AvatarMe: Realistically Renderable 3D Facial Reconstruction "in-the-wild"

Alexandros Lattas[1,2] Stylianos Moschoglou[1,2] Baris Gecer[1,2] Stylianos Ploumpis[1,2]
Vasileios Triantafyllou[2] Abhijeet Ghosh[1] Stefanos Zafeiriou[1,2]

[1]Imperial College London, UK [2]FaceSoft.io

[1]{a.lattas,s.moschoglou,b.gecer,s.ploumpis,ghosh,s.zafeiriou}@imperial.ac.uk [2]v.triantafyllou@facesoft.io

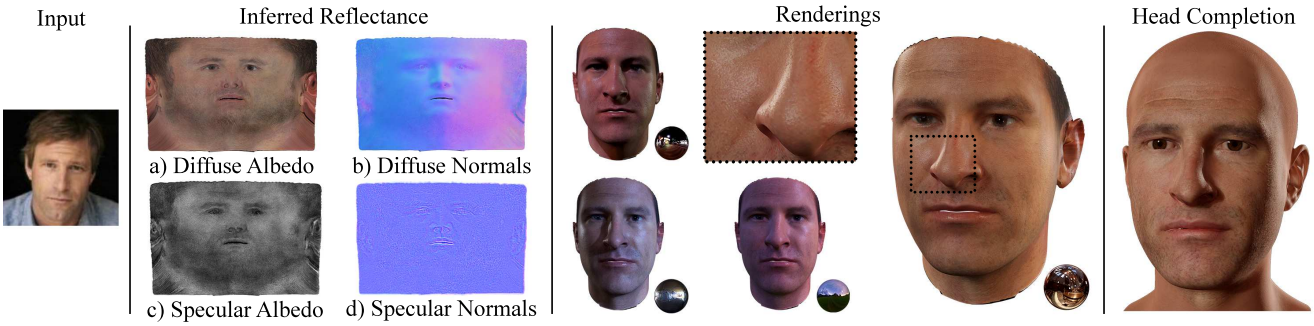


Figure 1: From left to right: Input image; Predicted reflectance (diffuse albedo, diffuse normals, specular albedo and specular normals); Rendered reconstruction in different environments, with detailed reflections; Rendered result with head completion.

Abstract

Over the last years, with the advent of Generative Adversarial Networks (GANs), many face analysis tasks have accomplished astounding performance, with applications including, but not limited to, face generation and 3D face reconstruction from a single "in-the-wild" image. Nevertheless, to the best of our knowledge, there is no method which can produce high-resolution photorealistic 3D faces from "in-the-wild" images and this can be attributed to the: (a) scarcity of available data for training, and (b) lack of robust methodologies that can successfully be applied on very high-resolution data. In this paper, we introduce AvatarMe, the first method that is able to reconstruct photorealistic 3D faces from a single "in-the-wild" image with an increasing level of detail. To achieve this, we capture a large dataset of facial shape and reflectance and build on a state-of-the-art 3D texture and shape reconstruction method and successively refine its results, while generating the per-pixel diffuse and specular components that are required for realistic rendering. As we demonstrate in a series of qualitative and quantitative experiments, AvatarMe outperforms the existing arts by a significant margin and reconstructs authentic, 4K by 6K-resolution 3D faces from a single low-resolution image that, for the first time, bridges the uncanny valley.

1. Introduction

The reconstruction of a 3D face geometry and texture is one of the most popular and well-studied fields in the intersection of computer vision, graphics and machine learning. Apart from its countless applications, it demonstrates the power of recent developments in scanning, learning and synthesizing 3D objects [3, 44]. Recently, mainly due to the advent of deep learning, tremendous progress has been made in the reconstruction of a smooth 3D face geometry, even from images captured in arbitrary recording conditions (also referred to as "in-the-wild") [13, 14, 33, 36, 37]. Nevertheless, even though the geometry can be inferred somewhat accurately, in order to render a reconstructed face in arbitrary virtual environments, much more information than a 3D smooth geometry is required, i.e., skin reflectance as well as high-frequency normals. In this paper, we propose a meticulously designed pipeline for the reconstruction of high-resolution render-ready faces from "in-the-wild" images captured in arbitrary poses, lighting conditions and occlusions. A result from our pipeline is showcased in Fig. 1.

The seminal work in the field is the 3D Morphable Model (3DMM) fitting algorithm [3]. The facial texture and shape that is reconstructed by the 3DMM algorithm always lies in a space that is spanned by a linear basis which is learned by Principal Component Analysis (PCA). The linear basis,

even though remarkable in representing the basic characteristics of the reconstructed face, fails in reconstructing high-frequency details in texture and geometry. Furthermore, the PCA model fails in representing the complex structure of facial texture captured "in-the-wild". Therefore, 3DMM fitting usually fails on "in-the-wild" images. Recently, 3DMM fitting has been extended so that it uses a PCA model on robust features, i.e., Histogram of Oriented Gradients (HoGs) [8], for representing facial texture [4]. The method has shown remarkable results in reconstructing the 3D facial geometry from "in-the-wild" images. Nevertheless, it cannot reconstruct facial texture that accurately.

With the advent of deep learning, many regression methods using an encoder-decoder structure have been proposed to infer 3D geometry, reflectance and illumination [6, 14, 33, 35, 36, 37, 39, 44]. Some of the methods demonstrate that it is possible to reconstruct shape and texture, even in real-time on a CPU [44]. Nevertheless, due to various factors, such as the use of basic reflectance models (e.g., the Lambertian reflectance model), the use of synthetic data or mesh-convolutions on colored meshes, the methods [33, 35, 36, 37, 39, 44] fail to reconstruct highly-detailed texture and shape that is render-ready. Furthermore, in many of the above methods the reconstructed texture and shape lose many of the identity characteristics of the original image.

