Monocular Identity-Conditioned Facial Reflectance Reconstruction

Xingyu Ren Jiankang Deng* Yuhao Cheng Jia Guo Chao Ma*
Yichao Yan Wenhan Zhu Xiaokang Yang

MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
{rxy_sjtu,chengyuhao,chaoma,yanyichao,zhuwenhan823,xkyang}@sjtu.edu.cn
{jiankangdeng,guojia}@gmail.com

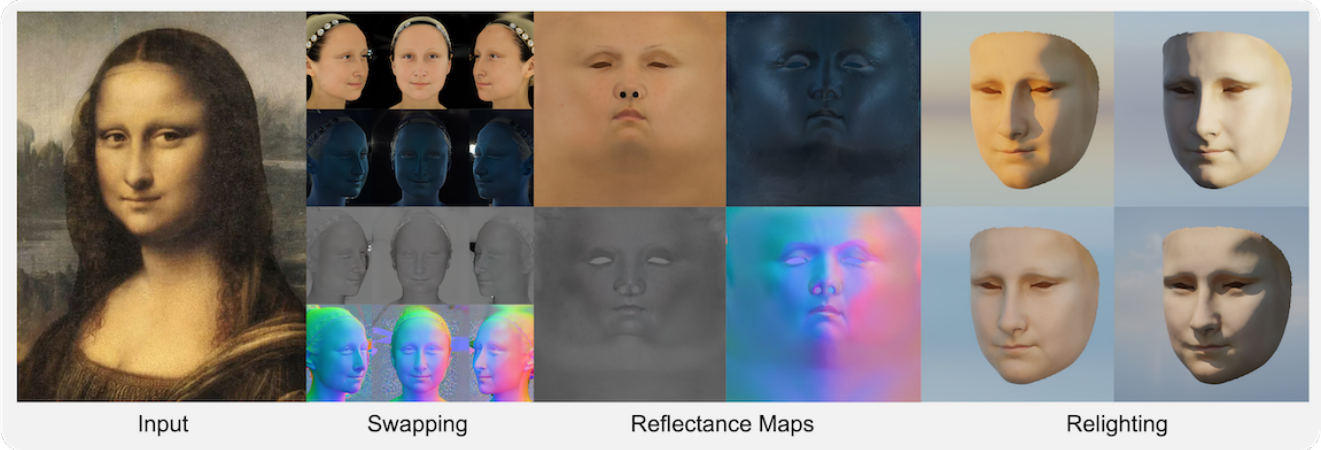

Figure 1. We present ID2Reflectance, a high-quality, identity-conditioned reflectance reconstruction method. ID2Reflectance learns multi-domain face codebooks by using limited captured data and generates multi-view domain-specific reflectance images guided by facial identity. Our approach greatly reduces the dependency on captured data and generates high-fidelity reflectance maps for realistic rendering.

Abstract

Recent 3D face reconstruction methods have made remarkable advancements, yet there remain huge challenges in monocular high-quality facial reflectance reconstruction. Existing methods rely on a large amount of light-stage captured data to learn facial reflectance models. However, the lack of subject diversity poses challenges in achieving good generalization and widespread applicability. In this paper, we learn the reflectance prior in image space rather than UV space and present a framework named ID2Reflectance. Our framework can directly estimate the reflectance maps of a single image while using limited reflectance data for training. Our key insight is that reflectance data shares facial structures with RGB faces, which enables obtaining expressive facial prior from inexpensive RGB data thus reducing the dependency on reflectance data. We first learn a high-quality prior for facial reflectance. Specifically, we pretrain multi-domain facial feature codebooks and design a codebook fusion method to align the reflectance and RGB domains. Then, we propose an identity-conditioned swapping module that injects facial identity from the target image into the pre-trained autoencoder to modify the identity of the source reflectance image. Finally, we stitch multi-view swapped reflectance images to obtain renderable assets. Extensive experiments demonstrate that our method exhibits excellent generalization capability and achieves state-of-the-art facial reflectance reconstruction results for in-the-wild faces. Our project page is https://xingyuren.github.io/id2reflectance.

1. Introduction

Facial reflectance reconstruction aims at predicting reflectance components (*e.g.*, diffuse and specular albedo) and high-frequency normals of the input in-the-wild face image. The recovered 3D faces can be realistically rendered in arbitrary illumination environments. Facial reflectance reconstruction is one of the fundamental problems in computer vision and graphics, with applications ranging from avatar creation [70], telecommunication [49], video games, films, and interactive AR/VR.

To achieve a realistic facial reflectance model, it is

*Corresponding authors

necessary to collect a large-scale and high-quality reflectance dataset from various individuals. However, capturing high-fidelity facial reflectance data is costly and time-consuming, requiring specialized scanning equipment (*e.g.*, Light Stage [8, 21]) and skilled artists for post-processing. Therefore, recent efforts [32, 38, 58] only manage to collect facial reflectance datasets with less than three hundred subjects. As a result, these reflectance models [32, 38, 47, 58] can not generalize very well across diverse real-world identities. To this end, AvatarMe++ [33] employs an image-to-image translation network [27] to synthesize diffuse, specular, and normal maps from large-scale facial texture maps [4]. Based on the complete pairs of facial reflectance, FitMe [34] and Relightify [50] achieve good performance in facial reflectance reconstruction by using StyleGAN [30] and latent diffusion model [54] to learn facial reflectance prior. Nevertheless, none of the facial reflectance data is released from these works [33, 34, 50].

Since the current facial reflectance models [32, 38, 58] are mainly trained in the UV space, a considerable amount of facial reflectance data is needed to learn facial structure, chromaticity, complex details, and other features from scratch. To this end, we train the facial reflectance model with limited light stage captures in the image space instead of the UV space, thus we can take advantage of large-scale, high-quality, and diverse RGB images (*e.g.*, FFHQ [29]). As shown in Fig. 2, facial reflectance data (*e.g.*, diffuse albedo, specular albedo, roughness, and surface normal) share the same facial structure as the RGB faces in the image space. Through joint training, basic facial structure priors can be learned from inexpensive RGB data, leading to a significant reduction in the necessity of reflectance data.

