Relightify: Relightable 3D Faces from a Single Image via Diffusion Models

Foivos Paraperas Papantoniou[1,2] Alexandros Lattas[1,2] Stylianos Moschoglou[1,2] Stefanos Zafeiriou[1,2]
[1]Imperial College London [2]Huawei Noah's Ark Lab
{f.paraperas,a.lattas,s.moschoglou,s.zafeiriou}@imperial.ac.uk

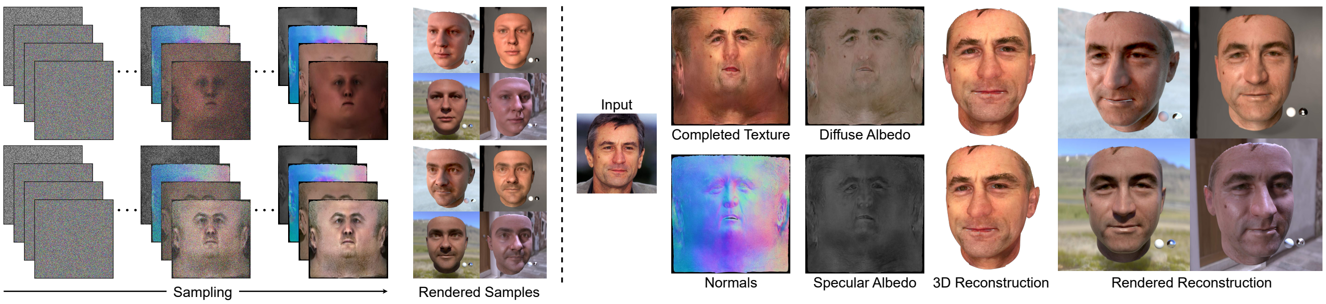

Figure 1. Left: We train an unconditional diffusion model on a high-quality dataset of UV textures and their accompanying facial reflectance maps. Right: Using this model, we perform both texture completion as well as accurate reflectance prediction from monocular images by inpainting in UV space. Our 3D facial reconstruction requires only a single image and allows the realistic rendering of the 3D avatar.

Abstract

Following the remarkable success of diffusion models on image generation, recent works have also demonstrated their impressive ability to address a number of inverse problems in an unsupervised way, by properly constraining the sampling process based on a conditioning input. Motivated by this, in this paper, we present the first approach to use diffusion models as a prior for highly accurate 3D facial BRDF reconstruction from a single image. We start by leveraging a high-quality UV dataset of facial reflectance (diffuse and specular albedo and normals), which we render under varying illumination settings to simulate natural RGB textures and, then, train an unconditional diffusion model on concatenated pairs of rendered textures and reflectance components. At test time, we fit a 3D morphable model to the given image and unwrap the face in a partial UV texture. By sampling from the diffusion model, while retaining the observed texture part intact, the model inpaints not only the self-occluded areas but also the unknown reflectance components, in a single sequence of denoising steps. In contrast to existing methods, we directly acquire the observed texture from the input image, thus, resulting in more faithful and consistent reflectance estimation. Through a series of qualitative and quantitative comparisons, we demonstrate superior performance in both texture completion as well as reflectance reconstruction tasks.

1. Introduction

Creating digital avatars of real people is of paramount importance for a range of applications, including VR, AR or the film industry. Human faces have been studied extensively over the years, attracting attention at the intersection of Computer Vision, Graphics and Machine Learning research. Although vast literature exists around the estimation of the 3D shape and reflectance of a face from unconstrained inputs such as "in-the-wild" RGB images, it still remains a challenging problem in the field. In particular, the recent breakthrough in image synthesis using diffusion generative models creates a new perspective towards photo-realistic 3D face reconstruction, which has not been explored so far and stems from the state-of-the-art performance of these models in solving inverse problems without supervised training.

Facial reflectance capture typically requires a controllable illumination system equipped with multiple cameras, first introduced as a Light Stage [12]. Polarized illumination and gradient patterns can be employed for diffuse-specular separation [48, 26], using which, spatially varying facial reflectance maps can be acquired, that describe BRDF parameters, including the diffuse and specular albedo and normals. Although recent works attempt to simplify the capturing apparatus and process using inverse rendering [28, 55] or commodity devices [38], such methods still require a laborious capturing process and expensive equipment.

Since their introduction by Blanz and Vetter [3], 3D Morphable Models (3DMMs) [54, 11, 44, 5, 4] have been established as a robust methodology for monocular 3D face reconstruction [18, 69] by regularizing the otherwise ill-posed optimization problem towards a known statistical prior of the facial geometry, which is usually defined by the linear space of a PCA model. In addition to the coarse geometry estimation, 3DMMs have been used in conjunction with powerful CNN-based texture models, leading to impressively detailed avatar reconstructions even from low-resolution images [57, 23, 24]. Furthermore, another line of research [6, 32, 68, 2, 17, 16, 39, 41] revolves around the reconstruction of rendering assets such as reflectance components (diffuse and specular albedo) and high-frequency normals of the facial surface. As a result, the recovered 3D faces can be realistically rendered in arbitrary illumination environments. However, prior work either contains scene illumination inhibiting relighting [13, 21, 23] or is restricted by the models' generalization, lowering the identity similarity [23, 39, 47]. Our work shares the same objective in that we couple a 3DMM with high-quality UV reflectance maps, but attempts to solve both of these issues, by preserving the observed texture details from the input image and jointly inferring the facial reflectance.

In fact, the visible pixels of the facial texture by the given camera pose are directly recoverable from the input image via inverse rasterization of the fitted 3D mesh. Therefore, we cast the 3D face reconstruction problem as an image inpainting task in the UV space; *i.e*. the goal is to fill in the missing pixels in a consistent manner with respect to some statistical prior. In particular, we propose to use a diffusion model as the generative backbone of our method. Diffusion models [61] are naturally associated with guided image synthesis since they treat image generation as a sequence of denoising steps in the form of a learnable Markov process. This allows to directly interfere with the sampling process, given that samples at each part of the chain are distorted versions of real images with known noise variances. Thus, by properly modifying the sampling process, a single unconditional diffusion model can be used for different inverse problems, such as image editing [50], inpainting [46, 10], restoration [36] or super-resolution [9, 8], without problem-specific training.

