}}}^{\text{ref}}$ parameterized by internal parameters θ_{ref} :

$$\mathbf{x}_{\text{ref}} = \mathbf{x} + \delta, \quad (2a)$$

$$f_{\theta_{\text{ref}}}^{\text{ref}} : \mathbb{R}^3 \rightarrow \mathbb{R}, \quad \mathbf{x}_{\text{ref}} \mapsto s. \quad (2b)$$

Putting them together we obtain the composed SDF decoder

$$f_{\theta_{\text{sdf}}, \mathbf{z}_{\text{sdf}}}^{\text{sdf}} : \mathbf{x} \mapsto f_{\theta_{\text{ref}}}^{\text{ref}}(\mathbf{x}_{\text{ref}}), \quad (3)$$

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

$$r_{\theta_r, \mathbf{z}_r} : (\mathbf{x}_{\text{ref}}, \mathbf{n}, \mathbf{v}, \gamma) \mapsto \mathbf{c} \quad (4)$$

parameterised by internal parameters θ_r and a latent vector \mathbf{z}_r . This function assigns an RGB color \mathbf{c} , to every combination of 3D coordinate in the reference space \mathbf{x}_{ref} , unit normal vector \mathbf{n} , and unit viewing direction vector \mathbf{v} .

3.1.2 SA-SM training

To train our SA-SM we follow an auto-decoder framework, where each scene is assigned a set of latents $\mathbf{z}_{\text{sa}}^{(i)} =$

$\{\mathbf{z}_{\text{sdf}}^{(i)}, \mathbf{z}_r^{(i)}\}$, which are optimized together with the statistical model parameters θ_{sa} . After training, we obtain the parameters $\theta_{\text{sa},0} = \{\theta_{\text{sdf},0}, \theta_r, 0\}$ such that any combination of latents $(\mathbf{z}_{\text{sdf}}, \mathbf{z}_r)$ inside the latent space correspond to a well-behaved SDF, $f_{\theta_{\text{sdf},0}, \mathbf{z}_{\text{sdf}}}^{\text{sdf}}$, and appearance, $f_{\theta_r, 0, \mathbf{z}_r}^{\text{rend}}$, of a human head. We drop the dependence on the decoder internal parameters.

To learn a space of head shapes, for each scene, we sample a set of points on the surface, $\mathcal{P}_s^{(i)}$, and compute the surface error loss $\mathcal{L}_{\text{Surf}}^{(i)} = \sum_{\mathbf{x}_j \in \mathcal{P}_s^{(i)}} |f_{\mathbf{z}_{\text{sdf}}^{(i)}}^{\text{sdf}}(\mathbf{x}_j)|$. We also sample another set uniformly taken from the scene volume, $\mathcal{P}_v^{(i)}$, and compute the Eikonal loss [11] $\mathcal{L}_{\text{Eik}}^{(i)} = \sum_{\mathbf{x}_k \in \mathcal{P}_v^{(i)}} (\|\nabla_{\mathbf{x}} f_{\mathbf{z}_{\text{sdf}}^{(i)}}^{\text{sdf}}(\mathbf{x}_k)\| - 1)^2$. We promote small-magnitude and zero-mean deformations, which avoids solutions where the deformations compensate for an unnecessarily offset or scaled reference SDF:

$$\mathcal{L}_{\text{Def}}^{(i)} = \frac{1}{|\mathcal{P}_s^{(i)}|} \left(\sum_{\mathbf{x}_j \in \mathcal{P}_s^{(i)}} \|\delta_j^{(i)}\|_2 + \left\| \sum_{\mathbf{x}_j \in \mathcal{P}_s^{(i)}} \delta_j^{(i)} \right\|_2 \right), \quad (5)$$

where $\delta_j^{(i)}$ is the deformation vector applied to the 3D point \mathbf{x}_j of the scene i .

Similarly to [49], we use a landmark consistency loss. We automatically annotate a set of 3D face landmarks $\{\mathbf{x}_l^{(i)}\}$ with $l = 1 \dots L$ for each scene, i , and use their deformed coordinate mismatch between pairs of scenes as a loss, $\mathcal{L}_{\text{Lm}}^{(i)} = \sum_{j \neq i} \sum_l^L \|\mathbf{x}_{\text{ref},l}^{(i)} - \mathbf{x}_{\text{ref},l}^{(j)}\|^2$, where $\mathbf{x}_{\text{ref},l}^{(i)}$ is the position of landmark l of scene i in the reference space.

