MoSAR: Monocular Semi-Supervised Model for Avatar Reconstruction using Differentiable Shading

Abdallah Dib[1*] Luiz Gustavo Hafemann[1*] Emeline Got[1] Trevor Anderson[1] Amin Fadaeinejad[1,2]
Rafael M. O. Cruz[3] Marc-André Carbonneau[1]

[1]Ubisoft LaForge York University[2] Ecole de Technologie Supérieure[3]

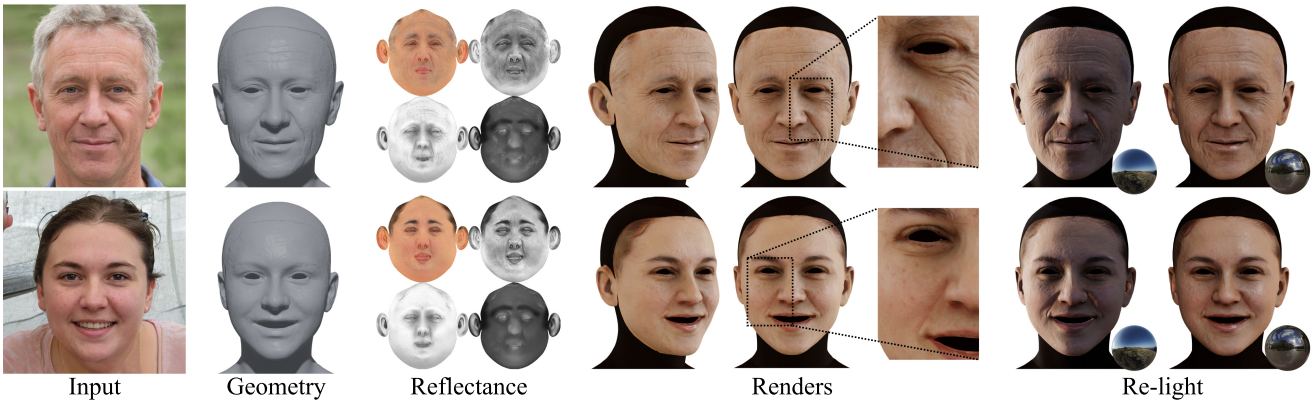

Figure 1. Our method estimates detailed geometry and reflectance maps, yielding convincing rendering under new lighting conditions.

Abstract

Reconstructing an avatar from a portrait image has many applications in multimedia, but remains a challenging research problem. Extracting reflectance maps and geometry from one image is ill-posed: recovering geometry is a one-to-many mapping problem and reflectance and light are difficult to disentangle. Accurate geometry and reflectance can be captured under the controlled conditions of a light stage, but it is costly to acquire large datasets in this fashion. Moreover, training solely with this type of data leads to poor generalization with in-the-wild images. This motivates the introduction of MoSAR, a method for 3D avatar generation from monocular images. We propose a semi-supervised training scheme that improves generalization by learning from both light stage and in-the-wild datasets. This is achieved using a novel differentiable shading formulation. We show that our approach effectively disentangles the intrinsic face parameters, producing relightable avatars. As a result, MoSAR[1] estimates a richer set of skin reflectance maps and generates more realistic avatars than existing state-of-the-art methods. We also release a new dataset, that provides intrinsic face attributes (diffuse, specular, ambient occlusion and translucency maps) for 10k subjects.

1. Introduction

Avatars are an important component of virtual worlds, being widely used in virtual reality, multimedia and video-games. Realistic personalized avatars, often called digital doubles, serve as bridge between the physical world and digital realities. The light stage [13] has long been regarded as the primary solution to obtain high-quality realistic avatars, however their high cost restricts their availability to the general public. They also require physical presence of users for scanning, which is impractical for many applications.

In contrast, creating avatars from a single monocular image enables a wide range of applications [6, 11, 16, 17, 19, 20, 34, 53–55, 66]. These methods are fast and robust to arbitrary capture conditions, however, they fall short in terms of quality for high-level multimedia production standards.

Moreover, these approaches often produce avatars that cannot seamlessly be integrated in tools used by content creators and artists, which limits the ability to edit the captured

*Equal contribution
[1]Project page: https://ubisoft-laforge.github.io/character/mosar

avatars. For example, modern graphic engines render realistic faces by relying on advanced materials and texture maps that these methods do not support. Even state-of-the-art methods [30–32, 40, 63] are limited to estimating diffuse, specular, and normal maps, while modern engines use a broader range of maps including ambient occlusion and translucency to simulate global illumination and subsurface scattering.

In this paper, we propose a new method for generating 3D avatars from a single monocular image. It estimates rich reflectance maps including ambient occlusion and translucency, in addition to diffuse, specular, and normal maps, all at a 4K resolution. Our method can separate light contributions from the ambient occlusion and translucency from the diffuse maps because it relies on high-quality light stage data. It generalizes well to uncontrolled capture conditions because it also trains on a large quantity of in-the-wild monocular images using self-supervision. This is made possible by our new differentiable shading formulation. This paper makes the following main contributions:

- We propose a semi-supervised training scheme leveraging high-quality light stage data, and large quantities of in-the-wild images. This allows for the estimation of a rich set of reflectance maps, as well as accurate geometry.
- We establish non-linear morphable models as a promising avenue to recover geometry from a single image by ranking in 2nd place on the REALY benchmark [4], so far largely dominated by linear models.
- We introduce a new differentiable shading formulation that incorporates, for the first time, the ambient occlusion and translucency maps with the spherical harmonics lighting model.

These contributions culminate in a more accurate and detailed head model than other state-of-the-art monocular avatar reconstruction methods. Figure 1 shows results from our method. In addition to our method's contribution, we introduce a new dataset, named *FFHQ-UV-Intrinsics*, the first public dataset providing intrinsic face attributes at scale (diffuse, specular, ambient occlusion and translucency maps) for a total of 10k subjects. We built it by applying our method to the dataset *FFHQ-UV*[2].

2. Related works

3D Geometry reconstruction Monocular face reconstruction is commonly treated in a self-supervised way [67], by modeling a parametric scene with geometry, light, reflectance, and camera parameters. These are used in conjunction with a differentiable renderer [28], enabling learning from images only. There are three main types of approaches for the parametric geometry model: (i) using linear morphable models (3DMM), (ii) learning unconstrained geometry, and (iii) learning non-linear 3DMMs.