In this paper, we build a high-quality statistical model of facial texture and reflectance by means of a diffusion model and adopt an inpainting approach to complete the partially reconstructed UV texture produced by a 3DMM fitting step. We further extend the sampling process to recover the missing reflectance components by enforcing consistency with the input texture. As a result, our method, dubbed *Relightify*, generates accurate and render-ready 3D faces from unconstrained images, as shown in Fig. 1.

In summary, we make the following contributions:

- We present the first, to the best of our knowledge, diffusion-based approach for relightable 3D face reconstruction from images. By training on a pseudo-ground-truth dataset of facial reflectance, while directly recovering texture parts from the input, we achieve high-quality rendering assets that preserve important details of the input face (*e.g*. wrinkles, moles).
- We propose an efficient way of predicting different modalities in a consistent way by learning a generative model on concatenated reflectance maps and casting the reconstruction as an inpainting problem, spatially, but also channel-wise.
- We qualitatively and quantitatively demonstrate the superiority of our approach against previous methods regarding both the completed textures as well as the recovered reflectance maps.

2. Related Work

2.1. Diffusion Models for Inverse Problems

Diffusion models [61] are latent variable generative models which artificially corrupt the data distribution by adding noise and attempt to approximate the reverse process. They have lately emerged as a powerful image synthesis model [30, 15, 63] outperforming previous state-of-the-art approaches in both conditional and unconditional tasks. While they achieve excellent image quality and are robust to multi-modal distributions, they are computationally demanding to sample from, since they require a large sequence of denoising steps (*e.g*. 1000), each of which operates in the high dimensional image space. To alleviate this, a number of works [62, 37, 58] have proposed alternative strategies to accelerate sampling by reducing the steps of the reverse process. Another line of research [67, 59] proposes to train an encoding model and learn a diffusion model on its lower-dimensional latent space. Recently, Rombach *et al*. [56] have further explored the use of a VQGAN [19] as the auto-encoding model, showing that a mild compression is enough to reduce the training/sampling time without sacrificing sample quality. The latter approach is our method of choice for this work, as we elaborate on a high-resolution UV image space, which would otherwise significantly increase the computational overhead.

One of the most interesting aspects of diffusion models is that they can be used as unsupervised solvers for different inverse problems, where the goal is to reconstruct a sample from some distorted observation, *i.e*. conditioning input. Song *et al*. [63] propose a conditioning mechanism during inference that allows applications such as class-conditional generation, inpainting and colorization. Similarly, [8] uses a low-pass filtered version of the conditioning image to guide the denoising process at each step and SDEdit [50] addresses image translation and editing using a diffused version of the input image to initialize sampling from an intermediate timestep. RePaint [46] achieves state-of-the-art results on image inpainting by repeating multiple forward and backward diffusion steps to enforce harmonization. Despite its improved performance, this resampling strategy significantly increases the computational time. In contrast, CCDF [9] and DDRM [36] propose efficient techniques for reducing the length of the reverse process while retaining image quality at a high level. More recently, MCG [10] introduced a novel manifold constraint step, which combined with the standard reverse diffusion outperforms the aforementioned methods on a number of inverse tasks, including inpainting. We adopt this approach in our work to accurately fill in the missing pixels of both texture and reflectance maps of a face from a given image via diffusion-based inpainting, while fully preserving the observed ones. Note also that this approach does not assume any specific distribution of visibility masks, as it is trained unconditionally on complete textures.

2.2. Facial Reconstruction

3DMMs [3] are the typical models for facial reconstruction from "in-the-wild" images, using a linear model for the identity, and additional linear models for expression or color. Current facial 3DMMs include the Basel Face Model (BFM) [54] and the Large Scale Facial Model (LSFM) [4]. Egger *et al*. [18] provide a thorough review on the subject. AlbedoMM [60] first created a 3DMM of facial reflectance, which can be relighted, but is restricted to a linear and per-vertex color model. Dib *et al*. [16, 17] improved on prior works' simplistic shading models and used inverse ray tracing to acquire photorealistic facial reflectance. Recently, GANFit [23, 24] introduced a potent method for fitting 3DMMs with a GAN-based [27] facial texture generator, achieving high-fidelity facial avatars, but lacking relighting capabilities due to baked illumination in the textures. AvatarMe++ [39, 41] overcame this issue by translating the reconstructed textures to facial reflectance using a conditional GAN, while adding extra processing steps. While we use AvatarMe++ to augment our training data, our method significantly outperforms them by using a powerful diffusion model and inferring only the occluded facial texture.

TBGAN [22] first introduced a deep generative network for facial reflectance, based on ProgressiveGAN [33] and [43] introduced a more powerful model, based on StyleGAN [34]. However, both works did not showcase fitting capabilities. An extension of the latter [47], introduced a set of multiple networks, with a StyleGAN2 [35] base, that can be used to generate shape and albedo from images with arbitrary illumination and expression. While close to our work, our method uses a single and more powerful diffusion model, inferring not only the diffuse albedo, but also the specular albedo and normals. Moreover, our work inpaints only the occluded facial areas, preserving the visible part of the texture and achieves higher reconstruction fidelity.

Although our method is applied to facial reconstruction, we simultaneously solve a facial texture inpainting problem in UV space. Initially explored in 2D facial images [45] and expanded to UV completion using deep encoder-decoder architectures (UV-GAN [13]), such works recover the facial texture from partial and masked facial images. Recently, OSTeC [21], used a pre-trained StyleGAN in 2D to recover multiple poses of the input subject so as to create a complete UV facial texture. While prior works achieve impressive results, all are restricted facial textures with baked illumination. In contrast, we jointly recover the facial reflectance, making the reconstruction relightable in standard rendering engines.

3. Method

We propose a diffusion-based inpainting approach to estimate both the UV texture with existing baked illumination and the actual reflectance of a face in a single process. At the core of our approach lies an unconditional diffusion generative model trained on pairs of textures and their accompanying reflectance. This coupled texture-reflectance modeling along with the sequential denoising process of diffusion models allows us to reconstruct the reflectance from a partial texture of the input face, as shown in Fig. 2. Our method, thus, generates high-quality 3D face avatars from "in-the-wild" images, which can be realistically relighted.

In the following sections, we first analyze the training of our diffusion model, and then explain the 3D shape reconstruction and texture inpainting strategies in further detail.