The SA-SM learns a distribution of head appearances from the set of posed images accompanying every training scene. To evaluate the rendering function (eq. 4), we compute the coordinate in the reference space \mathbf{x}_{ref} associated to the surface point (eq. 2a), as well as the high-level descriptor γ (eq. 1). We also obtain the surface normals, \mathbf{n} , as the normalized gradient of the SDF [48]. With these, we define the color loss

$$\mathcal{L}_{\text{Col}}^{(i)} = \sum_{\mathbf{x} \in \mathcal{P}_s^{(i)}} \sum_{(\mathbf{c}, \mathbf{v}) \in \mathcal{C}_{\mathbf{x}}^{(i)}} \|r_{\mathbf{z}_r^{(i)}}(\mathbf{x}_{\text{ref}}, \mathbf{n}, \mathbf{v}, \gamma) - \mathbf{c}\|. \quad (6)$$

Finally, $\mathcal{L}_{\text{Emb}}^{(i)}$ enforces a zero-mean multivariate-Gaussian distribution with spherical covariance σ^2 over the spaces of shape and appearance latent vectors: $\mathcal{L}_{\text{Emb}}^{(i)} = \frac{1}{\sigma^2} (\|\mathbf{z}_{\text{sdf}}^{(i)}\|_2 + \|\mathbf{z}_r^{(i)}\|_2)$. Putting them together, we minimize the following:

$$\arg \min_{\{\mathbf{z}_{\text{sa}}^{(i)}\}, \theta_{\text{sa}}} \sum_i \mathcal{L}_{\text{Surf}}^{(i)} + \lambda_1 \mathcal{L}_{\text{Eik}}^{(i)} + \lambda_2 \mathcal{L}_{\text{Def}}^{(i)} + \lambda_3 \mathcal{L}_{\text{Lm}}^{(i)} + \lambda_4 \mathcal{L}_{\text{Col}}^{(i)} + \lambda_5 \mathcal{L}_{\text{Emb}}^{(i)} \quad (7)$$

where λ_{1-5} are scalar hyperparameters.

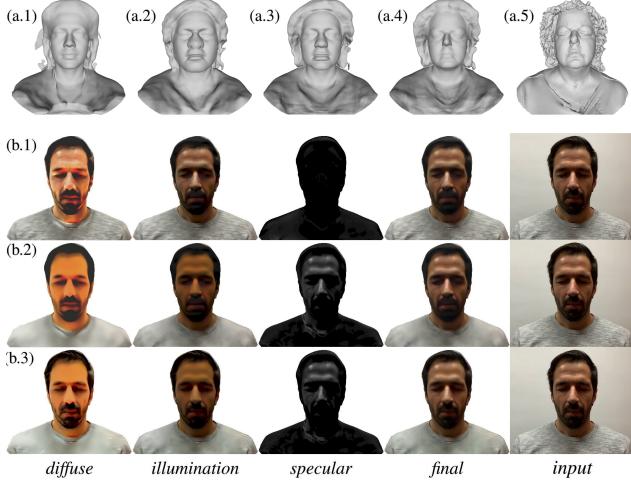


Figure 4: **Ablation study:** In (a) we ablate the 3D reconstruction method. In (b) we ablate the inverse-rendering method. See text for details.

3.1.3 AF-SM architecture

We build a statistical model of illumination and materials to enable the downstream task of single-view appearance factorization (Fig. 2-right). We use the physically-based differentiable rendering model introduced in [50] to capture the complex appearance of human faces, which can include shadows and view-dependent specular reflections [44].

We compute the radiance r^{pb} emitted from a surface point \mathbf{x} with normal \mathbf{n} in the viewing direction ω_o using the non-emitting rendering equation

$$\begin{aligned} r^{\text{pb}}(\omega_o, \mathbf{x}) &= r^d(\omega_o, \mathbf{x}) + k_s r^s(\omega_o, \mathbf{x}) = \\ &= \int_{\Omega} l(\omega_i) (f^d(\mathbf{x}) + k_s f^s(\mathbf{x}, \omega_i, \omega_o)) (\omega_i \cdot \mathbf{n}) d\omega_i, \end{aligned} \quad (8)$$

where $l(\omega_i)$ is the incident light from direction ω_i , the functions f^d, f^s are the diffuse and specular components of the BRDF respectively, and the scalar $k_s \in [0, 1]$ controls their relative weight. The functions r^d and r^s represent the integrated radiance corresponding to the diffuse and specular parts of the BRDF respectively. The integral is computed over the hemisphere $\Omega = \{\omega_i : \omega_i \cdot \mathbf{n} > 0\}$.