Methods based on linear 3DMMs [5, 12, 15, 18, 20, 34, 49, 53, 65] are bound by the statistical space of the 3DMM, which restricts their expressiveness. Methods such as [47, 51, 64] estimate unrestricted geometry. These methods output dense meshes that require registration to a standard topology to be used in most applications.

Other methods such as [57, 58] learn a non-linear morphable model. These methods train a decoder that estimates the geometry under the weak supervision of another decoder that predicts 3DMM coefficients. The geometry produced by these methods often contains artifacts as noted in [18]. Graph neural networks (GNN) have been used to refine the geometry obtained from 3DMMs [21, 36], or to directly estimate geometry and reflectance [33]. In [1], a non-linear part-based GNN auto-encoder is trained for local shape editing.

Our method is not restricted nor regularized by linear 3DMMs and instead implements a non-linear morphable model using GNN auto-encoders [46], inspired by [33] and [1]. Our proposed semi-supervised training scheme mitigates artifacts in the geometry while keeping it unbounded.

Skin reflectance To reconstruct a relightable avatar from an image, the intrinsic skin reflectance attributes must be disentangled. Recovering face attributes from single in-the-wild images is an ill-posed problem. Most existing methods [20, 53] use a Lambertian Bidirectional Reflectance Distribution Function (BRDF) [8] and a statistical PCA basis for albedo estimation, which largely limits their expressivity.

Some methods go beyond a statistical basis. In [17–19], skin reflectance is modeled using a Microfacet BRDF [56], with diffuse and specular components only regularized by the PCA model. Nevertheless, this leads to blurry textures and fails to capture fine-grained details of the skin. Lee *et al.*[33] learn an unbounded skin reflectance model, however, the estimated texture map bakes light, geometry, and skin color (diffuse and specular) making it unsuitable for relighting. Finally, other methods [30–32, 40, 63] train a conditional GAN to estimate diffuse, specular, and normal maps for relightable avatars, but use only light stage data.

Our proposed method extends the Microfacet BRDF and estimates a richer skin reflectance than existing methods. It produces Ambient Occlusion and Translucency maps, all in 4K, enabling more realistic avatars. These maps are also compatible with modern technical artist tools and rendering engines. Our semi-supervised training scheme, which uses a novel differentiable shading formulation, allows for disentangling diffuse from ambient occlusion, as well as specular and translucency maps, from in-the-wild images.

3. Method

An overview of our method is shown in Figure 2. First, the *3D shape reconstruction* module (Section 3.1) estimates

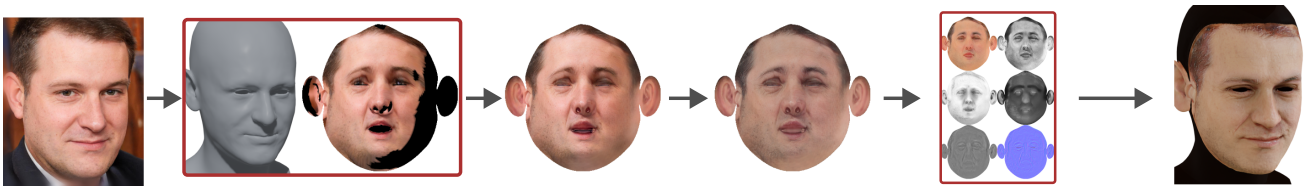

Figure 2. Overview of the proposed pipeline. In the first step, the 3D geometry of the face is estimated, and the image is projected onto the UV space. This texture is then used to estimate reflectance and displacement maps in 4K in a series of steps described in Section 3.

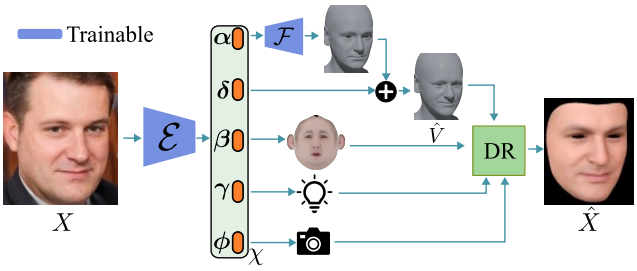

Figure 3. 3D Face reconstruction pipeline: An encoder $\mathcal{E}$ estimates the scene parameters χ, used by a Differential Renderer (DR) to obtain the estimated vertex colors, displayed here as an image.

the geometry and camera position used to project the input image onto UV space. Then, the *texture completion* network (Section 3.2) estimates the missing parts of the texture. Next, light is normalized across the entire face (Section 3.3), and the *intrinsic texture map estimation* (Section 3.4) recovers various texture maps (i.e diffuse, specular, displacement, ambient occlusion and translucency). Finally, a *super-resolution* network upscales the maps to 4K. We train this architecture in a semi-supervised manner, using (i) light stage data, that contains raw images, ground truth geometry, and reflectance maps, and (ii) in-the-wild images, for which we do not have ground truth.

3.1. 3D Shape Reconstruction

We represent the scene with a parametric model, defined by $\chi = \{\boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\phi}\}$, with neutral shape (identity) parameters $\boldsymbol{\alpha}$, expression coefficients $\boldsymbol{\delta}$, reflectance coefficients $\boldsymbol{\beta}$, light coefficients $\boldsymbol{\gamma}$ and camera parameters $\boldsymbol{\phi}$, as shown in Figure 3. This decomposition is similar to previous work [17, 22, 53], but we propose a novel procedure to learn a non-linear representation of identity.

We use a part-based GNN model inspired by [1, 46] to generate a detailed neutral mesh. While this method was proposed for artist-controlled geometry editing, our proposed semi-supervised training procedure makes it effective for monocular face reconstruction, improving modeling accuracy when compared to linear 3DMMs.

The GNN model is composed of a collection of encoders $\mathcal{G}_i$ each responsible for translating the vertices of a specific part of the face into a latent representation $\boldsymbol{\alpha}_i$ where $i \in [0, \ldots, 7]$. These are concatenated in a single vector $\boldsymbol{\alpha}$, and fed to a decoder $\mathcal{F}(\boldsymbol{\alpha})$, that reconstructs the entire input mesh. We use the same architecture as [1].