3.1. Diffusion Models: Background

Given a distribution of real images x, diffusion models [61] define a forward diffusion process which gradually adds Gaussian noise to the input image in T consecutive steps. This corresponds to a fixed Markov Chain, where starting from a clean image x_0, the noisy samples x_t at each timestep t are drawn from the following distributions (with timestep-depending variances β_t) conditioned on the previous samples:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

This is equivalent to directly sampling x_t conditioned on the clean image x_0 via:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{2}$$

where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. Given large enough T, this process leads to normally distributed noise x_T. Then, the goal is to learn the reverse Markov process:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3}$$

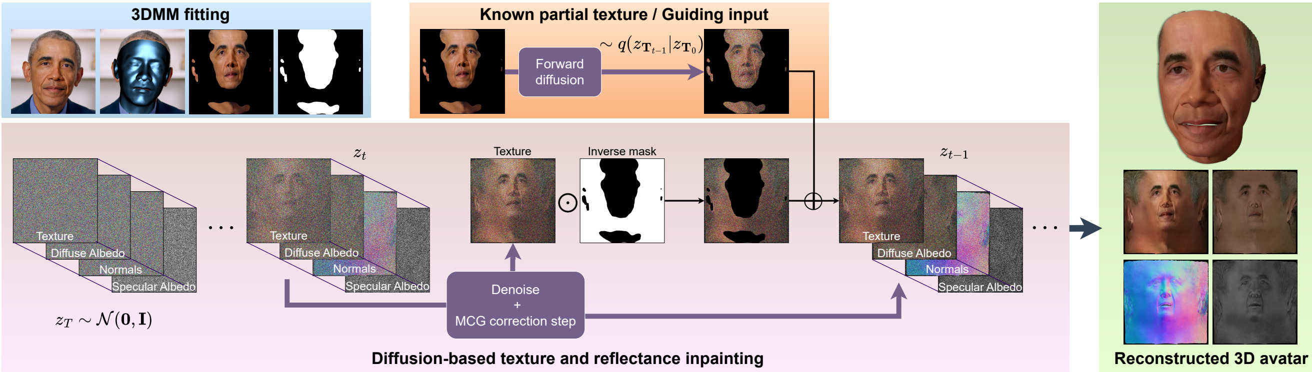

Figure 2. Overview of our method during inference. Please note that we use a latent diffusion model [56], yet we illustrate the denoising process in the original image space for visualization purposes. We perform standard 3DMM fitting to get a partial UV texture via image-to-uv rasterization. Then, starting from random noise, we utilize the known texture to guide the sampling process of a texture/reflectance diffusion model towards completing the unobserved pixels. Each denoising step, from z_t to z_{t-1} ($t \in \{1, \ldots, T\}$), follows an inpainting approach similar to MCG [10] (see Eq. 9): 1) The reflectance maps and unobserved texture pixels are updated based on reverse diffusion sampling and manifold constraints, while 2) the known pixels are directly sampled from the input texture via forward diffusion ($\odot$ and $\oplus$ denote the Hadamard product and addition respectively). Note that masking is only applied to the texture, while the reflectance maps (diffuse/specular albedo, normals) are entirely predicted from random noise. At the end of the process, we acquire high-quality rendering assets, making our 3D avatar realistically renderable.

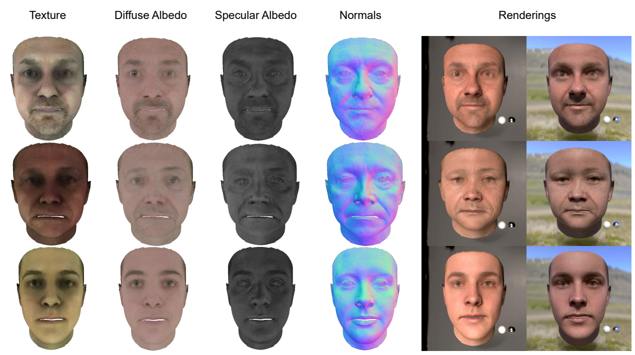

Figure 3. Samples of texture and reflectance maps generated by our diffusion model (left) and renderings in different scenes (right). Random 3DMM shapes are used for the visualization.

which gradually denoises the random noise x_T towards a realistic image, by minimizing the variational bound of the negative log likelihood [30, 15]. Following the reparameterization proposed in [30], the model consists of time-conditioned denoising autoencoders $\epsilon_\theta(x_t, t); t \in \{1, 2, \ldots, T\}$, which are trained to predict the noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ that was added to the input image x_0 to account for the noisy version x_t:

$$L = E_{x_0, \epsilon, t}\left[||\epsilon - \epsilon_\theta(x_t, t)||^2\right] \quad (4)$$

Once trained, we can generate images by starting from random noise $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and sequentially drawing denoised images around the mean:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) \quad (5)$$

3.2. Training of our Diffusion Model

In this work, we harness the power of diffusion models to learn a strong generative prior over the domain of facial texture/reflectance. In particular, we adopt a physically-based perspective by separating the facial reflectance into different UV maps, namely diffuse albedo ($\mathbf{A}_d$), specular albedo ($\mathbf{A}_s$) and surface normals ($\mathbf{N}$) with high-frequency details. This allows realistic rendering under different illumination conditions. We learn our prior using a high-quality dataset consisting of *complete* pairs of facial reflectance, and a corresponding rendered texture $\mathbf{T}$ under arbitrary illumination. More details on the data we use are provided in section 4.1. We train an unconditional diffusion model (as described in section 3.1) on the quadruples:

$$x = [\mathbf{T}, \mathbf{A}_d, \mathbf{A}_s, \mathbf{N}] \in \mathbb{R}^{512 \times 512 \times 10} \quad (6)$$

where we concatenate the components of Eq. 6 across channels (each of the 4 UV images measures 512×512 pixels and 3 channels, except for the single-channel image $\mathbf{A}_d$). By sampling from this model, we can synthesize pairs of shaded RGB textures ($\mathbf{T}$) and reflectance components ($\mathbf{A}_d, \mathbf{A}_s, \mathbf{N}$) which are in correspondence, meaning that the texture is a rendered version of the UV reflectance under some illumination environment.