Following [50] and [43], we approximate the incident light, specular reflectance, and clamped cosine decay factor, with spherical gaussian decompositions, which allows us to efficiently compute the integral in closed form. We represent the environment map $l(\omega_i)$ as a mixture of N_l spherical gaussians. The diffuse component of the BRDF is a scaled spatially varying RGB albedo, $\mathbf{a} \in \mathbb{R}^3$, with no angular dependence: $f^d(\mathbf{x}) = \mathbf{a}(\mathbf{x})/\pi$. As for the specular component of the BRDF, f^s , we particularise the simplified Disney BRDF [13] used in [50] by fixing the value of the roughness parameter. For a given point in the surface and a given viewing direction, having fixed the roughness, the only free

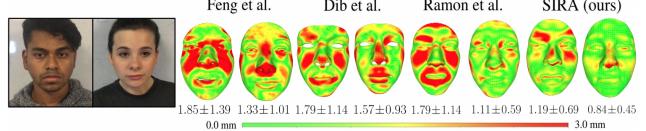


Figure 5: **Single-view 3D reconstruction:** Subjects from 3DFAW dataset [26]. Comparison: Feng 2021 [9], Dib 2021 [6], Ramon 2021 [30], SIRA (ours).

parameter of f^d is a spatially-varying monochrome specular albedo $s(\mathbf{x}) \in \mathbb{R}$. We provide more details about the rendering model in the supplementary material.

Determining the appearance of a scene then boils down to regressing the spatially-varying diffuse and specular albedos, \mathbf{a} and s , as well as the lighting parameters $\{\xi_l, \lambda_l, \mu_l\}$, where $\xi_l \in \mathbb{S}^2$ is the direction of the lobe, $\lambda_l \in \mathbb{R}_+$ is the lobe sharpness, and $\mu_l \in \mathbb{R}_+^n$ the lobe amplitude. We represent the AF-SM with three decoders. First, a diffuse and specular albedo decoders,

$$\mathbf{a}(\mathbf{x}) = \mathbf{a}_{\theta_a, \mathbf{z}_a}(\mathbf{x}_{\text{ref}}, \gamma) : \mathbb{R}^3 + N_\gamma \rightarrow \mathbb{R}^3 \quad (9a)$$

$$s(\mathbf{x}) = s_{\theta_s, \mathbf{z}_s}(\mathbf{x}_{\text{ref}}) : \mathbb{R}^3 \rightarrow \mathbb{R}, \quad (9b)$$

parameterised respectively by the internal parameters θ_a and θ_s , and with associated latent vectors \mathbf{z}_a and \mathbf{z}_s . These latent vectors are decoded to a space of continuous albedo functions, which we evaluate and supervise at the surface points. Instead of using the raw surface points as inputs to the decoders, we first process them with f^{def} , initialised with the parameters $\theta_{\text{def}, 0}$ and latent vectors corresponding to the geometry of each scene in the collection, $\mathbf{z}_{\text{def}}^{(i)}$, to obtain their correspondences in the learnt reference space, \mathbf{x}_{ref} (eq. 2a), as well as their associated descriptor γ (eq. 1). We provide these to the AF-SM to allow it to re-use the semantic information γ extracted by f^{def} for each surface point. In addition, we have an illumination decoder $l_{\theta_1, \mathbf{z}_1} \in \mathbb{R}^{N_l \times 7}$, parameterised by θ_1 and with associated latent vector \mathbf{z}_1 , which outputs the parameters of the N_l spherical gaussians that describe the environment map of a scene.

The color associated to surface point \mathbf{x} seen from viewing direction ω_o is computed by evaluating the physically-based rendering equation (eq. 8) with the parameters obtained from these decoders. Here, we denote this rendered color by $r_{\theta_{\text{pb}}, \mathbf{z}_{\text{pb}}}^{\text{pb}}(\mathbf{x}, \omega_o)$, where $\theta_{\text{pb}} = \{\theta_a, \theta_s, \theta_1\}$, and $\mathbf{z}_{\text{pb}} = \{\mathbf{z}_a, \mathbf{z}_s, \mathbf{z}_1\}$.