We represent facial expressions as a set of linear displacement blendshapes $\boldsymbol{E}$. Given parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$, we obtain the geometry as $\hat{V} = \mathcal{F}(\boldsymbol{\alpha}) + \sum_i \delta_i E_i$, where E_i is the i^{th} blendshape in the set.

For reflectance, we use a Lambertian BRDF model, with a PCA basis on per-vertex albedo colors. We model light using 3rd-order spherical harmonics, and we use a pinhole camera model.

Training We train two components at this stage: the GNN model producing neutral meshes and a scene-encoder $\mathcal{E}$ translating images into scene parameters χ. We first pretrain the GNN model in a supervised way with the light stage data, following the work of [1]. Next, we train the scene encoder $\mathcal{E}$ and the GNN decoder $\mathcal{F}$ in a semi-supervised manner. We combine the in-the-wild data with light stage data, for which we have ground truth vertex positions used for supervision.

When training with in-the-wild data, we rely solely on self-supervision. The differentiable renderer uses the estimated geometry $\hat{V}$, reflectance coefficients $\boldsymbol{\beta}$, light coefficients $\boldsymbol{\gamma}$ and camera parameters $\boldsymbol{\phi}$ to compute the final color (irradiance) for each vertex, as in [53]. These are then compared to their corresponding pixels in the input image X. Additionally, we extract landmarks l using an off-the-shelf detector [3]. We minimize the following loss function:

$$\mathcal{L}_u = L_{\text{photo}} + L_{\text{landmark}} + L_{\text{reg}}, \tag{1}$$

where the photo-consistency loss L_{photo} is the ℓ_1-norm between the predicted vertex colors and their associated pixel colors in the input image; L_{landmark} is the ℓ_2-norm of the distance between the landmarks l and the perspective projection of their associated mesh vertices. L_{reg} is the regularization term defined as:

$$L_{\text{reg}} = L_{\text{lap}} + L_{\text{light}} + L_{\text{exp}} + L_{\text{alb}}. \tag{2}$$

L_{lap} is the Laplacian smoothing operator from [23] applied to the estimated mesh $\hat{V}$, which minimizes the mean surface curvature. L_{exp} is the ℓ_1-norm of the expression coefficient vector acting as a regularization term. L_{light} encourages monochromatic light (similar to [17]), while L_{alb} regularizes against implausible face deformations using prior albedo statistics as defined in [53]. We note that the term L_{lap} if required since $\mathcal{F}$ is not restricted to a parametric model, and thus requires regularization to avoid producing implausible geometry.

When training with light stage data, we leverage the ground-truth mesh v_i associated with the image X_i. In that case we complement $\mathcal{L}_u$ with $L_{\text{supervised}}$ and L_{nrm} which are respectively the ℓ_2-distance between the vertices of the estimated and GT mesh, and the cosine distance between their normal vectors:

$$\mathcal{L}_s = \mathcal{L}_u + L_{\text{supervised}} + L_{\text{nrm}}. \tag{3}$$

We omitted weighting coefficients (λ) of the loss term for better readability. They are specified in the supplementary material.

Finally, at inference time, we further refine the scene parameters for the input image. Similarly to [17], we directly optimize the parameters χ to minimize Equation 1, using the Adam optimizer [29], with the values given by the scene encoder $\chi = \mathcal{E}(X)$ as initialization.

3.2. Texture completion

We project the image onto the UV space using the estimated geometry and camera information. This produces an incomplete texture because certain regions of the face are occluded, as illustrated in Figure 2.

Similarly to [14], we estimate a complete texture map from an incomplete map alongside its mirrored version. To train this model, we create synthetic textures simulating in-the-wild settings from light stage data. We render faces from random camera angles and HDR maps, then project onto UV space. This creates self-occlusions and non-visible areas similar to real images. We train an image-to-image translation network, by minimizing the ℓ_1-distance between the estimated texture and the ground truth.

3.3. Light normalization

The light normalization step aims to produce uniform light on the subject, mimicking the controlled light conditions of the light stage. This enables using simple light models for the remaining of the pipeline, such as Spherical Harmonics [44, 52], even for in-the-wild images, that can present complex lighting conditions. If this step is omitted, unexplained light contributions (e.g. strong directional lights) end up baked in the final estimated reflectance maps.

Figure 4 illustrates the training procedure for the light normalization network. We first perform the shading in UV space, using the mesh geometry V and its associated texture maps. We use random environment maps, camera positions with the rendering Equation 5. The resulting texture is fed to a U-Net architecture that generates a normalized texture $\mathcal{M}$. The network is trained to minimize the ℓ_1-distance between the estimated normalized texture and the target ground truth. The target texture is obtained by projecting the images from all light stage camera views onto UV space, and then aggregating them in a single image.

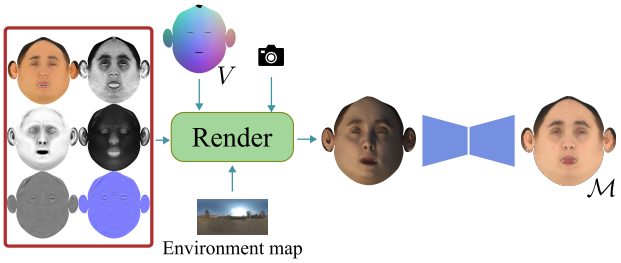

Figure 4. Light normalization step: We render faces in UV space with random light and camera positions. The network then estimates a light-normalized map $\mathcal{M}$.

3.4. Intrinsic texture maps estimation

The normalized texture $\mathcal{M}$ and the estimated geometry V are used to estimate different face attribute maps: Diffuse $\mathcal{D}$, Specular $\mathcal{S}$, Ambient Occlusion $\mathcal{A}$, Translucency $\mathcal{T}$ and Normal map $\mathcal{N}$, as depicted in Figure 5.