To learn a joint facial reflectance and RGB model, we combine multi-domain data to train VQGAN [15, 42, 61, 72], which employs discrete generative priors, in terms of codebooks, for high-quality image synthesis. However, it is difficult to rely on a single codebook to reconstruct vastly different facial reflectance images as obvious artifacts can be observed in the reconstructions (Fig. 12). To this end, we design a multi-domain codebook learning scheme and each codebook represents a domain-specific discretization of the latent space. For an input image, the final latent representation is a weighted combination derived from these codebooks, resulting in a more expressive and robust representation.

After we train the facial reflectance model, we further apply it for unconstrained facial reflectance reconstruction. Inspired by ID2Albedo [53], we employ identity-conditioned reflectance prediction instead of employing the iterative fitting [34] or conditional inpainting [50]. To inject identity information from the target face into the pre-trained quantized autoencoder, we propose an identity integration module by using AdaIN [26] and train identity swapping only in the RGB domain. As the facial reflectance domain and RGB domain are aligned in our VQGAN model, the identity-swapping capacity learned from the RGB domain can be automatically transferred to the facial reflectance domain. To obtain the complete reflectance maps in UV space, we synthesize three-view identity-conditioned reflectance images in the wrapped space and finally stitch them together for realistic rendering, as illustrated in Fig. 1.

In summary, we make the following contributions:

- We propose a novel facial reflectance reconstruction framework that utilizes multi-domain codebooks to align the facial reflectance domain with the RGB domain to significantly reduce the requirement of captured data.
- We propose a lightweight face swapper module to inject the identity feature into the pre-trained decoder to achieve identity-conditioned facial reflectance generation.
- Extensive experiments demonstrate that the proposed ID2Reflectance can predict high-fidelity facial reflectance from in-the-wild face images.

2. Related Work

Facial Reflectance Reconstrucion. 3D Morphable Models (3DMMs) [3, 14] are typical approaches for face reconstruction from unconstrained images. The linear parametric face model constrains face reconstruction in a low-dimensional space by encoding facial shapes and textures with Principal Component Analysis (PCA), thus neglecting high-frequency facial details. To achieve high-fidelity facial representation, non-linear 3DMM methods [17, 36, 59, 60] are introduced. These models are formulated as neural network decoders where the 3D faces are generated directly from latent vectors. To obtain high-quality facial texture, GANFit [19] employs ProgressiveGAN [28] as the texture generator [9, 20]. However, GANFit lacks relighting capabilities due to backed illumination in the texture.

To overcome the backed illumination problem, the Light Stage [8] is employed to capture high-quality facial reflectance data [32, 38, 58]. By using multiple gradient illuminations with polarization [21], the diffuse and specular components of reflection can be separated. Given reflectance data, face reconstruction is upgraded into facial reflectance reconstruction [1, 13, 25, 32, 34, 38, 47, 50, 58, 69, 70]. AlbedoMM [58] first proposes a drop-in replacement to the 3DMM statistical albedo model with separate diffuse and specular albedo priors, but AlbedoMM is still based on a linear per-vertex color model. Since the capture process by the light stage is expensive and time-consuming, the identity number is usually limited to several hundreds [32, 38, 58]. To this end, AvatarMe++ [33] synthesizes diffuse and specular colors for high-quality textures [4] by training an image-to-image translation network using limited lightstage data. Based on the data from AvatarMe++ [33], FitMe [34] proposes a BRDF generative

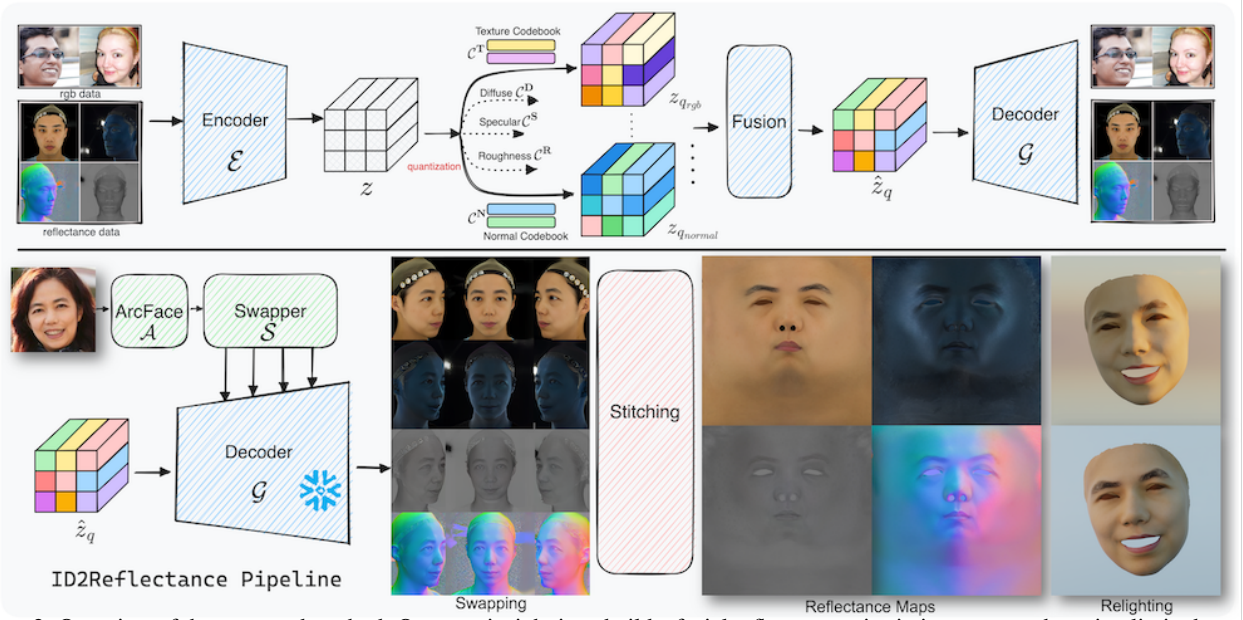

Figure 2. Overview of the proposed method. Our core insight is to build a facial reflectance prior in image space by using limited captures and to recover the reflectance maps for any unconstrained face. We first train multi-domain facial codebooks using a large amount of RGB data and limited reflectance data. Then, given an input unconstrained face, we extract the identity feature from the pre-trained ArcFace [10] model. This feature is fed into the swapper module, which guides the decoder to perform identity injection for all domains. We finally stitch three-view identity-conditioned reflectance images to acquire high-quality rendering assets and renderable 3D faces.

network and a two-stage fitting method to predict facial reflectance for unconstrained images. Relightify [50] utilizes a powerful diffusion model to infer diffuse, specular, and normal.