In practice, to reduce the computational requirements to a reasonable level, we follow the paradigm of **latent diffusion models** proposed by Rombach *et al.* [56], where the images are first compressed to a latent space $z = \mathcal{E}(x) \in \mathbb{R}^{h \times w \times c}$ by training a perceptual auto-encoder, consisting of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. Using perceptual and adversarial losses similar to VQGAN [19], the autoen-

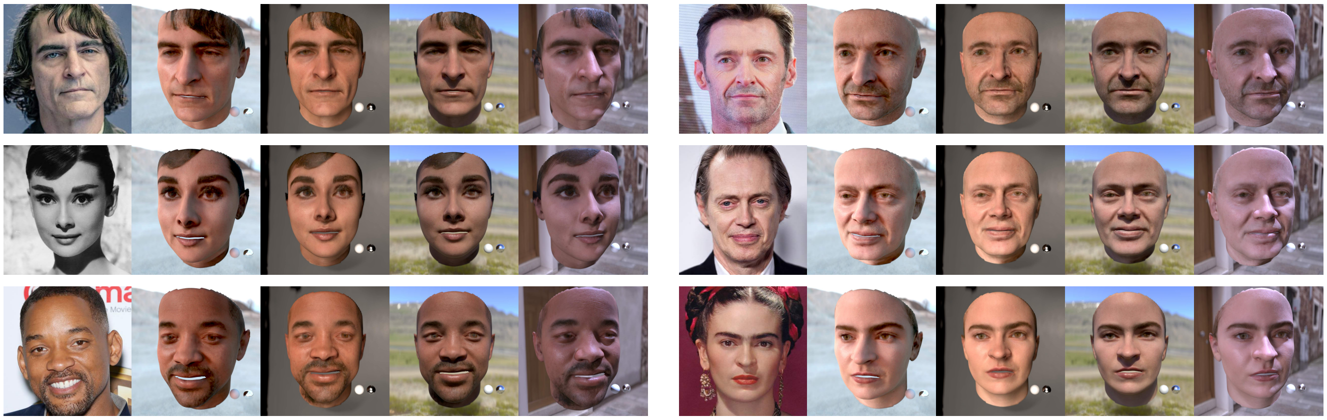

Figure 4. Examples of 3D reconstructions by our method, rendered using different environment maps in a commercial renderer [49].

coder achieves an excellent quality of reconstructed samples $\tilde{x} = \mathcal{D}(\mathcal{E}(x))$, while allowing to efficiently train the diffusion model on the lower dimensional pixel-space of the learned embeddings. In our case, we train four similar auto-encoders, one for each of $\mathbf{T}, \mathbf{A}_d, \mathbf{A}_s$ and $\mathbf{N}$, all of them reducing the input resolution to latent dimensions of $h = w = 64$, $c = 3$. Therefore, our latent diffusion model [56] is trained on the concatenation of the 4 embeddings:

$$z = [z_{\mathbf{T}}, z_{\mathbf{A}_d}, z_{\mathbf{A}_s}, z_{\mathbf{N}}] \in \mathbb{R}^{64\times64\times12} \tag{7}$$

Samples from our diffusion model (after being decoded through each $\mathcal{D}$) can be seen in Fig. 1 (left part) and Fig. 3.

3.3. Inference

We use the aforementioned trained diffusion model to perform inpainting on both the texture and reflectance UV maps based on a partial UV texture obtained by 3DMM fitting. We provide a detailed description below.

3DMM Fitting and Texture Initialization. We rely on 3DMMs to recover a rough 3D shape of the face from a 2D image as a mesh $\mathbf{S} \in \mathbb{R}^{n\times3}$ with n vertices. Specifically, we employ a linear 3DMM:

$$\mathbf{S}(\mathbf{p}_s, \mathbf{p}_e) = \mathbf{m} + \mathbf{U}_s\mathbf{p}_s + \mathbf{U}_e\mathbf{p}_e \tag{8}$$

consisting of the LSFM [4] shape eigenbasis $\mathbf{U}_s \in \mathbb{R}^{3n\times158}$ and the expression eigenbasis $\mathbf{U}_e \in \mathbb{R}^{3n\times29}$ from the 4DFAB database [7]. We fit the 3DMM to the input image by optimizing the shape coefficients $\mathbf{p}_s$, expression coefficients $\mathbf{p}_e$ and camera parameters $\mathbf{p}_c$ using an off-the-shelf framework [1]. Note that any 3DMM fitting framework works as a "plug and play" solution to our method. Thus, one may trivially use a more sophisticated algorithm (*e.g.* GANFit [23]) for precise shape reconstruction.

[1] https://github.com/ascust/3DMM-Fitting-Pytorch

We use a standard UV topology for texturing the 3D mesh, where each vertex is assigned to a fixed 2D coordinate on the UV plane. By rasterizing the fitted 3D mesh and using barycentric interpolation, we can reverse the rendering process and unfold the face in UV, hence reconstructing the visible parts of the texture directly from the input image. This initial texture is accompanied by a UV visibility mask, with 1 for pixels that are observed from the input image, and 0 for those that are occluded and, thus, need to be inpainted by our model.

Texture Completion and Reflectance Prediction. Starting from the partially completed UV texture $\mathbf{T}_0$ of the face and a binary visibility mask m produced by the previous step, our goal is to inpaint the remaining pixels along with the pixels of the 3 reflectance maps. We use the latent representation $z_{\mathbf{T}_0} = \mathcal{E}(\mathbf{T}_0) \in \mathbb{R}^{h\times w\times c}$ of this texture image to constrain the reverse diffusion process. Note that the mask m is downsampled to the same resolution $h = w = 64$ of the latent space for the next steps. Our inpainting algorithm starts with a random noise image $z_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and uses the denoising procedure of MCG [10], consisting of the following repeated steps:

$$z_{t-1}^{\text{unknown}} \sim \mathcal{N}(\mu_\theta(z_t, t), \Sigma_\theta(z_t, t)) \tag{9a}$$

$$z_{\mathbf{T}_{t-1}}^{\text{known}} \sim \mathcal{N}(\sqrt{\bar{\alpha}_{t-1}} z_{\mathbf{T}_0}, (1-\bar{\alpha}_{t-1})\mathbf{I}) \tag{9b}$$

$$\hat{z}_0 = \left(z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t, t)\right) / \sqrt{\bar{\alpha}_t} \tag{9c}$$

$$\mathcal{L} = \| (z_{\mathbf{T}_0} - \hat{z}_{\mathbf{T}_0}) \odot m\|_2^2 \tag{9d}$$

$$z_{\mathbf{T}_{t-1}} = m \odot z_{\mathbf{T}_{t-1}}^{\text{known}} + (1-m) \odot \left(z_{\mathbf{T}_{t-1}}^{\text{unknown}} - \alpha\frac{\partial\mathcal{L}}{\partial z_{\mathbf{T}_t}}\right) \tag{9e}$$

$$z_{k_{t-1}} = z_{k_{t-1}}^{\text{unknown}} - \alpha\frac{\partial\mathcal{L}}{\partial z_{k_t}}, \quad k = \{\mathbf{A}_d, \mathbf{A}_s, \mathbf{N}\} \tag{9f}$$