A separate network is used to estimate each of the reflectance maps. This architecture could be trained using only light stage data by supervising the output of each network. To achieve better generalization outside the light stage, we seek training on in-the-wild images as well. For this, we add an encoder C to estimate the residual light $\hat{\gamma}$ and the camera position $\hat{\phi}$, and introduce a new differentiable shading formulation (DS in Figure 5). DS outputs a new shaded texture $\hat{\mathcal{M}}$ which can then be compared to $\mathcal{M}$. This allows our model to be trained in a semi-supervised manner, combining light stage and in-the-wild datasets. The encoder C captures residual light that was not removed during the Light Normalization step (Section 3.3). This prevents the model from baking this light into one of the texture maps.

To estimate the normal map $\mathcal{N}$, we first predict the displacement map following a patch-based approach as in [7]. First, a PCA basis is built on patches of displacement maps from the light stage data. A coarse displacement encoder (refer to Figure 5), estimates the PCA coefficients that best reproduce a patch of the texture $\mathcal{M}$. This PCA captures repeated patterns (e.g. wrinkles) and low to medium frequency details. This approximation is further refined by a second U-Net model that complements the patch with high-frequency details (e.g. skin pores and moles). Finally, the

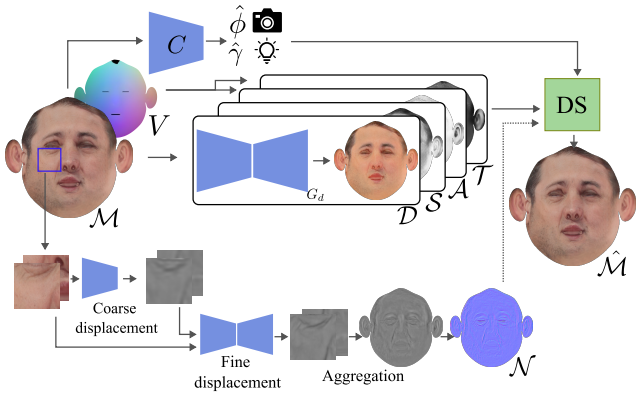

Figure 5. Texture estimation pipeline: Reflectance and normal maps are estimated by reconstructing the light-normalized texture $\mathcal{M}$ using Differentiable Shading (DS).

normal map is obtained from the displacement map using a Fast Fourier convolution.

To estimate the remaining maps $(\mathcal{D}, \mathcal{S}, \mathcal{A}, \mathcal{T})$, we use image-to-image translation networks. For the diffuse $\mathcal{D}$ and specular $\mathcal{S}$, the generators predict the maps from the normalized texture image $\mathcal{M}$. For $\mathcal{A}$ and translucency $\mathcal{T}$, the generators also use the geometry V projected onto UV space. We use adversarial training, by defining a discriminator for each texture map. These networks are trained jointly, with the following loss:

$$\mathcal{L} = L_{\text{shading}} + \sum_{\mathcal{D},\mathcal{S},\mathcal{A},\mathcal{T}} (L_{\text{GAN}}(G, D) + L_{\text{sup}}), \quad (4)$$

where L_{shading} is the ℓ_1 reconstruction loss between $\mathcal{M}$ and $\hat{\mathcal{M}}$, $L_{\text{GAN}}(G, D)$ is the adversarial loss described in [25], applied for each texture map separately. Finally, when training with light stage data, we also compute L_{sup} as the ℓ_1 loss between the ground truth and predicted maps. We omitted weighting coefficients (λ) of the loss term for better readability. They are specified in the supplementary material.

Differentiable shading. The differentiable shading *DS* (in Figure 5), takes the predicted textures ($\mathcal{D}$, $\mathcal{S}$, $\mathcal{A}$, $\mathcal{T}$ and $\mathcal{N}$), camera parameters $\hat{\phi}$, light coefficients $\hat{\gamma}$ and outputs a new texture map $\hat{\mathcal{M}}$. We formulate our shading equation, for each pixel $\hat{m}$ in $\hat{\mathcal{M}}$, extending the Microfacet BRDF model [56] as follows:

$$\hat{m} = \mathcal{B}_d + \mathcal{B}_{sss} + \mathcal{B}_s, \quad (5)$$

where $\mathcal{B}_d$, $\mathcal{B}_{sss}$ and $\mathcal{B}_s$ are the diffuse, subsurface scattering and specular contribution respectively, described in the equations below.

To model the diffuse component of the skin $\mathcal{B}_d$, we extend the diffuse Lambertian model with an additional term a that captures residual diffuse ambient occlusion in areas that receive less light such as wrinkles, folds, and nostrils. $\mathcal{B}_d$ is calculated as follows:

$$\mathcal{B}_d = d \cdot a \cdot \sum_{l=0}^{2} \sum_{m=-l}^{l} A_l \cdot \hat{\gamma}_l^m \cdot Y_l^m(n), \quad (6)$$

where d, is the diffuse albedo obtained from the diffuse $\mathcal{D}$ map, and a is the residual ambient occlusion term from $\mathcal{A}$. A_l are the Lambertian BRDF coefficients [43, 44]. $\hat{\gamma}$ are the SH coefficients, and Y_l^m are the SH basis [44]. n is the pixel normal (using normal mapping). For the diffuse component, we use 3rd-order spherical harmonics ($l = 2$) as it has been shown in [43] that it captures 99% of the reflected radiance.

Next, we introduce a new subsurface scattering contribution $\mathcal{B}_{sss}$ which indicates where thinner parts of the face (e.g. nose, lips, and ears) scatter more light than other thicker areas, defined as:

$$\mathcal{B}_{sss} = d \cdot \sum_{l=1}^{2} \sum_{m=-l}^{l} S_l \cdot \hat{\gamma}_l^m \cdot Y_l^m(n), \quad (7)$$

with $S_l = e^{-l^2/t^4}$, with t being the translucency value from $\mathcal{T}$. Intuitively, thicker parts of the face would have a low value in $\mathcal{T}$ which in turn would negate the $\mathcal{B}_{sss}$ contribution through S_l nearing 0.