Arguably, the first generic method that demonstrated that it is possible to reconstruct high-quality texture and shape from single "in-the-wild" images is the recently proposed GANFIT method [14]. GANFIT can be described as an extension of the original 3DMM fitting strategy but with the following differences: (a) instead of a PCA texture model, it uses a Generative Adversarial Network (GAN) [23] trained on large-scale high-resolution UV-maps, and (b) in order to preserve the identity in the reconstructed texture and shape, it uses features from a state-of-the-art face recognition network [11]. However, the reconstructed texture and shape is not render-ready due to (a) the texture containing baked illumination, and (b) not being able to reconstruct high-frequency normals or specular reflectance.

Early attempts to infer photorealistic render-ready information from single "in-the-wild" images have been made in the line of research of [6, 20, 32, 42]. Arguably, some of the results showcased in the above noted papers are of high-quality. Nevertheless, the methods do not generalize since: (a) they directly manipulate and augment the low-quality and potentially occluded input facial texture, instead of reconstructing it, and as a result, the quality of the final reconstruction always depends on the input image. (b) the employed 3D model is not very representative, and (c) a very small number of subjects (e.g., 25 [42]) were available for training for the high-frequency details of the face. Thus, while closest to our work, these approaches focus on easily creating a digital avatar rather than high-quality render-ready face reconstruction from "in-the-wild" images which is the goal of our work.

In this paper, we propose the first, to the best of our knowledge, methodology that produces high-quality render-ready face reconstructions from arbitrary images. In particular, our method builds upon recent reconstruction methods (e.g., GANFIT [14]) and contrary to [6, 42] does not apply algorithms for high-frequency estimation to the original input, which could be of very low quality, but to a GAN-generated high-quality texture. Using a light stage, we have collected a large scale dataset with samples of over 200 subjects' reflectance and geometry and we train image translation networks that can perform estimation of (a) diffuse and specular albedo, and (b) diffuse and specular normals. We demonstrate that it is possible to produce render-ready faces from arbitrary faces (pose, occlusion, etc.) including portraits and face sketches, which can be realistically relighted in any environment.

2. Related Work

2.1. Facial Geometry and Reflectance Capture

Debevec et al. [9] first proposed employing a specialized light stage setup to acquire a reflectance field of a human face for photo-realistic image-based relighting applications. They also employed the acquired data to estimate a few view-dependent reflectance maps for rendering. Weyrich et al. [41] employed an LED sphere and 16 cameras to densely record facial reflectance and computed view-independent estimates of facial reflectance from the acquired data including per-pixel diffuse and specular albedos, and per-region specular roughness parameters. These initial works employed dense capture of facial reflectance which is somewhat cumbersome and impractical.

Ma et al. [27] introduced polarized spherical gradient illumination (using an LED sphere) for efficient acquisition of separated diffuse and specular albedos and photometric normals of a face using just eight photographs, and demonstrated high quality facial geometry, including skin mesostructure as well as realistic rendering with the acquired data. It was however restricted to a frontal viewpoint of acquisition due to their employment of view-dependent polarization pattern on the LED sphere. Subsequently, Ghosh et al. [15] extended polarized spherical gradient illumination for multi-view facial acquisition by employing two orthogonal spherical polarization patterns. Their method allows capture of separated diffuse and specular reflectance and photometric normals from any viewpoint around the equator of the LED sphere and can be considered the state-of-the art in terms of high quality facial capture.

Recently, Kampouris et al. [22] demonstrated how to employ unpolarized binary spherical gradient illumination for

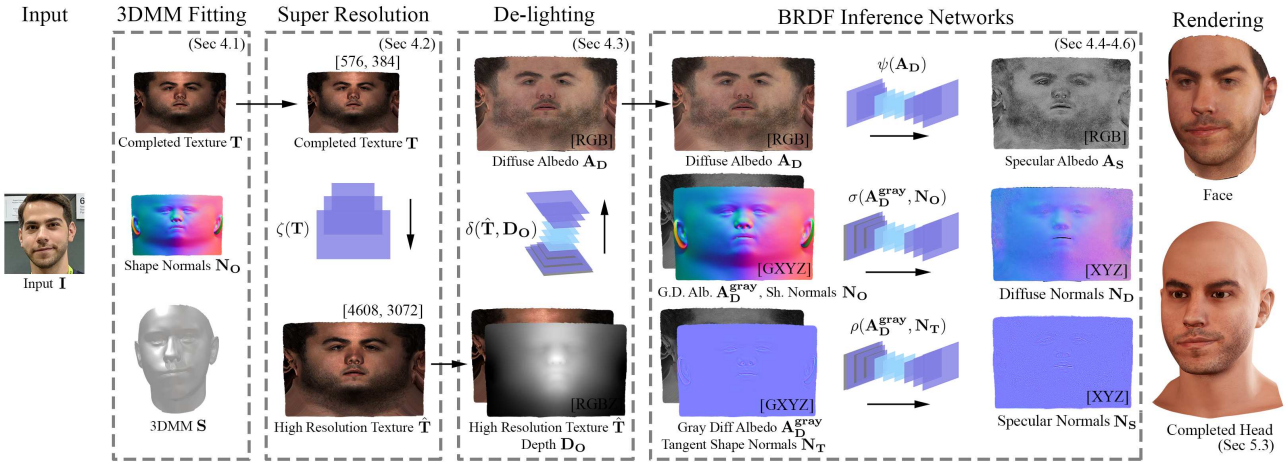


Figure 2: Overview of the proposed method. A 3DMM is fitted to an "in-the-wild" input image and a completed UV texture is synthesized, while optimizing for the identity match between the rendering and the input. The texture is up-sampled 8 times, to synthesize plausible high-frequency details. We then use an image translation network to de-light the texture and obtain the diffuse albedo with high-frequency details. Then, separate networks infer the specular albedo, diffuse normals and specular normals (in tangent space) from the diffuse albedo and the 3DMM shape normals. The networks are trained on 512×512 patches and inferences are ran on 1536×1536 patches with a sliding window. Finally, we transfer the facial shape and consistently inferred reflectance to a head model. Both face and head can be rendered realistically in any environment.