3.1.4 AF-SM training

These decoders are trained in an auto-decoder setup, assigning latent vectors $\mathbf{z}_{\text{pb}}^{(i)}$ to each scene. The appearance factorization model is trained in a self-supervised manner, since only multi-view images with known cameras are assumed. After training, we obtain the optimized parameters $\theta_{\text{pb}, 0} = \{\theta_{a, 0}, \theta_{s, 0}, \theta_{1, 0}\}$ such that any combination

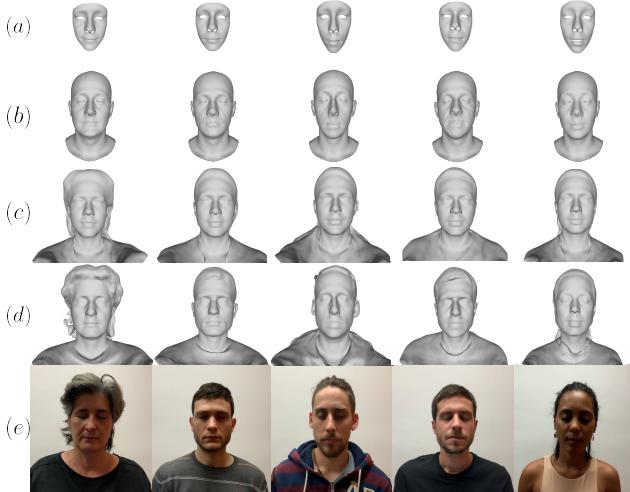


Figure 6: Single-view 3D reconstruction: Subjects from the H3DS dataset. (a) Dib 2021 [6], (b) Feng 2021 [9], (c) Ramon 2021 [30], (d) SIRA(Ours) and (e) input image.

of latents \mathbf{z}_{pb} inside the latent space correspond to a well-behaved illumination, as well as diffuse and specular albedos. During training, we set the parameter that controls the relative weight of the diffuse and specular components to $k_s = 1$, and minimize the following loss:

$$\arg \min_{\{\mathbf{z}_{\text{pb}}^{(i)}\}, \theta_{\text{pb}}} \mathcal{L}_{\text{Col}}^{(i)} + \lambda_6 \mathcal{L}_{\text{Emb}}^{(i)} + \lambda_7 \mathcal{L}_{\text{Reg}}^{(i)} \quad (10)$$

where λ_6 and λ_7 are hyperparameters. The first component of the loss, $\mathcal{L}_{\text{Col}}^{(i)}$, is the photometric error between the color rendered by r^{PB} and the ground truth images from different views. It is defined analogously to eq. 6. We also use a latent vector regularization, \mathcal{L}_{Emb} , defined analogously to the equivalent loss used to train the SA-SM. Finally, to avoid baking shadows and reflections in the diffuse albedo, we include a regularisation loss that encourages it to be spatially smooth: $\mathcal{L}_{\text{Reg}}^{(i)} = \sum_{\mathbf{x} \in \mathcal{P}_i} \|\mathbf{a}(\mathbf{x}) - \mathbf{a}(\mathbf{x} + \boldsymbol{\epsilon})\|$ where $\boldsymbol{\epsilon}$ is a 3D perturbation set as a hyperparameter.

3.2. Single-view inverse rendering

With our pre-trained statistical models at hand, we can tackle the task of obtaining a 3D reconstruction and a factorized appearance from a single portrait image \mathbf{I} with associated camera parameters \mathbf{C} and foreground mask \mathbf{M} .

3.2.1 Reconstructing geometry from a single image

To obtain 3D reconstructions of new scenes, we render the geometry described by f^{sdf} using the differentiable rendering function r of eq. 4, and minimize a photoconsistency error. For a pixel coordinate p of the input image \mathbf{I} , we march a ray $\mathbf{r} = \{\mathbf{c} + t\mathbf{v} | t \geq 0\}$, where \mathbf{c} is the position of the associated camera \mathbf{C} , and \mathbf{v} the viewing direction. We find the intersection coordinates with the composed SDF (eq. 3)

using sphere tracing. This intersection point can be made differentiable w.r.t \mathbf{z}_{sdf} and θ_{sdf} using implicit differentiation [21, 48]. The differentiable intersection coordinates \mathbf{x}_s are used to obtain their associated 3D displacement δ and feature vector γ (eq. 1), as well as their corresponding coordinates in the reference space \mathbf{x}_{ref} (eq. 2a), and normal vector $\mathbf{n} = \nabla_{\mathbf{x}} f^{\text{sdf}}$. Then, the color associated to the ray is computed as $\mathbf{c} = r(\mathbf{x}_{\text{ref}}, \mathbf{n}, \mathbf{v}, \gamma)$.

In order to optimize \mathbf{z}_{sa} and θ_{sa} , we minimize the same photoconsistency, mask, and eikonal losses as in [48]. See supplementary material for a more detailed explanation.