Finally, we compute the specular component $\mathcal{B}_s$ as the spatial convolution of the SH light representation with the BRDF roughness kernel. This kernel is constant in the simplified Microfacet BRDF model that we use. Thus our specular contribution $\mathcal{B}_s$ is defined as:

$$\mathcal{B}_s = f \cdot \sum_{l=0}^{8} \sum_{m=-l}^{l} R_l \cdot \hat{\gamma}_l^m \cdot Y_l^m(r), \quad (8)$$

where $f = s + (1 - s)(1 - \cos\theta)^5$ is the Fresnel reflection [59] calculated using Schlick's approximation [50]. f quantifies the proportion of light that is reflected and depends on the angle at which light hits the surface. θ is approximated with the angle between the normal vector and the camera view direction [45]. $s \in \mathbb{R}$ is specular intensity from $\mathcal{S}$. r_i is the reflection vector of the viewing vector (obtained from the camera position $\hat{\phi}$) according to the surface normal [59], and R_l are the SH coefficients of the BRDF function corresponding to the roughness [37, 45]. For the specular contribution, we use 9th-order spherical harmonics ($l = 8$) for better approximation of the specular reflection [18].

Equation 5 is fully differentiable with respect to all texture maps, allowing for gradient-based optimization of the generators. We note that the diffuse d and ambient occlusion a have the same contribution, and therefore cannot be separated using only self-supervised training. For instance,

the model can bake all the information in a single diffuse map which results in a sub-optimal separation. Our semi-supervised training scheme (Equation 4) allows for this separation, by leveraging ground truth data from light stage in conjunction with the unlabeled data.

3.5. Super-resolution

We use super-resolution at the last step, to obtain 4K texture maps. We train two ESRGAN models [61]: one that upsamples the input to 1K and one that upsamples to 4K, as in [35]. Both are trained with data from the light stage.

4. Experimental protocol

Datasets The light stage dataset is composed of 890 subjects, captured with a neutral pose. For every subject, 12 cameras capture frontal and side views. The geometry is obtained using a Multi-View reconstruction pipeline, followed by registration to a standard topology. Each subject has diffuse, specular, displacement, ambient occlusion and translucency maps in 4K. We keep 50 subjects for validation and use the remaining for training.

We train the 3D reconstruction model with the light stage dataset and 70k images from the FFHQ dataset [26]. We crop and resize the images to the resolution of 512×512. We apply standard data augmentation (brightness, rotation, scale, and flip). For texture estimation, we use the light stage dataset and FFHQ-UV [2], which contains 54k textures obtained from StyleGAN [27]. We normalize the FFHQ-UV textures using our light normalization network. The textures are processed in resolution 512×512.

Training For the 3D geometry estimation (Section 3.1), $\mathcal{E}$ is a ResNet-50 [24]. For $\mathcal{F}$ we use the same architecture as [1] from their open-source implementation. We train both networks for 20 epochs, using a batch size of 32 images, equally divided between light stage and in-the-wild data.

The texture completion network is trained for 20 epochs on the light stage data following the data augmentation described in Section 3.2. We train the light normalization network for 20 epochs, using random environment maps sampled from a collection of 190 HDR maps[2].

The displacement map estimation is done in patches in the full resolution (4K). We manually selected 20 wrinkled subjects from the training set to obtain the PCA basis for coarse displacement, as done in [7]. After training the coarse displacement encoder for 5 epochs, we jointly train the coarse and fine displacement networks for 15 epochs.

The texture estimation networks are trained on the light stage data and the FFHQ-UV dataset. We use a batch size of 2 containing a labeled and an unlabeled image. All the generators are 5-layer U-Nets with skip connections [48]. We use the patch GAN multi-scale discriminators from [41].

[2] https://polyhaven.com

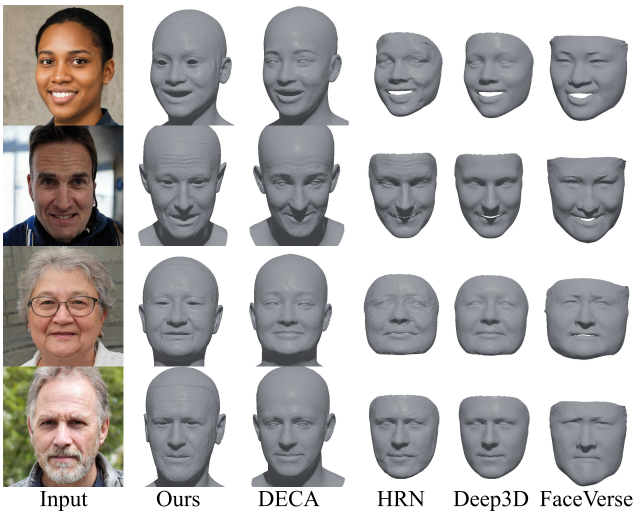

Figure 6. Comparison of 3D face reconstruction methods.

5. Results

5.1. Geometry estimation

In this section, we evaluate our method for 3D shape reconstruction. We compare to different state-of-the-art methods, including methods that specialize in geometry prediction without supporting texture map generation.

First we visually compare our results to DECA [20], HRN [34], Deep3D [15] and FaceVerse [60]. Figure 6 shows the estimated geometry of each method under the same rendering conditions. Our method produces meshes with more similar likeness to the input images, notably on the nose, eyes, cheeks, and the shape of the face. It also better captures higher-frequency details such as wrinkles and folds than the other methods. While HRN captures some of the details, they are not as fine and precise as ours. Moreover, HRN produces noticeable artifacts around the eyes and nose. The details captured by DECA are not as visible and sometimes missing. Deep3D does not predict displacement maps and does not capture any of the folds and wrinkles. DECA, Deep3D, and FaceVerse avoid baking glasses in the geometry, while HRN and our method do. The supplementary material contains additional comparisons.