Given a sample z_t at timestep t, we first sample the next denoised sample z_{t-1} using the original reverse diffusion step

(Eq. 9a). We term this as z_{t-1}^{unknown} (borrowing the notation from [46]) as it does not take into account the known parts of the observed texture. To exploit the known texture, we sample a noisy version of it $z_{\mathbf{T}_{t-1}}^{\text{known}}$ at timestep $t-1$ via a forward diffusion step (Eq. 9b). Then, we directly impose this known noisy texture $m \odot z_{\mathbf{T}_{t-1}}^{\text{known}}$ ($\odot$ denotes the Hadamard product) as in the first half of Eq. 9e. Finally, for the unknown pixels, we add the manifold constraint introduced in MCG [10]; *i.e.* we make a prediction of the clean sample $\hat{z}_0$ (Eq. 9c) based on the previous timestep z_t, compare this (ℓ_2 loss) with the ground truth in the known regions (Eq. 9d), and use the gradient of this loss to update the unknown pixels of z_{t-1} (Eq. 9e and 9f) so as to minimize this distance.

Note on inpainting algorithm. We have chosen to adopt the recently proposed MCG [10] inpainting algorithm, which outperforms related state-of-the-art diffusion-based methods (*e.g.* RePaint [46], DDRM [36]), as we empirically found it to produce excellent results. Motivated by the original algorithm, which aims at inpainting standard RGB images, we expand it to account for different input domains: by treating our images as concatenated texture/reflectance maps, we force the model to perform not only spatial inpainting, but also "channel-wise inpainting", by filling the missing pixels in a manner that closely aligns with the training distribution. This essentially encourages the model to learn an inverse rendering transformation during testing, thus predicting accurate reflectance maps from just a partial illuminated version of them, despite not directly imposing physically-based constraints.

4. Experiments

4.1. Dataset and Implementation Details

We create a high-quality dataset that consists of facial textures and their corresponding reflectance. Each item includes a texture $\mathbf{T}$, shaded in some illumination, diffuse albedo $\mathbf{A}_d$, specular albedo $\mathbf{A}_s$ and normals $\mathbf{N}$. To achieve this, firstly, we acquire the public MimicMe dataset [51], which contains $\tilde{\mathbf{T}} = \{\mathbf{T}_0, \ldots, \mathbf{T}_{n_T}\}, n_T = 4,700$ diverse facial textures, whose statistics are reported in [51]. However, such textures contain the illumination of the scanning apparatus and are not relightable. Hence, we then train an image-to-image translation network based on AvatarMe++ model using the available dataset [41], which translates the textures $\tilde{\mathbf{T}}$ to facial reflectance: $\alpha(\tilde{\mathbf{T}}) \rightarrow \{\mathbf{A}_D, \mathbf{A}_S, \mathbf{N}\}$. Moreover, we augment the skin-tone diversity, using histogram matching albedo augmentation following [40]. Given the memory requirement of our network, all textures have a resolution of 512×512. Finally, to enable the diffusion model to perform well in "in-the-wild" images, we use the shapes $\mathbf{S}$ of MimicMe and the acquired reflectance, to re-render the textures under arbitrary realistic environments, directly on the UV space: $\rho(\mathbf{A}_D, \mathbf{A}_S, \mathbf{N}, \mathbf{S}) \rightarrow \mathbf{T}$. For an evaluation of the model without re-rendered textures, please refer to the Supp. Material. Although AvatarMe++ uses a similar method to augment training data, we do not require this process to be differentiable and use a ray-tracing renderer [49] (*Baker* algorithm) to achieve more realistic textures.

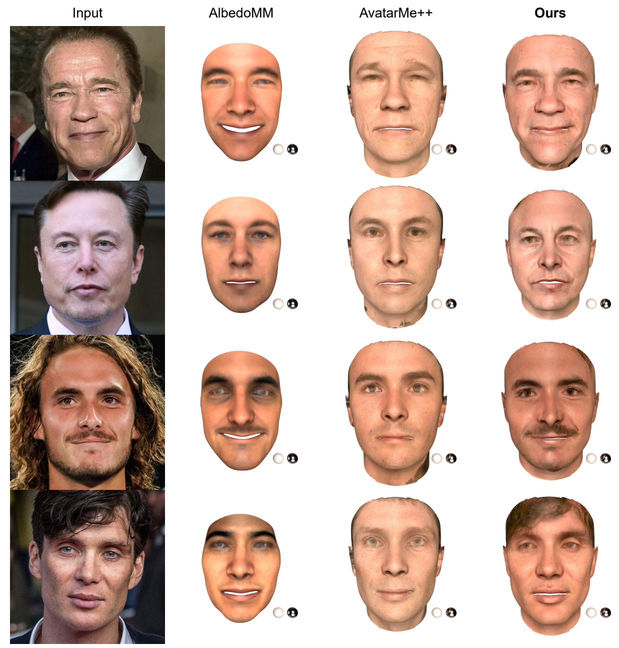

Figure 5. Rendered reconstructions of shape and reflectance by AlbedoMM [60] (using the open-source code), AvatarMe++ [41] (provided by authors) and our method, in the same illumination.

To train our model, we use a KL-regularized latent diffusion model with the default hyper-parameters proposed by the authors of [56]. Specifically, we use a downsampling factor of $f = 8$ for the perceptual auto-encoder and a diffusion length of $T = 1000$ for the denoising model. We train our model once and use it for texture and reflectance reconstruction from "in-the-wild" images. Below we provide comprehensive qualitative and quantitative evaluations.