estimating separated diffuse and specular albedo and photometric normals using color-space analysis. The method has the advantage of not requiring polarization and hence requires half the number of photographs compared to polarized spherical gradients and enables completely view-independent reflectance separation, making it faster and more robust for high quality facial capture [24].

Passive multiview facial capture has also made significant progress in recent years, from high quality facial geometry capture [2] to even detailed facial appearance estimation [17]. However, the quality of the acquired data with such passive capture methods is somewhat lower compared to active illumination techniques.

In this work, we employ two state-of-the-art active illumination based multiview facial capture methods [15, 24] for acquiring high quality facial reflectance data in order to build our training data.

2.2. Image-to-Image Translation

Image-to-image translation refers to the task of translating an input image to a designated target domain (e.g., turning sketches into images, or day into night scenes). With the introduction of GANs [16], image-to-image translation improved dramatically [21, 45]. Recently, with the increasing capabilities in the hardware, image-to-image translation has also been successfully attempted in high-resolution data [40]. In this work we utilize variations of pix2pixHD [40] to carry out tasks such as de-lighting and the extraction of reflectance maps in very high-resolution.

2.3. Facial Geometry Estimation

Over the years, numerous methods have been introduced in the literature that tackle the problem of 3D facial reconstruction from a single input image. Early methods required a statistical 3DMM both for shape and appearance, usually encoded in a low dimensional space constructed by PCA [3, 4]. Lately, many approaches have tried to leverage the power of Convolutional Neural Networks (CNNs) to either regress the latent parameters of a PCA model [38, 7] or utilize a 3DMM to synthesize images and formulate an image-to-image translation problem using CNNs [18, 31].

2.4. Photorealistic 3D faces with Deep Learning

Many approaches have been successful in acquiring the reflectance of materials from a single image, using deep networks with an encoder-decoder architecture [12, 25, 26]. However, they only explore 2D surfaces and in a constrained environment, usually assuming a single point-light source.

Early applications on human faces [34, 35] used image translation networks to infer facial reflection from an "in-the-wild" image, producing low-resolution results. Recent approaches attempt to incorporate additional facial normal and displacement mappings resulting in representations with high frequency details [6]. Although this method demonstrates impressive results in geometry inference, it tends to fail in conditions with harsh illumination and extreme head poses, and does not produce re-lightable results. Saito et al. [32] proposed a deep learning approach for data-

driven inference of high resolution facial texture map of an entire face for realistic rendering, using an input of a single low-resolution face image with partial facial coverage. This has been extended to inference of facial mesostructure, given a diffuse albedo [20], and even complete facial reflectance and displacement maps besides albedo texture, given partial facial image as input [42]. While closest to our work, these approaches achieve the creation of digital avatars, rather than high quality facial appearance estimation from "in-the-wild" images. In this work, we try to overcome these limitations by employing an iterative optimization framework as proposed in [14]. This optimization strategy leverages a deep face recognition network and GANs into a conventional fitting method in order to estimate the high-quality geometry and texture with fine identity characteristics, which can then be used to produce high-quality reflectance maps.

3. Training Data

3.1. Ground Truth Acquisition

(a) Diff. Alb. (b) Spec. Alb. (c) Diff. Nor. (d) Spec. Nor.

Figure 3: Two subjects' reflectance acquired with [15] (top) and [22, 24] (bottom). Specular normals in tangent space.

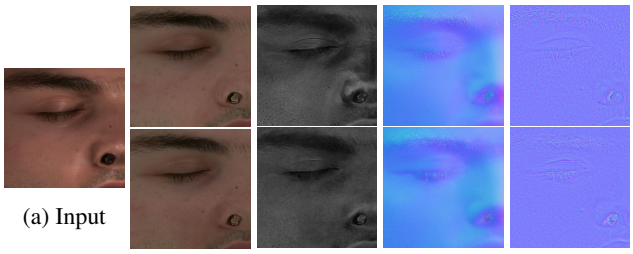
We employ the state-of-the-art method of [15] for capturing high resolution pore-level reflectance maps of faces using a polarized LED sphere with 168 lights (partitioned into two polarization banks) and 9 DSLR cameras. Half the LEDs on the sphere are vertically polarized (for parallel polarization), and the other half are horizontally polarized (for cross-polarization) in an interleaved pattern.

Using the LED sphere, we can also employ the color-space analysis from unpolarised LEDs [22] for diffuse-specular separation and the multi-view facial capture method of [24] to acquire unwrapped textures of similar quality (Fig. 3). This method requires less than half of data captured (hence reduced capture time) and a simpler setup (no polarizers), enabling the acquisition of larger datasets.

3.2. Data Collection

In this work, we capture faces of over 200 individuals of different ages and characteristics under 7 different expressions. The geometry reconstructions are registered to a standard topology, like in [5], with unwrapped textures as shown in Fig. 3. We name the dataset RealFaceDB. It is currently the largest dataset of this type and we intend to make it publicly available to the scientific community [1].