For a quantitative analysis, we compare our method to state-of-the-art methods using REALY [4], a publicly available benchmark. This evaluation framework uses multi-view rendered portrait images of 100 high-quality scans from the HeadSpace dataset [9, 10]. Geometric reconstruction errors are computed separately on different face regions which reduces imprecision caused by alignment errors and allows for a fine-grained analysis of the results. For each

Method	Error (mm) Nose	Mouth	Forehead	Cheeks	All
MICA [66]	1.585 ± 0.325	3.478 ± 1.204	2.374 ± 0.683	1.099 ± 0.324	2.134
DECA [20]	1.697 ± 0.355	2.516 ± 0.839	2.394 ± 0.576	1.479 ± 0.535	2.010
PSL [39]	1.708 ± 0.349	1.708 ± 0.563	2.350 ± 0.551	1.593 ± 0.540	1.882
AlbedoGan [42]	1.656 ± 0.374	2.087 ± 0.839	2.102 ± 0.512	1.141 ± 0.303	1.746
Deep3D [15]	1.719 ± 0.354	1.368 ± 0.439	2.015 ± 0.449	1.528 ± 0.501	1.657
HRN [34]	1.722 ± 0.330	1.357 ± 0.523	1.995 ± 0.476	1.072 ±0.333	1.537
HiFace (w/o synthetic) [5]	1.227 ± 0.407	1.787 ± 0.439	1.454 ± 0.382	1.762 ± 0.436	1.558
HiFace [5]	1.036 ± 0.280	1.450 ± 0.413	1.324 ± 0.334	1.291 ± 0.362	1.275
Baseline (linear 3DMM)	1.815 ± 0.516	1.725 ± 0.576	2.550 ± 0.825	1.469 ± 0.511	1.890
MoSAR (ours)	1.499 ± 0.366	1.424 ± 0.462	1.950 ± 0.559	1.128 ± 0.303	1.500

Table 1. Reconstruction errors on the REALY benchmark.

face region, the mean of the per-vertex ℓ_2-distance between the predictions and the ground truth is reported in Table 1.

Our method ranks second on reconstruction error averaged for all parts, being surpassed only by the full HiFace model [5] when trained with 200k synthetic images with ground-truth meshes. The addition of the synthetic dataset significantly improves HiFace's results, which otherwise would perform a bit worse than ours. We believe that our method would similarly benefit from this additional data.

Our results indicate that non-linear morphable models can perform competitively compared to the dominating linear 3DMM. We included in Table 1 an implementation of our method using a linear 3DMM, a PCA created with the same light stage dataset. Except for the geometry model, all other factors were kept constant (i.e. dataset, encoder architecture). These results show that the improvement stems from the higher expressivity of non-linear modeling.

5.2. Texture estimation ablation

This ablation study evaluates the effect of light normalization (Section 3.3) and the proposed semi-supervised training scheme (Section 3.4) on the quality of the predicted texture maps. As a baseline, we train our system in a fully-supervised manner without light normalization, using only light stage data. This means that our differential shading is not required for self-supervision, thus, we ignore the term L_{shading} on Equation 4. Then, we include the light normalization to the baseline model to measure its impact. For completeness, we train a model in a purely self-supervised manner, without any light stage data.

To evaluate these systems on settings similar to in-the-wild, while having access to ground truth maps, we render subjects from our light stage validation data under novel lighting conditions using random HDR maps. From 50 subjects, we generate 1000 renders. We report the Structural Similarity Index (SSIM) [62] between the predicted and ground-truth maps.

Results are shown in Table 2. The supervised method, trained using only light stage data underperforms, indicating poor generalization when confronted with the variety of lighting conditions in the wild. Our light normalization step reduces this problem by attenuating specular highlights and strong shadows. Consequently, it improves SSIM, especially on the specular map. On the other hand, the self-supervised method, which is trained without any ground-truth textures, fails to disentangle the contributions of different maps, as expected, since they are entangled in the rendering Equation 5. It bakes most of the information in the diffuse albedo map and is therefore unsuitable for relighting applications. Our full model obtains the best SSIM. It takes advantage of high-quality light stage data to better disentangle the contributions of different maps, as opposed to the self-supervised-only version. Moreover, training from in-the-wild images exposes the model to a wide variety of lighting conditions and allows for training with a larger quantity of data. This translates to improvements on every map, but especially the specular map.

Method	Diffuse	Specular	AO	Transl.
Supervised	0.69	0.28	0.61	0.67
+ Light Normalization	0.80	0.55	0.82	0.79
Self-supervised	0.80	0.19	0.36	0.65
Full (Semi-supervised)	**0.83**	**0.65**	**0.85**	**0.82**

Table 2. SSIM metric between estimated and ground truth texture maps, for different settings of our method (higher is better).

Aside from improving SSIM, our full model produces visually sharper and cleaner maps when applied to in-the-wild data, as shown in Figure 8. Our semi-supervised training helps to better separate the intrinsic face attributes. For instance, the supervised-only model bakes specular reflections in the diffuse map. This is observed on the specular highlights of the nose, which our full model successfully removed from the diffuse map. The supervised model also bakes into the diffuse map both light and shadows that were not completely removed by the light normalization step. Finally, the AO map produced by the full method better captures the shading of the wrinkles and folds. The supplementary material contains more results with challenging lighting conditions.

5.3. Avatar reconstruction

In this section, we perform a visual comparison with recent state-of-the-art methods that perform full-avatar reconstruction: FitMe [32] and Relightify [40]. Both methods estimate geometry, diffuse, specular and normal maps from a single image. Figure 7 shows, for each method, the estimated geometry, and a rendering of the avatar, under the same conditions (camera and light). We focus the comparison on the skin reflectance, and therefore we removed the scalp, eyes and mouth interior of FitMe and Relightify since our method do not estimate them. We refer the reader to the supplementary material for additional comparisons.

Our approach captures subtle geometric details, including wrinkles and folds, in contrast to other methods, which struggle to reproduce these nuances. Notably, the shadows caused by wrinkles (see the first row) are baked in the

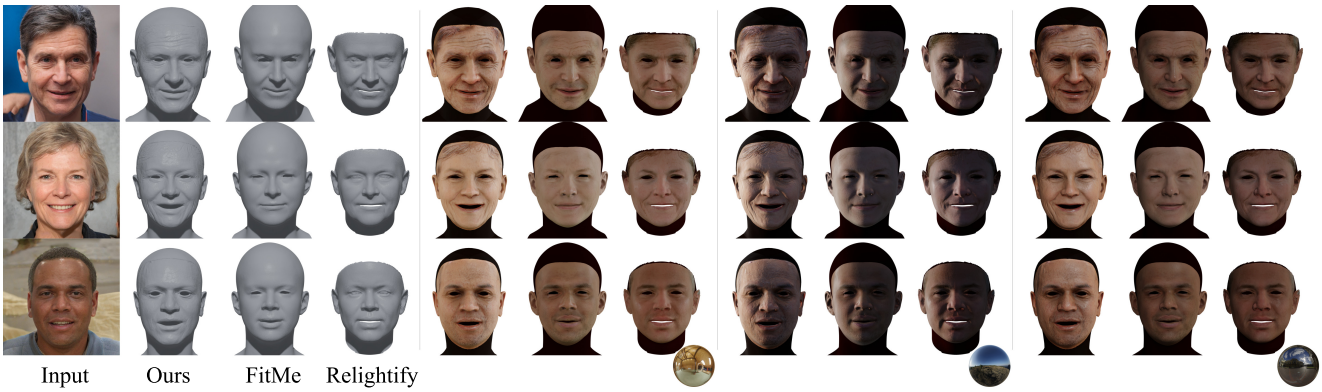

Figure 7. Comparison of the estimated geometry and renders under 3 lighting conditions, against FitMe [32] and Relightify [40]