4.2. Qualitative Results

As already described, we produce relightable 3D faces with reflectance assets that are compatible with commercial rendering engines. Fig. 4 shows examples of reconstructions from "in-the-wild" images and realistic renderings in varying environments (more results are included in the Supp. Material). Furthermore, we provide a visual comparison with the reflectance reconstruction methods of AlbedoMM [60] and AvatarMe++ [41] in Fig. 5. As can be seen, we recover 3D faces of higher consistency with respect to the input. Note that AvatarMe++ [41] starts from a GAN-generated texture as input, without di-

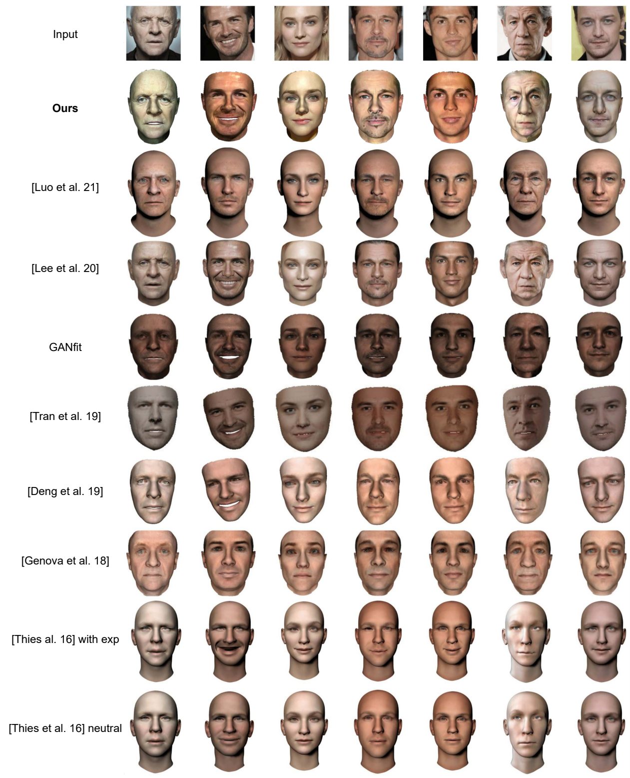

Figure 6. Visual comparison with state-of-the-art 3D face reconstruction methods [47, 42, 23, 65, 14, 25, 64]. Results for related methods are borrowed from [47].

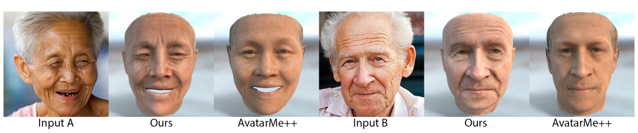

Figure 7. Comparison with AvatarMe++ [41] challenging cases.

rect feedback from the actual facial image. Despite using it to create our training data, our method clearly outperforms AvatarMe++ [41] during testing by conditioning the reflectance prediction on the genuine visible facial texture instead of a statistical approximation (fitting) of it (see Fig. 7 for some challenging subjects). We also show an extensive qualitative comparison with related 3D reconstruction methods in Fig. 6 (most of which can only recover the texture), where similar observations can be made. Finally, we test our method on images from the Digital Emily [1] and show the results in Fig. 10 together with related works [17, 41]. We yield similar results regardless of the lighting, thanks to our coupled texture/reflectance modeling that combines reflectance with randomly rendered textures during training.

4.3. Texture Completion

Following [21, 13], we evaluate our method on the task of texture completion using the Multi-PIE [29] subset of

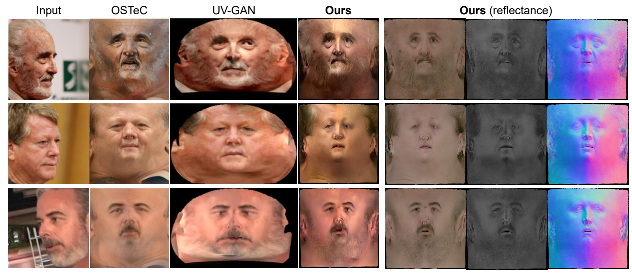

Figure 8. Examples of texture completion by OSTeC [21], UV-GAN [13] and our method from "in-the-wild" images (each method uses a different UV topology). In contrast to [21, 13] we additionally recover reflectance components for realistic rendering.

Methods	Metric	0°	±30°	±60°	±90°
CE [53]	PSNR	23.03	21.93	20.27	19.63
	SSIM	0.920	0.892	0.888	0.718
UV-GAN [13]	PSNR	23.36	22.25	20.53	19.83
	SSIM	0.924	0.897	0.892	0.725
OSTeC [21]	PSNR	23.95	22.54	21.04	20.44
	SSIM	**0.928**	0.902	0.898	0.746
Ours	PSNR	**26.00**	**24.73**	**24.65**	**20.58**
	SSIM	**0.928**	**0.916**	**0.917**	**0.874**

Table 1. Quantitative comparison between *Relightify* and [53, 13, 21] regarding UV texture completion on the MultiPIE dataset [29] for different viewing angles.

the UVDB dataset [13]. This consists of complete UV textures for 337 different identities, and corresponding 2D images of the faces from various camera poses. In accordance with [21, 13], we use the last 137 subjects for evaluation (as the first 200 were used as training data in prior works). We perform texture completion with our diffusion-based approach for each different viewing angle and compare it with existing texture completion methods, namely CE [53], UV-GAN [13] and OSTeC [21]. We use the widely adopted Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) metrics to compare the completed textures with the ground truth and report the results in Tab. 1. As can be seen, *Relightify* outperforms the related methods in almost all settings, especially for challenging angles. A visual comparison with [21, 13] is provided in Fig. 8. Note that in contrast to CE [53] and UV-GAN [13], our model was not trained on the Multi-PIE dataset.

4.4. Identity Preservation

We perform quantitative evaluations of our method's ability to preserve the subject's identity, by comparing the distribution of identity scores between the input image and rendered reconstruction, on the LFW dataset [31], against prior work [23, 24, 25, 66]. Following the existing benchmark [24], we evaluate our results using VGG-Face [52]. We present our analysis in Fig. 9, measuring the distance between the input image and reconstruction for all subjects. Our method shows a significant improvement in similarity, while also producing not just a facial texture, but a set of relightable reflectance textures.