4. Method



(a) Input (b) Diff Alb (c) Spec Alb (d) Diff Nor (e) Spec Nor

Figure 4: Rendered patch ([14]-like) of a subject acquired with [15], ground truth maps (top-row) and predictions with our network given rendering as input (bottom-row).

To achieve photorealistic rendering of the human skin, we separately model the diffuse and specular albedo and normals of the desired geometry. Therefore, given a single unconstrained face image as input, we infer the facial geometry as well as the *diffuse albedo* ($\mathbf{A_D}$), *diffuse normals* ($\mathbf{N_D}$) [2], *specular albedo* ($\mathbf{A_S}$), and *specular normals* ($\mathbf{N_S}$).

As seen in Fig. 2, we first reconstruct a 3D face (base geometry with texture) from a single image at a low resolution using an existing 3DMM algorithm [5]. Then, the reconstructed texture map, which contains baked illumination, is enhanced by a super resolution network, followed by a de-lighting network to obtain a high resolution diffuse albedo $\mathbf{A_D}$. Finally, we infer the other three components ($\mathbf{A_S}, \mathbf{N_D}, \mathbf{N_S}$) from the diffuse albedo $\mathbf{A_D}$ in conjunction with the base geometry. The following sections explain these steps in detail.

4.1. Initial Geometry and Texture Estimation

Our method requires a low-resolution 3D reconstruction of a given face image $\mathbf{I}$. Therefore, we begin with the estimation of the facial shape with n vertices $\mathbf{S} \in \mathbb{R}^{n\times 3}$ and texture $\mathbf{T} \in \mathbb{R}^{576\times 384\times 3}$ by borrowing any state-of-the-art 3D face reconstruction approach (we use GANFIT [14]). Apart from the usage of deep identity features, GANFIT synthesizes realistic texture UV maps using a GAN as a statistical representation of the facial texture. We reconstruct the initial base shape and texture of the input image $\mathbf{I}$ as follows

[1] For the dataset and other materials we refer the reader to the project's page https://github.com/lattas/avatarme.

[2] The diffuse normals $\mathbf{N_D}$ are not usually used in commercial rendering systems. By inferring $\mathbf{N_D}$ we can model the reflection as in the state-of-the-art specular-diffuse separation techniques [15, 24].

and refer the reader to [14] for further details:

$$\mathbf{T}, \mathbf{S} = \mathcal{G}(\mathbf{I}) \tag{1}$$

where $\mathcal{G} : \mathbb{R}^{k \times m \times 3} \mapsto \mathbb{R}^{576 \times 384 \times 3}, \mathbb{R}^{n \times 3}$ denotes the GANFIT reconstruction method for an $\mathbb{R}^{k \times m \times 3}$ arbitrary sized image, and n number of vertices on a fixed topology.

Having acquired the prerequisites, we procedurally improve on them: from the reconstructed geometry $\mathbf{S}$, we acquire the shape normals $\mathbf{N}$ and enhance the facial texture $\mathbf{T}$ resolution, before using them to estimate the components for physically based rendering, such as the diffuse and specular diffuse and normals.

4.2. Super-resolution

Although the texture $\mathbf{T} \in \mathbb{R}^{576 \times 384 \times 3}$ from GANFIT [14] has reasonably good quality, it is *below par* compared to artist-made render-ready 3D faces. To remedy that, we employ a state-of-the-art super-resolution network, RCAN [43], to increase the resolution of the UV maps from $\mathbf{T} \in \mathbb{R}^{576 \times 384 \times 3}$ to $\hat{\mathbf{T}} \in \mathbb{R}^{4608 \times 3072 \times 3}$, which is then re-topologized and up-sampled to $\mathbb{R}^{6144 \times 4096}$. Specifically, we train a super-resolution network ($\zeta : \mathbb{R}^{48 \times 48 \times 3} \mapsto \mathbb{R}^{384 \times 384 \times 3}$) with the texture patches of the acquired low-resolution texture $\mathbf{T}$. At the test time, the whole texture from GANFIT $\mathbf{T}$ is upscaled by the following:

$$\hat{\mathbf{T}} = \zeta(\mathbf{T}) \tag{2}$$

4.3. Diffuse Albedo Extraction by De-lighting

A significant issue of the texture $\mathbf{T}$ produced by 3DMMs is that they are trained on data with baked illumination (i.e. reflection, shadows), which they reproduce. GANFIT-produced textures contain sharp highlights and shadows, made by strong point-light sources, as well as baked environment illumination, which prohibits photorealistic rendering. In order to alleviate this problem, we first model the illumination conditions of the dataset used in [14] and then synthesize UV maps with the same illumination in order to train an image-to-image translation network from texture with baked-illumination to unlit diffuse albedo $\mathbf{A_D}$. Further details are explained in the following sections.

4.3.1 Simulating Baked Illumination

Firstly, we acquire random texture and mesh outputs from GANFIT. Using a cornea model [28], we estimate the average direction of the apparent 3 point light sources used, with respect to the subject, and an environment map for the textures $\mathbf{T}$. The environment map produces a good estimation of the environment illumination of GANFIT's data while the 3 light sources help to simulate the highlights and shadows. Thus, we render our acquired 200 subjects (Section 3), as if they were samples from the dataset used in the training of [14], while also having accurate ground truth of their albedo and normals. We compute a physically-based rendering for each subject from all view-points, using the predicted environment map and the predicted light sources with a random variation of their position, creating an illuminated texture map. We denote this whole simulation process by $\xi : \mathbf{A_D} \in \mathbb{R}^{6144 \times 4096 \times 3} \mapsto \mathbf{A_D^T} \in \mathbb{R}^{6144 \times 4096 \times 3}$ which translates diffuse albedo to the distribution of the textures with baked illumination, as shown in the following:

$$\mathbf{A_D^T} = \xi(\mathbf{A_D}) \sim \mathbb{E}_{\mathbf{t} \in \{\mathbf{T_1}, \mathbf{T_2}, \ldots, \mathbf{T_n}\}} \mathbf{t} \tag{3}$$

4.3.2 Training the De-lighting Network

Given the simulated illumination as explained in Sec. 4.3.1, we now have access to a version of RealFaceDB with the [14]-like illumination $\mathbf{A_D^T}$ and with the corresponding diffuse albedo $\mathbf{A_D}$. We formulate de-lighting as a domain adaptation problem and train an image-to-image translation network. To do this, we follow two strategies different from the standard image translation approaches.