Instead of optimizing all the parameters $\{\theta_{\text{sdf}}, \theta_r, \mathbf{z}_{\text{sdf}}, \mathbf{z}_r\}$ at once, we propose a two-step schedule, better suited for the underconstrained one-shot scenario. We initialise the geometry and rendering functions with the parameters obtained with the pretraining described in the last section, $\{\theta_{\text{sdf},0}, \theta_{r,0}\}$. The initial shape and appearance latents, \mathbf{z}_{sdf} and \mathbf{z}_r , are picked from a multivariate normal distribution with zero mean and small variance, ensuring that they start near the mean of the latent spaces. In a first optimization phase, we only optimize the shape and appearance latents. This yields an initial approximation within the previously learnt shape and appearance latent spaces. In a second phase, we unfreeze the parameters of the deformation and rendering nets, $\{\theta_{\text{def}}, \theta_r\}$ (eqs. 1, 4), but not those of the reference SDF, θ_{ref} .

This schedule is key to obtaining accurate results in the one-shot regime. While unfreezing the deformation and rendering networks allows us to reach highly-detailed solutions outside of the pre-learnt latent spaces, the fact that we express the shape as a deformed reference SDF acts as a regularization that allows correct training convergence. We refer to the fine-tuned shape parameters as $\theta_{\text{def,ft}}$ and $\mathbf{z}_{\text{sdf,ft}}$.

3.2.2 Appearance factorization from a single image

Once f^{sdf} has been optimized for image \mathbf{I} , we dispose of the non-disentangled renderer $r_{\theta_r, \mathbf{z}_r}$, and tackle the appearance factorization problem. We use ray marching to obtain the 3D surface coordinates \mathbf{x} , corresponding to each pixel $p \in \mathcal{P}$ with a non-zero foreground mask value. We process these coordinates with f^{def} to compute their correspondences in the reference space, \mathbf{x}_{ref} , and associated descriptors γ , which are the inputs to the AF-SM.

To better capture personalized details for each scene, outside of the pretrained latent space, we extend the AF-SM with a diffuse albedo refinement module. We express the diffuse component of the BRDF (eq. 9) as:

$$\mathbf{a}(\mathbf{x}) = \mathbf{a}_{\theta_a, \mathbf{z}_a}(\mathbf{x}_{\text{ref}}, \gamma) + k_r \mathbf{a}_{\theta_{\text{ar}}}^r(\mathbf{x}_{\text{ref}}, \gamma) \quad (11)$$

where \mathbf{a}^r is an albedo refinement neural field parameterized by θ_{ar} . The scalar k_r controls the weight of the albedo refinement field. This separation of the albedo into a base and

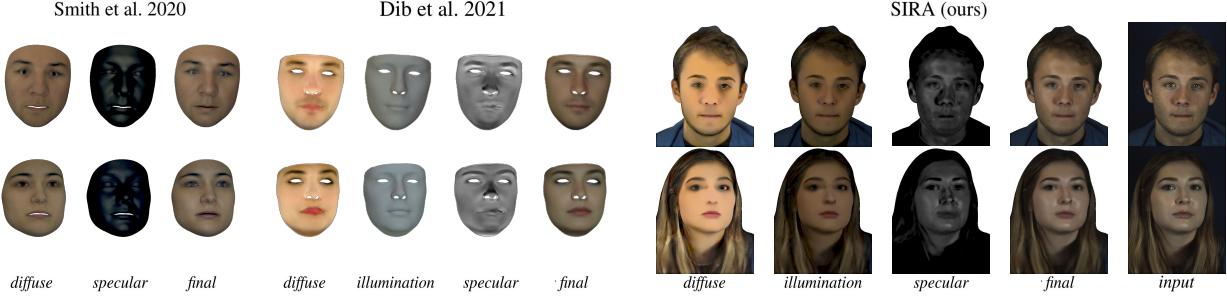


Figure 7: **Inverse rendering:** Subjects from 3DFAW-HR dataset. Comparison: Smith 2020 [35], Dib 2021 [6], and SIRA(ours). *Input* images are decomposed into *diffuse* albedo, diffuse *illumination*, *specular* radiance, and *final* render

refinement fields enables adding detail to the coarse albedo provided by the pretrained statistical model $\mathbf{a}_{\theta_a, \mathbf{z}_a}$, while using regularization losses that prevent the refinement \mathbf{a}^r from absorbing shading and specular information. In this section, we denote the rendered color using this modified model as $r^{pb}(\mathbf{x}, \omega_o)$, dropping the dependence on its internal parameters $\theta_{pb} = \{\theta_a, \theta_{ar}, \theta_s, \theta_l\}$ and latent vectors $\mathbf{z}_{pb} = \{\mathbf{z}_a, \mathbf{z}_s, \mathbf{z}_l\}$.