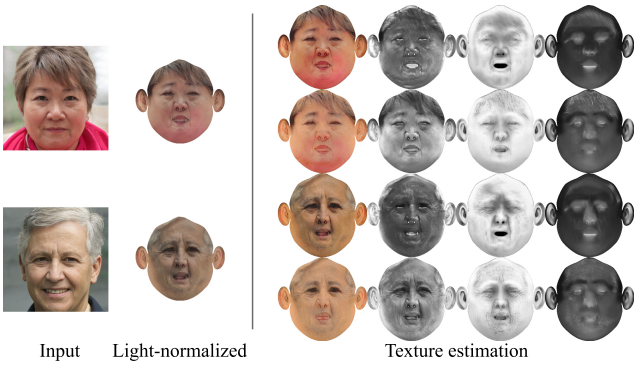

Figure 8. Reflectance estimation using the "supervised + LN" model (rows 1 and 3) vs the full method (rows 2 and 4).

Method	Diffuse	Specular	AO	Transl.
Relightify	0.704	-	-	-
FitMe	0.720	-	-	-
MoSAR (Ours)	0.754	0.515	0.766	0.806

Table 3. Comparison of SSIM between estimated and ground truth texture maps, for different methods.

albedo for both FitMe and Relightify, since they were not captured in the geometry, which causes unrealistic wrinkles in other light conditions. Our method is better at capturing the facial structure (cheek, nose, eyes) and expression, resulting in a likeness that more faithfully resembles the original image. For darker skin tones, while our method has a better shape estimation, it tends to yield lighter skin tones.

For quantitative analysis, we estimate reflectance maps from portrait images of 25 subjects for which we have ground truth reflectance maps (obtained separately in a light stage). For FitMe and Relightify, we do not compare the specular maps as they use a different formulation than ours (Eq. (8)). We compute the SSIM between estimated and ground truth texture maps, shown in Table 3. Our model achieves higher SSIM on the diffuse map, showing a better intrinsic separation of the face attributes.

6. Limitations

Our method does not explicitly model external occluders (e.g. eyeglasses, hair) which get baked into the texture maps. Also, our method is more successful in preserving the skin tones of caucasian subjects. This is mainly due to the representation imbalance in our in-the-wild dataset. As shown in [38], FFHQ is composed of around 69% white, 4% black, and 27% of other races. This biases the model towards generating lighter skin tones. Interestingly, we do not notice the same bias in the 3D shape reconstruction.

7. Conclusion

In this work, we presented MoSAR, a complete framework for creating realistic, relightable avatars from a single portrait image. It produces detailed geometry and skin reflectance maps at 4K resolution, compatible with modern rendering engines. We proposed a semi-supervised training procedure that includes a novel differentiable shading formulation that allows for estimating ambient occlusion and translucency. Our experiments highlighted the benefit of our semi-supervised training and showed that we obtain competitive results for geometry estimation using non-linear morphable models. This amounts to realistic and visually appealing renderings when compared to existing state-of-the-art methods. Finally, we introduced a dataset containing intrinsic facial attributes for 10k subjects of the FFHQ-UV dataset to accelerate research on this field. For future work, we intend to collect a more balanced dataset to reduce bias in our results. We want to model external occluders and reduce artifacts caused by hairs.

8. Acknowledgments

We would like to thank Dr. Alexandros Lattas and Foivos Papantoniou for running the FitMe and Relightify methods on the selected subjects. We thank Amaury Depierre for assistance with the super-resolution network.

References

[1] Mohammad Amin Aliari, Andre Beauchamp, Tiberiu Popa, and Eric Paquette. Face editing using part-based optimization of the latent space. In *Computer Graphics Forum*, pages 269–279. Wiley Online Library, 2023. 2, 3, 6

[2] Haoran Bai, Di Kang, Haoxian Zhang, Jinshan Pan, and Linchao Bao. Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2, 6

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 3

[4] Zenghao Chai, Haoxian Zhang, Jing Ren, Di Kang, Zhengzhuo Xu, Xuefei Zhe, Chun Yuan, and Linchao Bao. Realy: Rethinking the evaluation of 3d face reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 6

[5] Zenghao Chai, Tianke Zhang, Tianyu He, Xu Tan, Tadas Baltrusaitis, HsiangTao Wu, Runnan Li, Sheng Zhao, Chun Yuan, and Jiang Bian. Hiface: High-fidelity 3d face reconstruction by learning static and dynamic details. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9087–9098, 2023. 2, 7

[6] Prashanth Chandran, Sebastian Winberg, Gaspard Zoss, Jérémy Riviere, Markus Gross, Paulo Gotardo, and Derek Bradley. Rendering with style: combining traditional and neural approaches for high-quality face rendering. *ACM Transactions on Graphics (ToG)*, 40(6):1–14, 2021. 1

[7] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9429–9439, 2019. 4, 6

[8] Robert L Cook and Kenneth E. Torrance. A reflectance model for computer graphics. *ACM Transactions on Graphics (TOG)*, 1(1):7–24, 1982. 2

[9] Hang Dai, Nick Pears, William Smith, and Christian Duncan. Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision*, 2019. 6

[10] Hang Dai, Nick Pears, William Smith, and Christian Duncan. Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision*, 128:547–571, 2020. 6

[11] Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20322, 2022. 1

[12] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 2

[13] Paul Debevec. The light stages and their applications to photoreal digital actors. *SIGGRAPH Asia*, 2(4):1–6, 2012. 1