Firstly, we find that the occlusion of illumination on the skin surface is geometry-dependent and thus the resulting albedo improves in quality when feeding the network with both the texture and geometry of the 3DMM. To do so, we simply normalize the texture $\mathbf{A_D^T}$ channels to $[-1, 1]$ and concatenate them with the depth of the mesh in object space $\mathbf{D_O}$, also in $[-1, 1]$. The depth ($\mathbf{D_O}$) is defined as the Z dimension of the vertices of the acquired and aligned geometries, in a UV map. We feed the network with a 4D tensor of $[\mathbf{A^T_{D_R}}, \mathbf{A^T_{D_G}}, \mathbf{A^T_{D_B}}, \mathbf{D_O}]$ and predict the resulting 3-channel albedo $[\mathbf{A_{D_R}}, \mathbf{A_{D_G}}, \mathbf{A_{D_B}}]$. Alternatively, we can also use as an input the texture $\mathbf{A_D^T}$ concatenated with the normals in object space ($\mathbf{N_O}$). We found that feeding the network only with the texture map causes artifacts in the inference. Secondly, we split the original high resolution data into overlapping patches of 512×512 pixels in order to augment the number of data samples and avoid overfitting.

In order to remove existing illumination from $\hat{\mathbf{T}}$, we train an image-to-image translation network with patches $\delta : \mathbf{A_D^T}, \mathbf{D_O} \mapsto \mathbf{A_D} \in \mathbb{R}^{512 \times 512 \times 3}$ and then extract the diffuse albedo $\mathbf{A_D}$ by the following:

$$\mathbf{A_D} = \delta(\hat{\mathbf{T}}, \mathbf{D_O}) \tag{4}$$

4.4. Specular Albedo Extraction

Background: Predicting the entire specular BRDF and the per-pixel specular roughness from the illuminated texture $\hat{\mathbf{T}}$ or the inferred diffuse albedo $\mathbf{A_D}$, poses an unnecessary challenge. As shown in [15, 22] a subject can be realistically rendered using only the intensity of the specular reflection $\mathbf{A_S}$, which is consistent on a face due to the skin's refractive index. The spatial variation is correlated to facial skin structures such as skin pores, wrinkles or hair, which act as reflection occlusions reducing the specular intensity.

Methodology: In principle, the specular albedo can also be computed from the texture with the baked illumination, since the texture includes baked specular reflection. However, we empirically found that the specular component is strongly biased due to the environment illumination and occlusion. Having computed a high quality diffuse albedo $\mathbf{A_D}$ from the previous step, we infer the specular albedo $\mathbf{A_S}$ by a similar patch-based image-to-image translation network from the diffuse albedo ($\psi : \mathbf{A_D} \mapsto \mathbf{A_S} \in \mathbb{R}^{512\times512\times3}$) trained on RealFaceDB:

$$\mathbf{A_S} = \psi(\mathbf{A_D}) \tag{5}$$

The results (Figs. 4a, 4d) show how the network differentiates the intensity between hair and skin, while learning the high-frequency variation that occurs from the pore occlusion of specular reflection.

4.5. Specular Normals Extraction

Background: The specular normals exhibit sharp surface details, such as fine wrinkles and skin pores, and are challenging to estimate, as the appearance of some high-frequency details is dependent on the lighting conditions and viewpoint of the texture. Previous works fail to predict high-frequency details [6], or rely on separating the mid- and high-frequency information in two separate maps, as a generator network may discard the high-frequency as noise [42]. Instead, we show that it is possible to employ an image-to-image translation network with feature matching loss on a large high-resolution training dataset, which produces more detailed and accurate results.

Methodology: Similarly to the process for the specular albedo, we prefer the diffuse albedo over the reconstructed texture map $\hat{\mathbf{T}}$, as the latter includes sharp highlights that get wrongly interpreted as facial features by the network. Moreover, we found that even though the diffuse albedo is stripped from specular reflection, it contains the facial skin structures that define mid- and high-frequency details, such as pores and wrinkles. Finally, since the facial features are similarly distributed across the color channels, we found that instead of the diffuse albedo $\mathbf{A_D}$, we can use the luma-transformed (in sRGB) grayscale diffuse albedo ($\mathbf{A_D^{gray}}$).

Again, we found that the network successfully generates both the mid- and high-frequency, when it receives as input the detailed diffuse albedo $\mathbf{A_D}$ together with the lower-resolution geometry information (in this case, the shape normals). Moreover, the resulting high-frequency details are more accentuated, when using normals in tangent space ($\mathbf{N_T}$), which also serve as a better output, since most commercial applications require the normals in tangent space.