Even though the series of AvatarMe [32], FitMe [34], and Relightify [50] achieve good performance in facial reflectance reconstruction, neither the texture synthesized reflectance data nor the light stage captured reflectance data is released. In this paper, we define the facial reflectance model in the image space instead of the widely-used UV space, thus we can take advantage of large-scale and high-quality RGB faces to learn facial structure priors. By training multi-domain aligned codebooks, we only require limited reflectance training data for facial reflectance reconstruction.

Face Swapping. The task of face swapping is to transfer the facial identity of the source image/video into the target image/video. Early works mainly utilize traditional image processing [2] and 3D morphable models [62]. Recent methods [18, 40, 48, 52, 56, 65–68, 73] heavily rely on Generative Adversarial Networks [29, 30] and advanced face recognition models [10] to achieve photo-realistic and identity-preserved face swapping. However, all of these face-swapping methods are designed for the RGB domain and cannot be directly used when the target face is from the reflectance domain. In this paper, we first train multi-domain codebooks by using VQGAN [15]. Then, we design an identity injection module for the decoder by using AdaIN [7, 26, 37] to train face-swapping in the RGB domain. As the facial reflectance codebooks and RGB face codebook are aligned during our multi-domain codebook learning, the swapping capacity in the RGB domain can be automatically transferred to the reflectance domain. Therefore, we can decode high-quality identity-conditioned facial reflectance when the input of the encoder is from the facial reflectance domain.

3. Methodology

This work aims to reconstruct high-quality reflectance maps for a single unconstrained face image. To this end, we first train a high-quality facial reflectance model through a multi-domain codebook learning scheme (Sec. 3.1). Based on the pre-trained multi-domain codebooks and a pre-trained face recognition model [10], we train face swapping in the RGB domain and automatically transfer the swapping capacity to the reflectance domain (Sec. 3.2). As illustrated in Fig. 2, we finally design a simple yet efficient inference framework to achieve monocular high-quality facial reflectance map reconstruction (Sec. 3.3).

3.1. Codebook Learning

The main challenge in training expressive facial reflectance models is the absence of large-scale and high-quality reflectance maps scanned from diverse individuals. Existing reflectance models [32, 39] are learned from scratch in the unwrapped facial UV space with limited captured data

(several hundred subjects). As we can see from Fig. 2, facial reflectance data (*e.g.*, diffuse albedo $\mathbf{D}$, specular albedo $\mathbf{S}$, roughness $\mathbf{R}$ and surface normal $\mathbf{N}$) share the same facial structure prior as the common RGB faces in the image space. Since the face structure remains consistent in the image space regardless of different domains, we train a joint quantized autoencoder (*i.e.* VQGAN [15]) to learn the RGB and reflectance codebook simultaneously. Through joint training, the identity diversity in the large amount of inexpensive RGB data can be shared with the limited facial reflectance data. To achieve high-quality facial multi-modal codebooks, we adopt a two-stage training approach.

In the **first stage**, we simply train the quantized autoencoder with a shared codebook using both high-quality facial RBG and reflectance data. For the single-channel data, we duplicate roughness into three-channel images and put specular albedo in the blue channel to simplify the training. Following VQGAN [15], we employ a quantized autoencoder [61] architecture which consists of a encoder $\mathcal{E}$, a discrete codebook $\mathcal{C}$, a decoder $\mathcal{G}$, and a discriminator $\mathcal{D}$. Given a high dimensional image $x \in \mathbb{R}^{H\times W\times 3}$, the encoder $\mathcal{E}$ embeds the input image into the low dimensional code vector $z = \mathcal{E}(x) \in \mathbb{R}^{h\times w\times d}$, where $d = 256$ is the dimension of the latent vector. Then, each grid vector in z is replaced by the nearest vector from the learnable codebook $\mathcal{C} = \left\{c_n \in \mathbb{R}^d\right\}_{n=1}^{N}$:

$$z_q^{(i,j)} = \mathbf{q}(z^{(i,j)}) := \arg\min_{c_n\in\mathcal{C}} \|z^{(i,j)} - c_n\|, \qquad (1)$$

where $z_q \in \mathbb{R}^{h\times w\times d}$ is the quantized feature, $\mathbf{q}(\cdot)$ is a quantisation operation, and $N = 1,024$ is the number of codes in the codebook. Taking the quantized representation z_q as input, the decoder $\mathcal{G}$ can reconstruct the high-quality face image $\hat{x} = \mathcal{G}(\mathbf{q}(\mathcal{E}(x)))$.

To train the quantized autoencoder, we follow [15] to employ three image-level losses: (1) photo reconstruction loss $\mathcal{L}_{photo} = \|\hat{x} - x\|_1$, (2) perceptual loss [71] $\mathcal{L}_{per} = \sum_l \|\mathcal{V}_l(\hat{x}) - \mathcal{V}_l(x)\|_2^2$, where l denotes the different level of a pre-trained VGG [57] model $\mathcal{V}$, and (3) adversarial loss [22] $\mathcal{L}_{adv1} = \log\mathcal{D}(x) + \log(1 - \mathcal{D}(\hat{x}))$. As the quantization operation in Eq. 1 is non-differentiable, VQGAN [15] simply copies the gradients from the decoder to the encoder. The intermediate code-level loss is:

$$\mathcal{L}_{code} = \|\mathrm{sg}\left[\mathcal{E}(x)\right] - z_q\|_2^2 + \beta\|\mathrm{sg}\left[z_q\right] - \mathcal{E}(x)\|_2^2, \qquad (2)$$

where $\mathrm{sg} = [\cdot]$ denotes the stop-gradient operation and $\beta = 0.25$ is controlling the update frequency of the codebook.