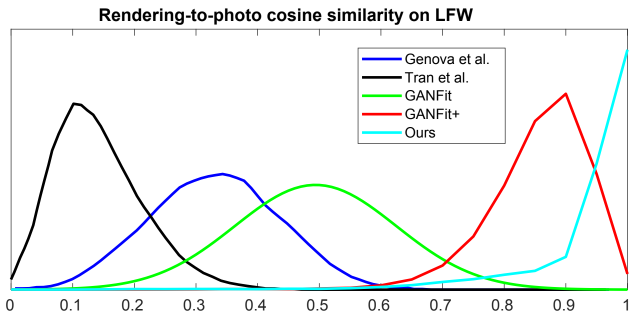

Figure 9. Quantitative evaluation of similarity scores on LFW [31], compared with prior work ([25, 66, 23, 24]), using VGG-Face [52]. We show the cosine similarity distribution between ground truth and reconstruction.

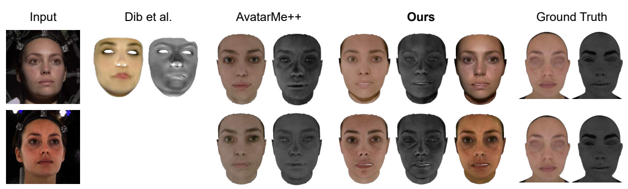

Figure 10. Reconstructions from images with different illumination (Digital Emily Project [1]) by our method as well as [17, 41] and ground truth. We show the diffuse and specular albedo for all methods (where available), plus the recovered texture for our method.

4.5. Reflectance Prediction

To further assess our method on the task of facial reflectance prediction from monocular images, we use six test subjects with captured ground truth reflectance using a Light Stage [26], and compare *Relightify* with the state-of-the-art method of AvatarMe++ [41]. More specifically, we apply both methods on 2D photos of these subjects and measure the PSNR of the recovered reflectance maps with respect to the ground truth maps. As shown in Tab. 2, our method produces significantly more accurate diffuse and specular albedos, while the normals closely match those of [41]. This demonstrates our method's ability to better capture subject-specific details by directly leveraging texture information from the input image. Note that AvatarMe++ reconstructions are additionally conditioned on the 3DMM shape normals, which may explain a slight increase in the corresponding PSNR.

4.6. Experimentation with Inpainting Algorithms

Although we adopt the MCG [10] approach for our texture/reflectance diffusion model, we have experimented

	PSNR (diffuse albedo)	PSNR (specular albedo)	PSNR (normals)
AvatarMe++ [41]	18.30	19.77	**27.26**
Ours	**22.47**	**27.17**	26.69

Table 2. Quantitative comparison of our method with [41] (results provided by authors). We calculate the average PSNR between the reconstructed and the ground truth reflectance maps for six subjects with ground truth, captured using a Light Stage [26].

with different inpainting algorithms. We compare four of them in Fig. 11 and Tab. 4. We also provide the runtime for each algorithm in Tab. 3. The baseline method of Score-SDE [63], which can be interpreted as Eq. 9 without the gradient term, produces sub-optimal results, *i.e.* the occluded areas are often inpainted in an inconsistent way with the observed ones, which is especially apparent in the texture (Fig. 11) and albedos (Tab. 4). RePaint [46] also produces unsatisfactory textures while at the same time increasing the reverse diffusion steps by a factor of n (we use $n = 10$ as suggested by the authors of [46]), which significantly affects the computational time. In contrast, MCG [10] preserves the original sampling length ($T = 1000$ timesteps), hence being much more efficient. However, it is still slower than Score-SDE [63] since it requires the computation of a gradient for the manifold constraint at each step. In general, we found MCG [10] to perform better in most cases. To further strengthen the efficiency of our method, we have additionally incorporated the DDIM [62] acceleration technique in the MCG algorithm, which allows reducing the denoising steps to $N < T$ (we use $N = 200$) without a significant drop in quality. In such case, our method can generate high-quality texture and reflectance assets from a partial UV texture in roughly 12 seconds, which is significantly faster than competing texture completion algorithms (*e.g.* OSTeC [21] requires around 10 minutes).

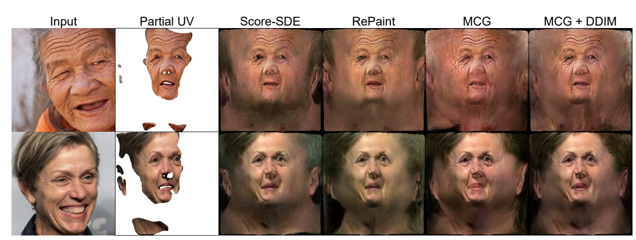

Figure 11. Texture completion with our diffusion model using different inpainting algorithms [63, 46, 10, 62]. All algorithms are implemented on top of the same unconditionally trained diffusion model, and only the reverse sampling process is modified.

5. Limitations

Our method outperforms prior works on texture completion as well as the challenging task of reflectance prediction. This is accomplished by explicitly recovering information

	Score-SDE	RePaint	MCG (Ours)	MCG (Ours) + DDIM
Time	17 sec	3 min	1 min	12 sec

Table 3. Sampling time during texture completion and reflectance prediction for different inpainting algorithms [63, 46, 10, 62] (using an Nvidia RTX 2080 TI GPU).

	Diffuse Albedo		Specular Albedo		Normals	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Score-SDE	20.80	0.808	26.69	0.845	26.86	0.784
RePaint	20.08	0.813	26.65	0.848	27.27	0.801
MCG	22.47	0.825	27.17	0.853	26.69	0.781
MCG + DDIM	21.94	0.817	26.88	0.846	26.45	0.774

Table 4. Comparison of inpainting algorithms [63, 46, 10, 62] applied on our diffusion model, following the evaluation of Tab. 2.

from the input image via inpainting. Nonetheless, similarly to related texture completion works [21, 13], this also implies that the reconstructed texture is affected by the quality of the input image. Although the partial texture is first projected in our latent diffusion space by the perceptual encoder, a low resolution input may still degrade the quality of our result. In these cases, an upsampling network could be employed as in [39] to improve the resolution and details of the predicted UV maps. Also, despite its relatively large size, the employed dataset [51] may still under-represent some ethnic groups and lack diverse facial expressions, reducing accuracy in those cases. Incorporating diverse high-quality ground truth data with captured reflectance would significantly improve the performance. Finally, our method may also suffer by the ambiguity between albedo and illumination, which is thoroughly described in TRUST [20]. In fact, their proposed solution could be combined with our method in future work.