We train a translation network $\rho : \mathbf{A_D^{gray}}, \mathbf{N_T} \mapsto \mathbf{N_S}$, $\in \mathbb{R}^{512\times512\times3}$ to map the concatenation of the grayscale diffuse albedo $\mathbf{A_D^{gray}}$ and the shape normals in tangent space $\mathbf{N_T}$ to the specular normals $\mathbf{N_S}$. The specular normals are extracted by the following:

$$\mathbf{N_S} = \rho(\mathbf{A_D^{gray}}, \mathbf{N_T}) \tag{6}$$

4.6. Diffuse Normals Extraction

Background: The diffuse normals are highly correlated with the shape normals, as diffusion is scattered uniformly across the skin. Scars and wrinkles alter the distribution of the diffusion and some non-skin features such as hair that do not exhibit significant diffusion.

Methodology : Similarly to the previous section, we train a network $\sigma : \mathbf{A_D^{gray}}, \mathbf{N_O} \mapsto \mathbf{N_D} \in \mathbb{R}^{512\times512\times3}$ to map the concatenation of the grayscale diffuse albedo $\mathbf{A_D^{gray}}$ and the shape normals in object space $\mathbf{N_O}$ to the diffuse normals $\mathbf{N_D}$. The diffuse normals are extracted as:

$$\mathbf{N_D} = \sigma(\mathbf{A_D^{gray}}, \mathbf{N_O}) \tag{7}$$

Finally, the inferred normals can be used to enhance the reconstructed geometry, by refining its features and adding plausible details. We integrate over the specular normals in tangent space and produce a displacement map which can then be embossed on a subdivided base geometry.

5. Experiments

5.1. Implementation Details

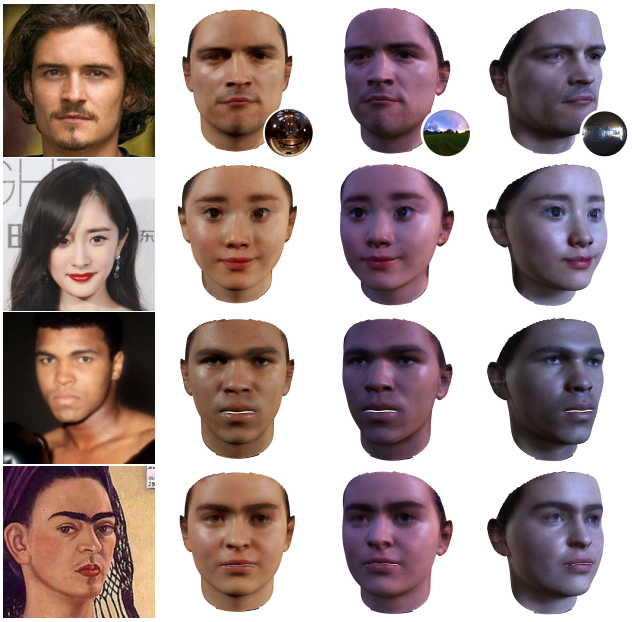
5.1.1 Patch-Based Image-to-image translation

(a) Input (b) Recon. (c) S.R. (d) Delight (e) Final

Figure 5: Rendering after (b) base reconstruction, (c) super resolution, (d) de-lighting, (e) final result.

The tasks of de-lighting, as well as inferring the diffuse and specular components from a given input image (UV) can be formulated as domain adaptation problems. As a result, to carry out the aforementioned tasks the model of our choice is pix2pixHD [40], which has shown impressive results in image-to-image translation on high-resolution data.

Nevertheless, as discussed previously: (a) our captured data are of very high-resolution (more than 4K) and thus cannot be used for training "as-is" utilizing pix2pixHD, due to hardware limitations (note not even on a 32GB GPU we can fit such high-resolution data in their original format), (b) pix2pixHD [40] takes into account only the texture information and thus geometric details, in the form of the shape



(a) Input (b) Cathedral (c) Sunset (d) Tunnel

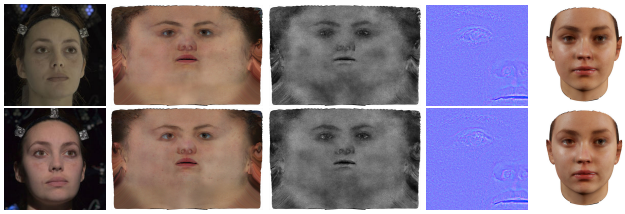
Figure 6: Reconstructions of our method re-illuminated under different environment maps [10] with added spot lights.

normals and depth cannot be exploited to improve the quality of the generated diffuse and specular components.

To alleviate the aforementioned shortcomings, we: (a) split the original high-resolution data into smaller patches of 512×512 size. More specifically, using a stride of size 256, we derive the partially overlapping patches by passing through each original UV horizontally as well as vertically, (b) for each translation task, we utilize the shape normals, concatenate them channel-wise with the corresponding grayscale texture input (e.g., in the case of translating the diffuse albedo to the specular normals, we concatenate the grayscale diffuse albedo with the shape normals channel-wise) and thus feed a 4D tensor ($[G, X, Y, Z]$) to the network. This increases the level of detail in the derived outputs as the shape normals act as a geometric "guide". Note that during inference that patch size can be larger (e.g. 1536×1536), since the network is fully-convolutional.

5.1.2 Training Setup

To train RCAN [43], we use the default hyper-parameters. For the rest of the translation of models, we use a custom translation network as described earlier, which is based on pix2pixHD [40]. More specifically, we use 9 and 3 residual blocks in the global and local generators, respectively. The learning rate we employed is 0.0001, whereas the Adam betas are 0.5 for β_1 and 0.999 for β_2. Moreover, we do not use the VGG features matching loss as this slightly deteriorated the performance. Finally, we use as inputs 3 and 4 channel tensors which include the shape normals $\mathbf{N_O}$ or depth $\mathbf{D_O}$ together with the RGB $\mathbf{A_D}$ or grayscale $\mathbf{A_D^{gray}}$ values of the



(a) Input (b) Diff Alb (c) Spec Alb (d) Norm (e) Render

Figure 7: Consistency of our algorithm on varying lighting conditions. Input images from the Digital Emily Project [1].

inputs. As mentioned earlier, this substantially improves the results by accentuating the details in the translated outputs.