To optimize r^{pb} , we minimize the loss $\mathcal{L} = \mathcal{L}_{RGB} + \mathcal{L}_{Reg}$. The rendering photoconsistency loss $\mathcal{L}_{RGB} = |\mathcal{P}|^{-1} \sum_{p \in \mathcal{P}} |\mathbf{I}(p) - \mathbf{c}(p)|$ is defined on physically-based rendered color $\mathbf{c} = r^{pb}(\mathbf{x}, \omega_o)$ evaluated at the coordinates \mathbf{x} and viewing directions ω_o corresponding to pixel p . We introduce a regularization loss, \mathcal{L}_{Reg} , designed to prevent the albedo refinement from explaining color variations that should be captured by the diffuse or specular shading. This loss is defined as

$$\mathcal{L}_{Reg} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \|\mathbf{a}^r(p)\| w(p) \quad (12)$$

$$w = \lambda_8 \max(0, \|\mathbf{a}_{\theta_a, \mathbf{z}_a}\|_1 - \|r_b^d\|_1) + \lambda_9 \|r^s\|_1 \quad (13)$$

where r^s is the specular component of the radiance in the rendering equation (eq. 8), and r_b^d is the diffuse component of the radiance evaluated with the base albedo $\mathbf{a}_{\theta_a, \mathbf{z}_a}$. The scalars λ_8 and λ_9 are hyperparameters. The weighting function w has been designed to have high values for pixels where there are shadows or reflections. By regularizing the norm of \mathbf{a}^r for those pixels (eq. 12), it prevents the albedo refinement field from absorbing shading and specular information.

We do not optimize all the parameters at once, as this would tend to bake shadows and specularities into the albedo. To prevent this from happening, we design a suitable scheduling of the learning rates, as well as of the rendering model hyperparameters k_s and k_r . Roughly, our scheduling follows this order: we first learn illumination, recovering coarse shadows on a fixed initial albedo, without specularities or albedo refinement ($k_s = k_r = 0$). Then, we optimize \mathbf{z}_a , allowing the model to learn an albedo within the latent space. Next, we gradually add and optimize specular reflections ($k_s = 1$). With this in place, we freeze

the coarse albedo and unfreeze the albedo refinement module. During this stage, the model captures photo-realistic details in the albedo refinement field, while avoiding baking shades and reflections thanks to \mathcal{L}_{Reg} . A more fine-grained description of our training schedule can be found in the supplementary material.

4. Experiments

We next evaluate our 3D reconstruction and appearance factorization on multiple real-world portrait photos from the datasets H3DS [30], 3DFAW [26] and Wikihuman Project [1]. We train the SA-SM and AF-SM priors on the dataset used in [30]. See supplementary material for a more detailed explanation of the training and evaluation datasets. Training the priors and fitting SIRA for a scene takes about 1 day and 10 min respectively, on a single RTX 2080Ti GPU.

4.1. Ablation

We conduct an ablation study on the H3DS dataset and show the qualitative results in Figure 4 for both 3D reconstruction and appearance factorization.

3D reconstruction. We select as baseline the architecture proposed in [30]. Note in Fig. 4-top, that this architecture underfits the scene when only the latent vector is optimized (a.1) and it is unstable when the decoder is fine-tuned (a.2). By splitting the geometry into a deformation field and a reference sdf (a.3), we gain more control over the resulting 3D surfaces, which leads to more plausible and stable solutions even when the deformation decoder is fine-tuned. Finally, by jointly modelling a distribution of 3D shapes and appearances with the SA-SM, SIRA (a.4) is able to better disentangle geometric and visual information, providing 3D models that highly resemble to input image. The qualitative results of Fig. 4 are aligned with the errors reported in Table 1 (top), in which SIRA outperforms our ablated baselines by a significant margin.

Appearance factorization. As shown in Fig. 4-right, directly fitting the physically-based rendering model of SIRA to a scene, without introducing any prior, yields baked lights and shadows in the diffuse albedo, and fails to recover a

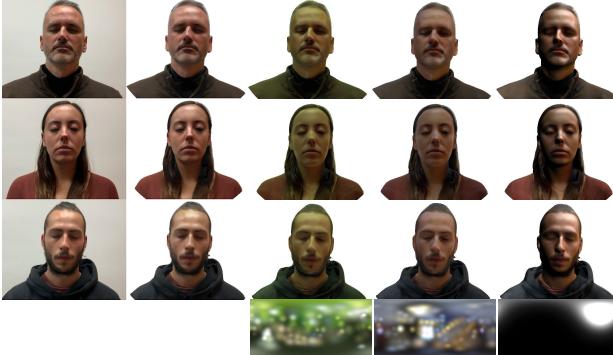


Figure 8: **Relighting** of inverse-rendered scenes. Subjects from the H3Ds dataset.

meaningful specular component (b.1). Using the AF-SM, together with a scheduling to guide the optimization (b.2), we correctly factor the appearance components. However, the results lack realism due to the albedo smoothness bias in the prior. To get sharper results (b.3), we introduce the albedo refinement module (Eq. 11).