With the above image-level and code-level losses, we summarize the training objective as:

$$\mathcal{L}_1 = \mathcal{L}_{photo} + \eta_1\mathcal{L}_{per} + \eta_2\mathcal{L}_{adv1} + \eta_3\mathcal{L}_{code}, \qquad (3)$$

where the loss weights η_1, η_2, and η_3 are set as 1.5, 0.2 and 1.0, respectively. After training, the shared codebook $\mathcal{C}$ presents the facial features containing context-rich structures and details. However, a single codebook makes it hard to reconstruct vastly different facial reflectance images, and the reconstruction results sometimes contain artifacts (Fig. 12).

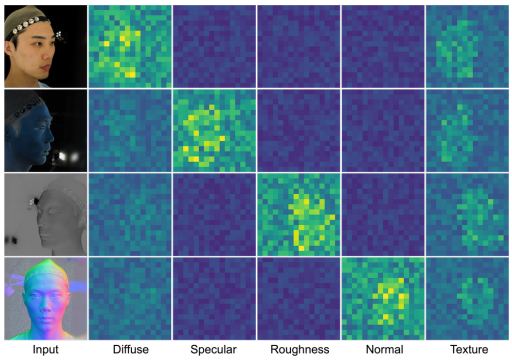

Figure 3. Visualization of codebook fusion weights. Our method uses multiple basis codebooks (especially the RGB texture codebook) for discrete representations, indicating the cross-domain correlation learned by our model.

In the **second stage**, we train domain-specific codebooks to further improve facial reflectance reconstruction. More specifically, we fix the encoder and decoder, and fine-tune the codebook $\mathcal{C}$ learned from the first stage by separately using the reflectance data and the RGB data. Therefore, we obtain another five codebooks, *i.e.* diffuse albedo codebook $\mathcal{C}^{\mathbf{D}}$, specular albedo codebook $\mathcal{C}^{\mathbf{S}}$, roughness codebook $\mathcal{C}^{\mathbf{R}}$, surface normal codebook $\mathcal{C}^{\mathbf{N}}$, and texture codebook $\mathcal{C}^{\mathbf{T}}$.

For a given input x, five quantized representations $\left\{z_{qk} \in \mathbb{R}^{h\times w\times d}\right\}_{k=1}^{K}$ can be generated by quantizing the latent code z with the five domain-specific codebooks $\left\{\mathcal{C}^{\mathbf{D}}, \mathcal{C}^{\mathbf{S}}, \mathcal{C}^{\mathbf{R}}, \mathcal{C}^{\mathbf{N}}, \mathcal{C}^{\mathbf{T}}\right\}$. To combine these five discrete representations $z_{q1,\cdots,K}$, we further train the swin transformer blocks [45] as the fusion weight prediction module, which takes $z \in \mathbb{R}^{h\times w\times d}$ as input and outputs a weight map $w \in \mathbb{R}^{h\times w\times K}$. During the training of the fusion module, the encode, the decoder, and all codebooks are fixed. The code fusion can be expressed by

$$\hat{z}_q = \mathcal{W}(z_{q1,\cdots,K}) = \sum_{k=1}^{K} w_k z_{qk}, \qquad (4)$$

where $\mathcal{W}(\cdot)$ denotes the code fusion operation. By decoding the fused code $\hat{z}_q$, we can get the reconstruction data $\hat{x} = \mathcal{G}(\hat{z}_q)$. As shown in Fig. 3, the reflectance data (*e.g.*, normal) uses multiple basis codebooks (*e.g.*, normal and RGB texture codebooks) for discrete representations, which indicates the alignment behavior between the reflectance domain and RGB domain.

3.2. Identity Swapping

Since the facial reflectance domain and RGB domain are aligned during our multi-domain codebooks learning in

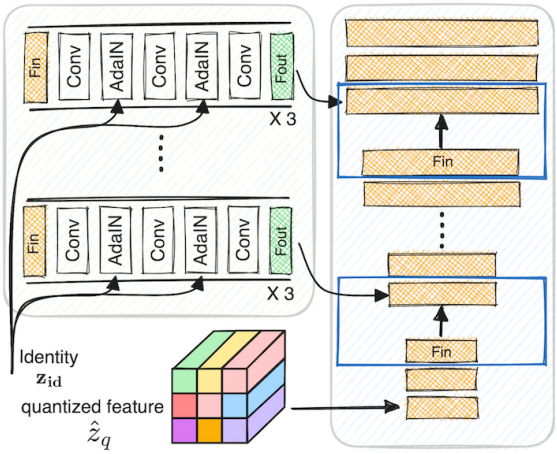

Figure 4. Detailed architecture of our swapper module. The yellow boxes represent the original multi-scale features from the decoder $\mathcal{G}$, and the green boxes represent the residual features generated by each identity injection branch. We use the small-scale feature map as input to generate identity-conditioned residual features, which will be added to the up-sampled feature map.

Sec. 3.1, we explore training identity swapping [37] in the RGB domain and automatically transferring the identity-swapping capacity to the facial reflectance domain. Specifically, we propose an identity-conditioned swapper module (denoted as $\mathcal{S}$) and integrate it with the above-learned quantized autoencoder.