6. Conclusion

In this paper we introduced *Relightify*, a method that achieves state-of-the-art facial texture completion and facial reflectance acquisition, from monocular "in-the-wild" images. To achieve this, we train a latent diffusion model with multiple encoder-decoder networks, on a synthetic facial texture and reflectance dataset, and use a diffusion-based inpainting method on the masked UV textures. Our results directly acquire the visible facial parts while also extrapolating to facial reflectance that exhibits a high likeness to the input image and can be trivially employed in commercial rendering applications.

Acknowledgements. A. Lattas was partly funded by the EPSRC Fellowship DEFORM (EP/S010203/1). S. Zafeiriou and part of the research was funded by the EPSRC Fellowship DEFORM (EP/S010203/1) and EPSRC Project GNOMON (EP/X011364/1).

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.

[2] Linchao Bao, Xiangkai Lin, Yajing Chen, Haoxian Zhang, Sheng Wang, Xuefei Zhe, Di Kang, Haozhi Huang, Xinwei Jiang, Jue Wang, Dong Yu, and Zhengyou Zhang. High-fidelity 3d digital human head creation from rgb-d selfies. *ACM Transactions on Graphics*, 2021.

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[4] 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 (CVPR)*, June 2016.

[5] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014.

[6] 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 (ICCV)*, pages 9429–9439, 2019.

[7] Shiyang Cheng, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. 4dfab: A large scale 4d database for facial expression analysis and biometric applications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[8] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14367–14376, October 2021.

[9] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12413–12422, June 2022.

[10] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *Advances in Neural Information Processing Systems*, 2022.

[11] Hang Dai, Nick Pears, William A. P. Smith, and Christian Duncan. A 3d morphable model of craniofacial shape and texture variation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[12] 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, 2000.

[13] 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 (CVPR)*, June 2018.

[14] 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 *IEEE Computer Vision and Pattern Recognition Workshops*, 2019.

[15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794, 2021.

[16] Abdallah Dib, Gaurav Bharaj, Junghyun Ahn, Cédric Thébault, Philippe Gosselin, Marco Romeo, and Louis Chevallier. Practical face reconstruction via differentiable ray tracing. In *Computer Graphics Forum*, volume 40, pages 153–164. Wiley Online Library, 2021.

[17] Abdallah Dib, Cedric Thebault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12819–12829, 2021.

[18] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020.

[19] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, June 2021.

[20] Haiwen Feng, Timo Bolkart, Joachim Tesch, Michael J Black, and Victoria Abrevaya. Towards racially unbiased skin tone estimation via scene disambiguation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 72–90. Springer, 2022.

[21] Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Ostec: One-shot texture completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7628–7638, June 2021.

[22] Baris Gecer, Alexandros Lattas, Stylianos Ploumpis, Jiankang Deng, Athanasios Papaioannou, Stylianos Moschoglou, and Stefanos Zafeiriou. Synthesizing coupled 3d face modalities by trunk-branch generative adversarial networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 415–433. Springer, 2020.

[23] 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 (CVPR)*, June 2019.

[24] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos P Zafeiriou. Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[25] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8377–8386, 2018.

[26] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. *ACM Transactions on Graphics (TOG)*, 30(6):1–10, 2011.

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[28] Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Practical dynamic facial appearance modeling and acquisition. *ACM Transactions on Graphics (TOG)*, 37(6):1–13, 2018.

[29] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, May 2010.

[30] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[31] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[32] 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 (CVPR)*, June 2018.

[33] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

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

[35] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020.

[36] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022.

[37] Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

[38] Alexandros Lattas, Yiming Lin, Jayanth Kannan, Ekin Ozturk, Luca Filipi, Giuseppe Claudio Guarnera, Gaurav Chawla, and Abhijeet Ghosh. Practical and scalable desktop-based high-quality facial capture. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 522–537. Springer, 2022.

[39] 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 (CVPR)*, June 2020.

[40] 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 (CVPR)*, June 2023.

[41] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos P Zafeiriou. Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[42] Gun-Hee Lee and Seong-Whan Lee. Uncertainty-aware mesh decoder for high fidelity 3d face reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[43] 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 (CVPR)*, pages 3410–3419, 2020.

[44] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017.

[45] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3911–3919, 2017.

[46] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471, June 2022.

[47] Huiwen Luo, Koki Nagano, Han-Wei Kung, Qingguo Xu, Zejian Wang, Lingyu Wei, Liwen Hu, and Hao Li. Normalized avatar synthesis using stylegan and perceptual refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11662–11672, 2021.

[48] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, Paul E Debevec, et al. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. *Rendering Techniques*, 2007(9):10, 2007.

[49] Marmoset LLC. Toolbag 4, version 4.051.

[50] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

[51] Athanasios Papaioannou, Baris Gecer, Shiyang Cheng, Grigorios Chrysos, Jiankang Deng, Eftychia Fotiadou, Christos Kampouris, Dimitrios Kollias, Stylianos Moschoglou, Kritaphat Songsri-In, et al. Mimicme: A large scale diverse 4d database for facial expression analysis. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 467–484. Springer, 2022.

[52] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12, September 2015.

[53] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[54] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009.

[55] Jérémy Riviere, Paulo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Single-shot high-quality facial geometry and skin appearance capture. *ACM Transactions on Graphics (TOG)*, 39(4):81–1, 2020.

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[57] 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 (CVPR)*, July 2017.

[58] Robin San-Roman, Eliya Nachmani, and Lior Wolf. Noise estimation for generative diffusion models. *CoRR*, abs/2104.02600, 2021.

[59] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-denoising models for few-shot conditional generation. *arXiv preprint arXiv:2106.06819*, 2021.

[60] William AP Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua B Tenenbaum, and Bernhard Egger. A morphable face albedo model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5011–5020, 2020.

[61] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 2256–2265, Lille, France, 07–09 Jul 2015.

[62] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[63] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[64] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niessner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[65] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[66] 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 (CVPR)*, pages 5163–5172, 2017.

[67] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In *Neural Information Processing Systems (NeurIPS)*, 2021.

[68] Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Trans. Graph.*, 37(4), jul 2018.

[69] 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. *Computer Graphics Forum*, 37:523–550, 2018.