4.2. 3D reconstruction comparison

We compare SIRA against the 3DMM-based methods Feng et al. 2021 [9] and Dib et al. 2021 [6], as well as the unconstrained method Ramon et al. 2021 [30].

Table 1 (top) reports the surface error in the facial area, using the unidirectional Chamfer distance from the predictions to the ground truth. SIRA achieves comparable results in all the datasets, outperforming the baselines on the 3DFAW low resolution subset and in H3DS. Moreover, as shown in Fig. 5, SIRA provides finer anatomical details in complex areas like the cheeks and nose.

Unlike 3DMM-based approaches [9, 6], SIRA recovers the geometry of the head, including hair and shoulders. This has an important perceptual impact, visible in Figure 6. While [30] reconstructs the same area, it underfits the input data. SIRA, in contrast, yields 3D shapes that clearly retain the identity of the person, yielding realistic relightable avatars.

4.3. Appearance factorization comparison

We next analyze our results for the task of appearance factorization. Fig. 7, shows qualitative comparison of SIRA against Smith et al. 2020 [35] and Dib et al. 2021 [6] on three cases from the 3DFAW high resolution dataset. SIRA performs similar to the baselines in the face region and additionally factorizes the intrinsic components of the hair and the upper body. Note that the appearance of the whole head is much more complex and diverse than the appearance of the skin in the facial region. Furthermore, SIRA models high frequency skin specularities, leading to photorealistic re-rendered images. This can be seen in Fig. 8, where three avatars reconstructed with SIRA are relighted under novel

	3DFAW ↓	3DFAW HR ↓	H3DS ↓
Feng [9]	1.53	1.46	1.62
Dib [6]	1.53	1.61	1.70
Ramon [30]	1.75	1.48	1.98
Ablation. Fig 4 a.2	-	-	2.17
Ablation. Fig 4 a.3	-	-	1.86
SIRA(Ours)	1.42	1.58	1.46

	Final (SSIM) ↑ (PSNR) ↑	Final (SSIM) ↑ (PSNR) ↑	Diffuse (SSIM) ↑ (PSNR) ↑	Diffuse (SSIM) ↑ (PSNR) ↑	Specular (SSIM) ↑ (PSNR) ↑	Specular (SSIM) ↑ (PSNR) ↑
Lattas [15]	-	-	0.83	21.7	0.58	17.5
Yamaguchi[47]	-	-	0.85	22.7	0.71	19.5
Dib [6]	0.91	27.5	0.83	20.0	0.62	14.6
Smith [35]	0.89	26.3	0.64	12.76	0.26	4.27
SIRA(Ours)	0.95	33.4	0.87	22.5	0.53	9.91

Table 1: **(top)** **3D reconstruction:** Average surface error in millimeters. **(bottom)** **De-rendering:** Evaluated on the Digital Emily scene. PSNR in dB.

lighting conditions. Note how the specularities move in the forehead when the light moves around the head (columns 5–7). Finally, it is also worth mentioning that SIRA is robust to extreme illumination conditions of the input image. This is shown in Fig. 3, where the input image is highly saturated by the scene illumination, but SIRA is still able to recover the intrinsic components correctly.

We report a quantitative analysis in Table 1 (bottom), using the Digital Emily scene. Compared to the baselines [6, 35, 47, 15], we achieve slightly better results in the final render, and comparable results in the diffuse and specular albedos, while being the only method that recovers the appearance of the entire head.

5. Conclusions

We have introduced SIRA, the first approach for building 3D avatars of human heads from a single image which, besides reconstructing high fidelity geometry, allows factorizing surface materials and global scene illumination. In order to tackle such an under-constrained problem, we have introduced two novel statistical models based on neural fields that encode shape and appearance into low dimensional latent spaces. A thorough evaluation has shown that SIRA provides SOTA results on full head geometry reconstruction, while also disentangling global illumination, and diffuse/specular albedos, yielding 3D relightable avatars from one single portrait image. Next avenues in this topic include speeding up the optimization process of neural fields.

Ethical considerations: Highly accurate photorealistic reconstructions can lead to identity impersonation concerns or image alteration. In addition, it is essential that the model is not biased and does not discriminate against any group, religion, colour, gender, age, or disability status. We include in the supplementary material the results on the Celeb-HQ dataset [16] to show that our model is diverse.