As illustrated in Fig. 4, our swapper module consists of serval parallel branches, each containing three identity injection blocks. Each block consists of convolution and AdaIN [26] operators and uses the Leaky ReLU as the activation function. We insert the swapper branch into each upsampling stage of the decoder $\mathcal{G}$. Specifically, we take the small-scale feature map as the input and use identity as the condition to generate a residual feature, which is finally incorporated into the up-sampled feature map. The identity embedding network is a ResNet-100 model trained on the large-scale WebFace dataset [74] using the ArcFace loss [10]. By using this pre-trained ArcFace model, we can extract face identity features that are robust to changes in illumination, pose, and occlusion. For identity embedding integration, it is achieved by using AdaIN [26] as follows:

$$AdaIN(f, z_{id}) = \sigma_{z_{id}} \frac{f - \mu_f}{\sigma_f} + \mu_{z_{id}}, \tag{5}$$

where $z_{id} \in \mathbb{R}^{1\times512}$ is the identity embedding, μ_f and σ_f are the channel-wise mean and standard deviation of the input feature f, and $\sigma_{z_{id}}$ and $\mu_{z_{id}}$ are two modulation parameters generated from z_{id} through FC layers.

To train our swapper module, we employ the RGB face recognition dataset [5]. During training, we fix the pre-trained encoder, codebooks, and decoder. We employ the identity loss function, which is the cosine distance between the input image $\mathbf{I_{id}}$ and the decoded face $\hat{x}$:

$$\mathcal{L}_{id} = 1 - \frac{\mathcal{A}(\mathbf{I_{id}})\mathcal{A}(\hat{x})}{\|\mathcal{A}(\mathbf{I_{id}})\|_2 \cdot \|\mathcal{A}(\hat{x})\|_2}, \tag{6}$$

where $\mathcal{A}$ is the pre-trained ArcFace model. Besides, we utilize a pyramid discriminator by referring to projected GAN [55].

$$\begin{aligned}\mathcal{L}_{adv2} = \min_{\mathcal{G}} \max_{\mathcal{D}} \sum_{l\in L} \Big(& E_x[\log \mathcal{D}_l(\mathcal{F}_l(x))] \\ & + E_{\hat{z}_q}[log(1 - \mathcal{D}_l(\mathcal{F}_l(\mathcal{G}(\hat{z}_q))))]\Big),\end{aligned} \tag{7}$$

where $l \in \{1, \cdots, L\}, L = 4$ indicates different feature levels, $\mathcal{F}$ is a fixed ImageNet model mapping the high-resolution image x or decoded image $\mathcal{G}(\hat{z}_q)$ into four-scale feature pyramids, and $\mathcal{D}_l$ is the corresponding discriminator applied to each feature level. The complete training objective for face swapping is as follows:

$$\mathcal{L}_2 = \mathcal{L}_{id} + \lambda_1 I \mathcal{L}_{photo} + \lambda_2 \mathcal{L}_{lpips} + \lambda_3 \mathcal{L}_{adv2}, \tag{8}$$

where I is the indicator function which is 1 when the input face x and the target face $\mathbf{I_{id}}$ are from the same identity and 0 otherwise. $\mathcal{L}_{lpips}$ denotes the LPIPS loss [71]. The loss weights λ_1, λ_2, λ_3 are set as 1.5, 0.1 and 0.1, respectively. Compared to general face-swapping networks [7, 37], the proposed codebook-based identity swapping explicitly decouples identity features from non-identity features, ensuring the swapped face $\hat{x}$ and input template x in the same domain as well as maintaining the high-resolution output of the original decoder.

3.3. Monocular Facial Reflectance Inference

Once multi-domain codebooks and the identity-conditioned swapper are trained, we design an ID2Reflectance framework to reconstruct the reflectance maps as shown in Fig. 2. Unlike previous facial reflectance prediction methods [32, 39] designed in the UV space, our framework first synthesizes multi-view identity-conditioned reflectance images in the wrapped space and then stitches them together to obtain the final reflectance maps.

Multi-view Reflectance Swapping. Given an input face $\mathbf{I_{id}}$, we employ the identity similarity to search for the closet reflectance template (*e.g.*, diffuse albedo $\mathbf{D}$, specular albedo $\mathbf{S}$, roughness $\mathbf{R}$, and surface normal $\mathbf{N}$). For each of these four reflectance domains, we select the three fixed views (*i.e.*, left, frontal, and right). These 12 template reflectance images provide the condition of the facial attribute (*e.g.*, domain and pose) for the quantized autoencoder, while the $\mathbf{I_{id}}$ provides the information of identity. As illustrated in Fig. 2, the input face $\mathbf{I_{id}}$ is passed through the ArcFace embedding network $\mathcal{A}$ to extract the

identity feature $z_{id} \in \mathbb{R}^{1\times512}$, while the 12 multi-view reflectance templates separately go through the encoder, multiple codebook quantization, latent representation fusion, and identity-conditioned generator to obtain multi-view identity-conditioned reflectance data.

Multi-view UV Stitching. To stitch the three-view reflectance images to get the reflectance map, we use Deep3D [12] to estimate the shape in each view of the diffuse albedos, establish dense correspondences among three views, and unfold the facial reflectance in the UV space. Besides, we employ a face parsing model [41] to predict the facial region, excluding non-facial areas (*e.g.*, hats) for the unwrapped texture UV maps. To blend three-view diffuse albedos, we perform color matching in the YUV space [1] to merge the maps from the left and right views with the map from the frontal view. To obtain the specular albedo, roughness, and surface normal maps, we employ the same dense correspondences estimated for multi-view diffuse albedos. In this way, we obtain high-quality and identity-preserved facial reflectance assets as shown in Fig. 2.

4. Experiments

4.1. Implementation Details

All our implementations are based on PyTorch [51] and Nvidia A6000 GPUs. We employ the Adam [31] optimizer with a batch size of 16 for all training tasks in this paper. For reflectance prior, we first train our models on FFHQ [29] and captured dataset, and all images are resized to 512×512 for training. We set the latent code size as 16×16, and the codebook size as 1024×256. Our captured dataset contains 135 participants with gender, age, and race diversity. We randomly select 115 subjects for training and the rest 20 subjects for testing. In the first stage, we train the model on a mixture of FFHQ and reflectance data, with an initial learning rate of 8e-5 and a total iteration number of 700K. In the second stage, we fine-tune each domain-specific codebook on the corresponding dataset and finally fine-tune the fusion module by a balanced sampling on FFHQ and captured data. For each fine-tuning step, the initial learning rate is set as 5e-5, and the iteration number is 100K.