[14] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7093–7102, 2018. 4

[15] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 6, 7

[16] Abdallah Dib, Gaurav Bharaj, Junghyun Ahn, Cedric Thebault, Philippe-Henri Gosselin, and Louis Chevallier. Face reflectance and geometry modeling via differentiable ray tracing. *ACM SIGGRAPH European Conference on Visual Media Production (CVMP)*, 2019. 1

[17] Abdallah Dib, Gaurav Bharaj, Junghyun Ahn, Cédric Thébault, Philippe-Henri Gosselin, Marco Romeo, and Louis Chevallier. Practical face reconstruction via differentiable ray tracing. *Computer Graphics Forum*, 2021. 1, 2, 3, 4

[18] Abdallah Dib, Cedric Thebault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 5

[19] Abdallah Dib, Junghyun Ahn, Cedric Thebault, Philippe-Henri Gosselin, and Louis Chevallier. S2f2: Self-supervised high fidelity face reconstruction from monocular image. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2023. 1, 2

[20] 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. 1, 2, 6, 7

[21] Zhongpai Gao, Juyong Zhang, Yudong Guo, Chao Ma, Guangtao Zhai, and Xiaokang Yang. Semi-supervised 3d face representation learning from unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 348–349, 2020. 2

[22] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Perez, and Christian Theobalt. Reconstruction of Personalized 3D Face Rigs from Monocular Video. {*ACM*} *Trans. Graph. (Presented at SIGGRAPH 2016)*, 35(3):28:1–28:15, 2016. 3

[23] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. 3d-coded: 3d correspondences by deep deformation. In *Proceedings of the european conference on computer vision (ECCV)*, pages 230–246, 2018. 4

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016. 6

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 5

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 6

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 6

[28] Hiroharu Kato, Deniz Beker, Mihai Morariu, Takahiro Ando, Toru Matsuoka, Wadim Kehl, and Adrien Gaidon. Differentiable rendering: A survey. *arXiv preprint arXiv:2006.12057*, 2020. 2

[29] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015. 4

[30] 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. 2

[31] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9269–9284, 2021.

[32] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Fitme: Deep photorealistic 3d morphable model avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8629–8640, 2023. 2, 7, 8

[33] Gun-Hee Lee and Seong-Whan Lee. Uncertainty-aware mesh decoder for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6100–6109, 2020. 2

[34] Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 394–403, 2023. 1, 2, 6, 7

[35] Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, et al. Learning formation of physically-based face attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3410–3419, 2020. 6

[36] Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5891–5900, 2020. 2

[37] Dhruv Mahajan, Ravi Ramamoorthi, and Brian Curless. A theory of frequency domain invariants: Spherical harmonic identities for brdf/lighting transfer and image consistency. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):197–213, 2007. 5

[38] Vongani H. Maluleke, Neerja Thakkar, Tim Brooks, Ethan Weber, Trevor Darrell, Alexei A. Efros, Angjoo Kanazawa, and Devin Guillory. Studying bias in gans through the lens of race. In *Computer Vision – ECCV 2022: 17th European Conference*, page 344–360, Berlin, Heidelberg, 2022. Springer-Verlag. 8

[39] Christopher Otto, Prashanth Chandran, Gaspard Zoss, Markus Gross, Paulo Gotardo, and Derek Bradley. A Perceptual Shape Loss for Monocular 3D Face Reconstruction. *Computer Graphics Forum*, 2023. 7

[40] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, and Stefanos Zafeiriou. Relightify: Relightable 3d faces from a single image via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 7, 8

[41] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 6

[42] Aashish Rai, Hiresh Gupta, Ayush Pandey, Francisco Vicente Carrasco, Shingo Jason Takagi, Amaury Aubel, Daeil Kim, Aayush Prakash, and Fernando De la Torre. Towards realistic generative 3d face models. *arXiv preprint arXiv:2304.12483*, 2023. 7

[43] Ravi Ramamoorthi and Pat Hanrahan. An Efficient Representation for Irradiance Environment Maps. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, pages 497–500, New York, NY, USA, 2001. ACM. 5

[44] Ravi Ramamoorthi and Pat Hanrahan. On the relationship between radiance and irradiance: determining the illumination from images of a convex Lambertian object. *JOSA A*, 18 (10):2448–2459, 2001. 4, 5

[45] Ravi Ramamoorthi and Pat Hanrahan. A signal-processing framework for inverse rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 117–128, 2001. 5

[46] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3D faces using convolutional mesh autoencoders. In *European Conference on Computer Vision (ECCV)*, pages 725–741, 2018. 2, 3

[47] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1259–1268, 2017. 2

[48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 6

[49] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019. 2

[50] Christophe Schlick. An inexpensive brdf model for physically-based rendering. In *Computer graphics forum*, pages 233–246. Wiley Online Library, 1994. 5

[51] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1576–1585, 2017. 2

[52] Peter-Pike Sloan, Jan Kautz, and John Snyder. Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 339–348. 2023. 4

[53] Ayush Tewari, Michael Zollöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 3, 4

[54] Ayush Tewari, Michael Zollöfer, Florian Bernard, Pablo Garrido, Hyeongwoo Kim, Patrick Perez, and Christian Theobalt. High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018.

[55] A Tewari, F Bernard, P Garrido, G Bharaj, et al. *FML: Face Model Learning from Videos*. CVPR, 2019. 1

[56] Kenneth E Torrance and Ephraim M Sparrow. Theory for off-specular reflection from roughened surfaces. *Josa*, 57 (9):1105–1114, 1967. 2, 5

[57] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7346–7355, 2018. 2

[58] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Long Beach, CA, 2019. 2

[59] Bruce Walter, Stephen Marschner, Hongsong Li, and Kenneth Torrance. Microfacet models for refraction through rough surfaces. pages 195–206, 2007. 5

[60] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20333–20342, 2022. 6

[61] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 6

[62] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[63] S Yamaguchi, S Saito, et al. *High-fidelity Facial Reflectance and Geometry Inference from an Unconstrained Image*. ACM TOG, 2018. 2

[64] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2315–2324, 2019. 2

[65] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*, 2017. 2

[66] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision*, pages 250–269. Springer, 2022. 1, 7

[67] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer graphics forum*, pages 523–550. Wiley Online Library, 2018. 2