References

- [1] Emily. the wikihuman project. <https://vgl.ict.usc.edu/Data/DigitalEmily2/>. Accessed: 2022-03-05.
- [2] Ziqian Bai, Zhaopeng Cui, Jamal Ahmed Rahim, Xiaoming Liu, and Ping Tan. Deep facial non-rigid multi-view stereo. In *CVPR*, 2020.
- [3] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *CVPR*, 2016.
- [4] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12684–12694, 2021.
- [5] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan Barron, and Hendrik Lensch. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. *Advances in Neural Information Processing Systems*, 34, 2021.
- [6] Abdallah Dib, Cedric Thebault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevalier. 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*, pages 12819–12829, 2021.
- [7] Pengfei Dou and Ioannis A Kakadiaris. Multi-view 3d face reconstruction with deep recurrent neural networks. *Image and Vision Computing*, 80:80–91, 2018.
- [8] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020.
- [9] 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.
- [10] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021.
- [11] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020.
- [12] IEEE. A 3D Face Model for Pose and Illumination Invariant Face Recognition, Genova, Italy, 2009.
- [13] Brian Karis and Epic Games. Real shading in unreal engine 4. *Proc. Physically Based Shading Theory Practice*, 4(3):1, 2013.
- [14] Zhengfei Kuang, Kyle Olszewski, Menglei Chai, Zeng Huang, Panos Achlioptas, and Sergey Tulyakov. Neroic: Neural rendering of objects from online image collections. *arXiv preprint arXiv:2201.02533*, 2022.
- [15] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme: “Realistically renderable 3d facial reconstruction” in-the-wild”. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 760–769, 2020.
- [16] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [17] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.
- [18] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017.
- [19] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel 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*, pages 7210–7219, 2021.
- [20] 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 *European conference on computer vision*, pages 405–421. Springer, 2020.
- [21] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020.
- [22] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019.
- [23] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021.
- [24] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.
- [25] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *AVSS*, pages 296–301, 2009.
- [26] Rohith Krishnan Pillai, László Attila Jeni, Huiyuan Yang, Zheng Zhang, Lijun Yin, and Jeffrey F Cohn. The 2nd 3d face alignment in the wild challenge (3dfaw-video): Dense reconstruction from video. In *ICCV Workshops*, 2019.
- [27] Stylianos Ploumpis, Haoyang Wang, Nick Pears, William AP Smith, and Stefanos Zafeiriou. Combining 3d morphable models: A large scale face-and-head model. In *CVPR*, 2019.

- [28] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.
- [29] Eduard Ramon, Janna Escur, and Xavier Giro-i Nieto. Multi-view 3d face reconstruction in the wild using siamese networks. In *ICCV Workshops*, 2019.
- [30] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5620–5629, 2021.
- [31] Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data. In *3DV*, 2016.
- [32] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, 2017.
- [33] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *ICCV*, 2017.
- [34] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019.
- [35] William AP Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua B Tenenbaum, and Bernhard Egger. A morphable face albedo model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5011–5020, 2020.
- [36] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021.
- [37] Tiancheng Sun, Kai-En Lin, Sai Bi, Zexiang Xu, and Ravi Ramamoorthi. Neff: Neural light-transport field for portrait view synthesis and relighting. *arXiv preprint arXiv:2107.12351*, 2021.
- [38] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. In *Computer Graphics Forum*, volume 39, pages 701–727. Wiley Online Library, 2020.
- [39] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV*, 2017.
- [40] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard G Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *CVPR*, 2018.
- [41] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *CVPR*, 2019.
- [42] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *CVPR*, 2017.
- [43] Jiaping Wang, Peiran Ren, Minmin Gong, John Snyder, and Baining Guo. All-frequency rendering of dynamic, spatially-varying reflectance. In *ACM SIGGRAPH Asia 2009 papers*, pages 1–10. 2009.
- [44] Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, et al. Analysis of human faces using a measurement-based skin reflectance model. *ACM Transactions on Graphics (ToG)*, 25(3):1013–1024, 2006.
- [45] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *CVPR*, 2019.
- [46] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *arXiv preprint arXiv:2111.11426*, 2021.
- [47] Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018.
- [48] 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*, 33:2492–2502, 2020.
- [49] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12803–12813, 2021.
- [50] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah 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*, pages 5453–5462, 2021.
- [51] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nefactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (TOG)*, 40(6):1–18, 2021.
- [52] Yufeng Zheng, Victoria Fernández Abrevaya, Xu Chen, Marcel C Bühler, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. *arXiv preprint arXiv:2112.07471*, 2021.