To train the face swapper module, we choose VGGFace2HQ [5] as our training set. To improve the training data quality, we remove small images, improve the resolution by GFPGAN [64], and resize images to 512×512 for training. The pre-processed dataset contains 1.77 Million images from 8K identities. Our swapper module $\mathcal{S}$ uses a fully connected architecture with random initialization. The target image is resized to 112×112 [11, 23, 43, 44] for ID feature embedding and the size of the final swapped reflectance is 512×512. Here, the initial learning rate is set as 7e-5, and the iteration number is set to 500K.

Our entire ID2Reflectance framework consists of identity embedding, multi-view reflectance swapping, dense correspondence estimation, and stitching. The total inference time is 3.8 seconds/face on the A6000 GPU. To visualize our relighting results, we employ the off-the-shelf geometry prediction network Deep3D [12] to estimate the face shape.

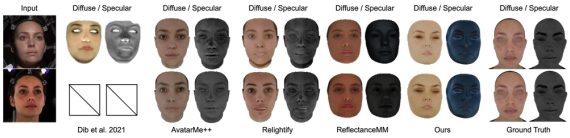

Figure 5. Comparison of diffuse and specular albedo reconstruction on Digital Emily project. From left to right: input image, Dib *et al*. [13], AvatarMe++ [33], Relightify [50], ReflectanceMM [24], ours and ground-truth.

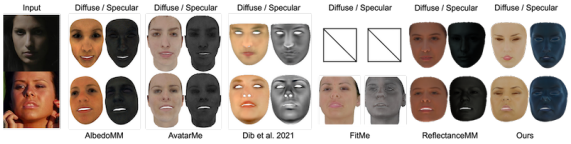

Figure 6. Comparison with recent single image reflectance prediction methods. From left to right: input image, AlbedoMM [58], AvatarMe [32], Dib *et al*. [13], FitMe [34], ReflectanceMM [24] and ours.

4.2. Reflectance Reconstruction Results

We first evaluate our method in the task of monocular facial reflectance prediction. Following recent state-of-the-art methods [13, 24, 33, 34, 50], we perform visual comparisons on the well-known Digital Emily and in-the-wild data, respectively. In Fig. 5 and 6, our method yields consistent and identity-preserved results regardless of the illumination variation. The minor lipstick artifacts in our results are due to the VGGFace2HQ dataset. This is because the restorer, GFPGAN, alters the makeup of some female subjects.

In our approach, we select the reflectance template with the highest similarity from the template library as the input of the encoder. To evaluate the impact of template selection, we randomly select a Chinese male adult template and provide step-wise qualitative results in Fig. 7. Even though the reflectance template is fixed, our method still attains a high level of identity preservation. The only difference is the skin tone colour in diffuse albedo. As shown in Fig. 8, we visualize some relighting results of our method. Given the ethnicity-diverse template library, our method is capable of reconstructing a wide range of races, producing relighting results at superior quality.

4.3. Albedo Reconstruction Results

We employ CelebAMask-HQ [35] and our test set (ground truth reflectance data captured from 20 participants) to evaluate our facial diffuse albedo prediction. As shown in Fig. 9 and 10, we compare these results with recent state-of-the-art albedo estimation methods. Even though some of the input images contain occlusions, the albedos estimated by our method are robust under these occlusions, exhibiting real-

Figure 7. Qualitative results on unconstrained images. Given an input image, we initially obtain multi-domain three-view images through reflectance swapping and then stitch them into high-quality reflectance maps for relighting.

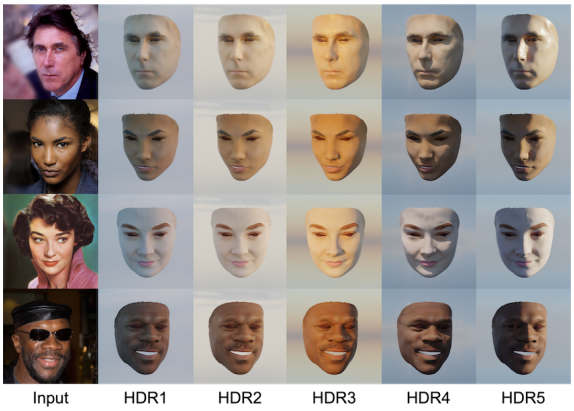

Figure 8. Relighting results for people from different races. Each column of images maintains the same HDR illumination.

ism, identity preservation, and race consistency. In addition, we quantitatively compare our method with previous methods using identity similarity, PSNR, SSIM, and LPIPS [71] metrics. To calculate identity similarity for TRUST [16], ID2Albedo [53], and our method, we first overlay the diffuse albedo onto the original input image and then compute the cosine similarity between the overlaid albedo and the input face by using a pre-trained CosFace [63] model. Tab. 1 demonstrates that our method achieves the best results in all image-level metrics.

4.4. Ablation Studies

ID2Reflectance Framework. In the first and second rows of Tab. 2, we reconstruct the three-view reflectance images directly using the captured data and stitch them together to get complete UV maps. As we can see, the multi-domain

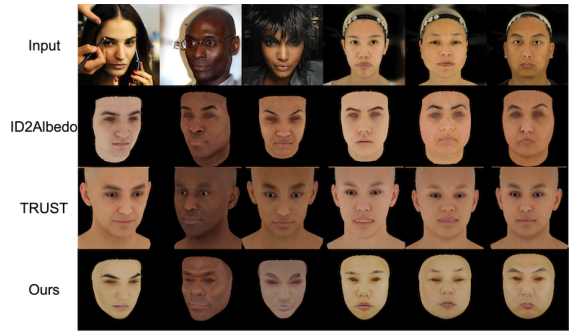

Figure 9. Comparisons of albedo estimation on in-the-wild and constrained images. Input images are from CelebAMask-HQ [35] and our captured test set.