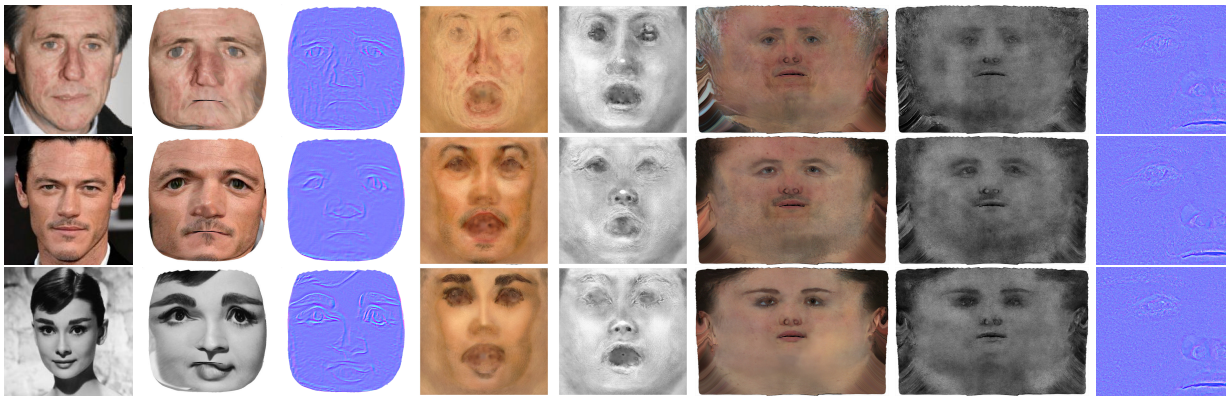
5.2. Evaluation

We conduct quantitative as well as qualitative comparisons against the state-of-the-art. For the quantitative comparisons, we utilize the widely used PSNR metric [19], and report the results in Table 1. As can be seen, our method outperforms [6] and [42] by a significant margin. Moreover using a state-of-the-art face recognition algorithm [11], we also find the highest match of facial identity compared to the input images when using our method. The input images were compared against renderings of the faces with reconstructed geometry and reflectance, including eyes.

For the qualitative comparisons, we perform 3D reconstructions of "in-the-wild" images. As shown in Figs. 8 and 9, our method does not produce any artifacts in the final renderings and successfully handles extreme poses and occlusions such as sunglasses. We infer the texture maps in a patch-based manner from high-resolution input, which produces higher-quality details than [6, 42], who train on high-quality scans but infer the maps for the whole face, in lower resolution. This is also apparent in Fig. 5, which shows our reconstruction after each step of our process. Moreover, we can successfully acquire each component from black-and-white images (Fig. 9) and even drawn portraits (Fig. 8).

Furthermore, we experiment with different environment conditions, in the input images and while rendering. As presented in Fig. 7, the extracted normals, diffuse and specular albedos are consistent, regardless of the illumination on the original input images. Finally, Fig. 6 shows different subjects rendered under different environments. We can realistically illuminate each subject in each scene and accurately reconstruct the environment reflectance, including detailed specular reflections and subsurface scattering.

In addition to the facial mesh, we are able to infer the entire head topology based on the Universal Head Model (UHM) [29, 30]. We project our facial mesh to a subspace, regress the head latent parameters and then finally derive the completed head model with completed textures. Some qualitative head completion results can be seen in Figs 1, 2.



(a) Input (b) Tex. [6] (c) Nor. [6] (d) Alb. [42] (e) S.A. [42] (f) Ours D.A. (g) Ours S.A. (h) Ours S.N.

Figure 8: Comparison of reflectance maps predicted by our method against state-of-the-art methods. [42] reconstruction is provided by the authors and [6] from their open-sourced models. Last column is cropped to better show the details.

(a) Input (b) [42] (c) [6] (d) Ours

Figure 9: Qualitative comparison of reconstructions of a subject from "in-the-wild" images, rendered in the Grace Cathedral environment [10]. [42] reconstructions provided by the authors and [6] from their open-sourced models.

Algorithm	[42]	[6]	Ours
PSNR (Albedo)	11.225	14.374	**24.05**
PSNR (Normals)	21.889	17.321	**26.97**
Rendered ID Match [11]	0.632	0.629	**0.873**

Table 1: Average PSNR computed for a single subject between 6 reconstructions of the same subject from "in-the-wild" images and the ground truth captures with [24]. We transform [6, 42] results to our UV topology and compute only for a $2K \times 2K$ centered crop, as they only produced the frontal part of the face and manually add eyes to [42].

5.3. Limitations

While our dataset contains a relatively large number of subjects, it does not contain sufficient examples of subjects from certain ethnicities. Hence, our method currently does not perform that well when we reconstruct faces of e.g. darker skin subjects. Also, the reconstructed specular albedo and normals exhibit slight blurring of some high frequency pore details due to minor alignment errors of the acquired data to the template 3DMM model. Finally, the accuracy of facial reconstruction is not completely independent of the quality of the input photograph, and well-lit, higher resolution photographs produce more accurate results.