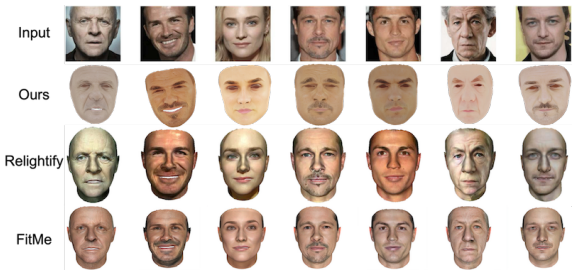

Figure 10. Comparisons of albedo estimation on in-the-wild and constrained images. From top to bottom: Input, Ours, Relightify [50], and FitMe [34].

codebook learning significantly outperforms the joint codebook learning. In the third and fourth rows, we compare reflectance swapping under the fixed template and the closest template by using multi-domain codebooks. The results indicate that the fixed template setting obviously underperforms the closest template setting on diffuse albedo estimation. As depicted in Fig. 7, using a fixed template results in

Methods	PSNR↑	SSIM↑	LPIPS↓	ID↑
TRUST [16]	21.63	0.852	0.2014	0.478
ID2Albedo [53]	23.72	0.884	0.1549	0.532
Ours	**28.47**	**0.923**	**0.1248**	**0.735**

Table 1. Comparisons of our method with previous methods [16, 53] on albedo estimation. Note that we compute PSNR and SSIM on the captured test set, while the ID similarity and LPIPS metrics are computed on the CelebAMask-HQ [35] dataset.

Configs	Diffuse	Specular	Roughness	Normal
Joint codebook	24.87	20.95	21.31	20.56
Multi-domain codebooks	31.62	30.96	31.59	30.32
Fixed Template	25.26	26.44	29.56	25.77
Closest Template	28.47	26.68	30.32	26.83

Table 2. Comparison of ID2Reflectance framework under different configurations. (1) joint codebook v.s. multi-domain codebooks for reflectance reconstruction, and (2) fixed swapping template v.s. closest swapping template for identity-conditioned reflectance prediction. We calculate the average PSNR between the reconstructed and the ground truth reflectance maps on our test set.

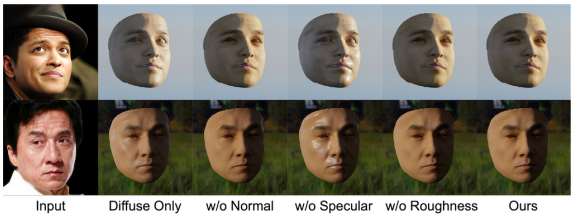

Figure 11. Ablation studies on each reflectance component predicted by our model.

a skin tone gap in diffuse albedos when the target and source faces are from two different races. In addition, we perform ablation studies on each reflectance component predicted by our method in Fig. 11. The results verify that the predicted normal, specular, roughness maps are beneficial to improve the quality of physically based rendering. While good results were achieved, it's important to note that our reconstructed PBR maps may not be physically plausible, as they weren't supervised using ground truth PBR maps.

Multi-domain Codebook. As shown in Fig. 12, the input and reconstruction results from the RGB and reflectance domains demonstrate that the joint codebook has limited capability in recovering cross-domain data, leading to visible artifacts and an over-smoothed face structure. By contrast, our multi-domain codebooks achieve finer facial structure and better reconstruction quality.

Swapper Module. In Fig. 13, we compare our swapping module with state-of-the-art methods. As we can see, SimSwap [7] and InfoSwap [18] encounter great difficulties in the reflectance domain, generating obvious noises and artifacts in the results. Moreover, StyleGAN-based swapping methods (*e.g.*, E4S [46] and 3dSwap [40]) can only run in the texture and diffuse domains as the patterns in StyleGAN [29] and EG3D [6] are insufficient to cover other reflectance domains. By contrast, our approach achieves consistent and high-quality swapping results on both RGB and reflectance domains.

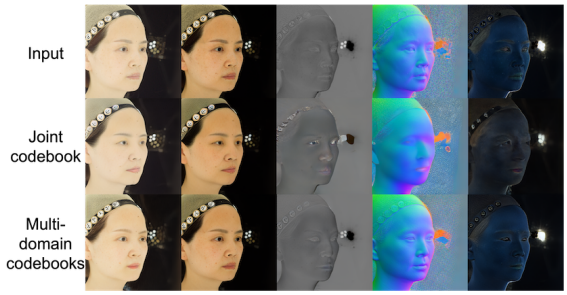

Figure 12. Reconstruction comparison of using joint and multi-domain codebooks. Inputs are the same faces from PBR domains.

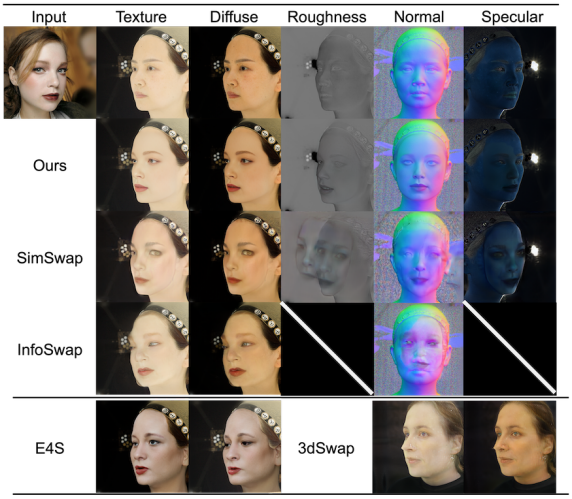

Figure 13. Cross-domain swapping comparison with other methods (*e.g.*, SimSwap [7], InfoSwap [18], E4S [46], and 3dSwap [40]). Input RGB face provides the source identity while other reflectance data are the target.