6. Conclusion

In this paper, we propose the first methodology that produces high-quality rendering-ready face reconstructions from arbitrary "in-the-wild" images. We build upon recently proposed 3D face reconstruction techniques and train image translation networks that can perform estimation of high quality (a) diffuse and specular albedo, and (b) diffuse and specular normals. This is made possible with a large training dataset of 200 faces acquired with high quality facial capture techniques. We demonstrate that it is possible to produce rendering-ready faces from arbitrary face images varying in pose, occlusions, etc., including black-and-white and drawn portraits. Our results exhibit unprecedented level of detail and realism in the reconstructions, while preserving the identity of subjects in the input photographs.

Acknowledgements

AL was supported by EPSRC Project DEFORM (EP/S010203/1) and SM by an Imperial College FATA. AG acknowledges funding by the EPSRC Early Career Fellowship (EP/N006259/1) and SZ from a Google Faculty Fellowship and the EPSRC Fellowship DEFORM (EP/S010203/1).

References

[1] Oleg Alexander, Mike Rogers, William Lambeth, Jen-Yuan Chiang, Wan-Chun Ma, Chuan-Chang Wang, and Paul Debevec. The digital emily project: Achieving a photorealistic digital actor. *IEEE Computer Graphics and Applications*, 30(4):20–31, 2010. 7

[2] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. High-quality single-shot capture of facial geometry. *ACM Transactions on Graphics (TOG)*, 29(3):40:1–40:9, 2010. 3

[3] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. In *Siggraph*, volume 99, pages 187–194, 1999. 1, 3

[4] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3d face morphable models" in-the-wild". In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5464–5473. IEEE, 2017. 2, 3

[5] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5543–5552, 2016. 4

[6] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 3, 6, 7, 8

[7] Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T Freeman. Synthesizing normalized faces from facial identity features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3703–3712, 2017. 3

[8] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. 2005. 2

[9] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156. ACM Press/Addison-Wesley Publishing Co., 2000. 2

[10] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156. ACM Press/Addison-Wesley Publishing Co., 2000. 7, 8

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 2, 7, 8

[12] Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau. Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics (ToG)*, 37(4):1–15, 2018. 3

[13] Baris Gecer, Alexander Lattas, Stylianos Ploumpis, Jiankang Deng, Athanasios Papaioannou, Stylianos Moschoglou, and Stefanos Zafeiriou. Synthesizing coupled 3d face modalities by trunk-branch generative adversarial networks. *arXiv preprint arXiv:1909.02215*, 2019. 1

[14] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1155–1164, 2019. 1, 2, 4, 5

[15] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. In *ACM Transactions on Graphics (TOG)*, volume 30, page 129. ACM, 2011. 2, 3, 4, 5

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3

[17] Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Practical dynamic facial appearance modeling and acquisition. *ACM Trans. Graph.*, 37(6), Dec. 2018. 3

[18] Yudong Guo, Jianfei Cai, Boyi Jiang, Jianmin Zheng, et al. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1294–1307, 2018. 3

[19] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369. IEEE, 2010. 7

[20] Loc Huynh, Weikai Chen, Shunsuke Saito, Jun Xing, Koki Nagano, Andrew Jones, Paul Debevec, and Hao Li. Mesoscopic facial geometry inference using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8407–8416, 2018. 2, 4

[21] 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. 3

[22] Christos Kampouris, Stefanos Zafeiriou, and Abhijeet Ghosh. Diffuse-specular separation using binary spherical gradient illumination. In *EGSR (EI&I)*, pages 1–10, 2018. 2, 4, 5

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2

[24] Alexander Lattas, Mingqian Wang, Stefanos Zafeiriou, and Abhijeet Ghosh. Multi-view facial capture using binary spherical gradient illumination. In *ACM SIGGRAPH 2019 Posters*, page 59. ACM, 2019. 3, 4, 8

[25] Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. *ACM Transactions on Graphics (TOG)*, 36(4):1–11, 2017. 3

[26] Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. Materials for masses: Svbrdf acquisition with a single mobile phone image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 72–87, 2018. 3

[27] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, and Paul Debevec. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 183–194. Eurographics Association, 2007. 2

[28] Ko Nishino and Shree K Nayar. Eyes for relighting. *ACM Transactions on Graphics (TOG)*, 23(3):704–711, 2004. 5

[29] Stylianos Ploumpis, Evangelos Ververas, Eimear O' Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William AP Smith, Baris Gecer, and Stefanos Zafeiriou. Towards a complete 3d morphable model of the human head. *arXiv preprint arXiv:1911.08008*, 2019. 7

[30] Stylianos Ploumpis, Haoyang Wang, Nick Pears, William AP Smith, and Stefanos Zafeiriou. Combining 3d morphable models: A large scale face-and-head model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10934–10943, 2019. 7

[31] Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 460–469, 2016. 3

[32] Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. Photorealistic facial texture inference using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5144–5153, 2017. 2, 3

[33] 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. 1, 2

[34] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6296–6305, 2018. 3

[35] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5541–5550, 2017. 2, 3

[36] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2549–2559, 2018. 1, 2

[37] Luan Tran and Xiaoming Liu. On learning 3d face morphable model from in-the-wild images. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1, 2

[38] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5163–5172, 2017. 3

[39] Mengjiao Wang, Zhixin Shu, Shiyang Cheng, Yannis Panagakis, Dimitris Samaras, and Stefanos Zafeiriou. An adversarial neuro-tensorial approach for learning disentangled representations. *International Journal of Computer Vision*, 127(6-7):743–762, 2019. 2

[40] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 3, 6, 7

[41] Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, and Markus Gross. Analysis of human faces using a measurement-based skin reflectance model. *ACM Transactions on Graphics (TOG)*, 25(3):1013–1024, July 2006. 2

[42] Shuco 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):162, 2018. 2, 4, 6, 7, 8

[43] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018. 5, 7

[44] Yuxiang Zhou, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1097–1106, 2019. 1, 2

[45] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3