5. Conclusions and Discussions

In this paper, we present a novel monocular facial reflectance reconstruction method. By learning high-quality multi-domain discrete codebooks, we can obtain a reliable reflection prior from limited captured data. Our model utilizes identity features as conditions to reconstruct multi-view reflectance images directly from the multi-domain codebooks and then stitches them together into a complete reflectance map. Experiments demonstrate that our method exhibits good qualitative and quantitative performance, and excellent generalization to real-world images.

Broader Impacts. Although not the purpose of our work, monocular facial reflectance reconstruction may potentially be abused. Nevertheless, our method can be used for high-quality reconstruction and rendering, opening new avenues for faster avatar creation and promotion of the metaverse.

Acknowledgements. This work was supported in part by NSFC (62322113, 62376156, 62101325, 62201342), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and the Fundamental Research Funds for the Central Universities.

References

[1] Haoran Bai, Di Kang, Haoxian Zhang, Jinshan Pan, and Linchao Bao. Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction. In *CVPR*, 2023. 2, 6

[2] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K Nayar. Face swapping: automatically replacing faces in photographs. In *SIGGRAPH*, 2008. 3

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999. 2

[4] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *CVPR*, 2016. 2

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 2018. 5, 6

[6] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022. 8

[7] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *ACM MM*, 2020. 3, 5, 8

[8] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Computer Graphics and Interactive Techniques*, 2000. 2

[9] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *CVPR*, 2018. 2

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 3, 5

[11] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020. 6

[12] 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 *CVPRW*, 2019. 6

[13] 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 *ICCV*, 2021. 2, 6

[14] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhöfer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3D morphable face models - past, present, and future. *TOG*, 2020. 2

[15] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2, 3, 4

[16] Haiwen Feng, Timo Bolkart, Joachim Tesch, Michael J Black, and Victoria Abrevaya. Towards racially unbiased skin tone estimation via scene disambiguation. In *ECCV*, 2022. 7, 8

[17] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *TOG*, 2021. 2

[18] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Information bottleneck disentanglement for identity swapping. In *CVPR*, 2021. 3, 8

[19] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *CVPR*, 2019. 2

[20] 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 *ECCV*, 2020. 2

[21] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. *TOG*, 2011. 2

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 4

[23] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. In *ICLR*, 2022. 6

[24] Yuxuan Han, Zhibo Wang, and Feng Xu. Learning a 3d morphable face reflectance model from low-cost data. In *CVPR*, 2023. 6

[25] Yuxuan Han, Zhibo Wang, and Feng Xu. Learning a 3d morphable face reflectance model from low-cost data. In *CVPR*, 2023. 2

[26] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2, 3, 5

[27] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2

[28] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 2

[29] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 3, 6, 8

[30] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2, 3

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[32] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. AvatarMe: Realistically renderable 3D facial reconstruction. In *CVPR*, 2020. 2, 3, 5, 6

[33] 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. *TPAMI*, 2021. 2, 6

[34] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Fitme: Deep photorealistic 3d morphable model avatars. In *CVPR*, 2023. 2, 3, 6, 7

[35] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 6, 7, 8

[36] Gun-Hee Lee and Seong-Whan Lee. Uncertainty-aware mesh decoder for high fidelity 3d face reconstruction. In *CVPR*, 2020. 2

[37] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. In *CVPR*, 2020. 3, 5

[38] Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, and Hao Li. Learning formation of physically-based face attributes. In *CVPR*, 2020. 2

[39] 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 *CVPR*, 2020. 3, 5

[40] Yixuan Li, Chao Ma, Yichao Yan, Wenhan Zhu, and Xiaokang Yang. 3d-aware face swapping. In *CVPR*, 2023. 3, 8

[41] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. Roi tanh-polar transformer network for face parsing in the wild. *IVC*, 2021. 6

[42] Kechun Liu, Yitong Jiang, Inchang Choi, and Jinwei Gu. Learning image-adaptive codebooks for class-agnostic image restoration. In *ICCV*, 2023. 2

[43] Yang Liu, Fei Wang, Jiankang Deng, Zhipeng Zhou, Baigui Sun, and Hao Li. Mogface: Towards a deeper appreciation on face detection. In *CVPR*, 2022. 6

[44] Yang Liu, Jiankang Deng, Fei Wang, Lei Shang, Xuansong Xie, and Baigui Sun. Damofd: Digging into backbone design on face detection. In *ICLR*, 2023. 6

[45] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 4

[46] Zhian Liu, Maomao Li, Yong Zhang, Cairong Wang, Qi Zhang, Jue Wang, and Yongwei Nie. Fine-grained face swapping via regional gan inversion. In *CVPR*, 2023. 8

[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 *CVPR*, 2021. 2

[48] Yuchen Luo, Junwei Zhu, Keke He, Wenqing Chu, Ying Tai, Chengjie Wang, and Junchi Yan. Styleface: Towards identity-disentangled face generation on megapixels. In *ECCV*, 2022. 3

[49] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. Pixel codec avatars. In *CVPR*, 2021. 1

[50] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, and Stefanos Zafeiriou. Relightify: Relightable 3d faces from a single image via diffusion models. In *ICCV*, 2023. 2, 3, 6, 7

[51] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6

[52] Xiaohang Ren, Xingyu Chen, Pengfei Yao, Heung-Yeung Shum, and Baoyuan Wang. Reinforced disentanglement for face swapping without skip connection. In *CVPR*, 2023. 3

[53] Xingyu Ren, Jiankang Deng, Chao Ma, Yichao Yan, and Xiaokang Yang. Improving fairness in facial albedo estimation via visual-textual cues. In *CVPR*, 2023. 2, 7, 8

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[55] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. In *NeurIPS*, 2021. 5

[56] Kaede Shiohara, Xingchao Yang, and Takafumi Taketomi. Blendface: Re-designing identity encoders for face-swapping. In *CVPR*, 2023. 3

[57] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4

[58] William A. P. Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua Tenenbaum, and Bernhard Egger. A morphable face albedo model. In *CVPR*, 2020. 2, 6

[59] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Fml: Face model learning from videos. In *CVPR*, 2019. 2

[60] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *CVPR*, 2018. 2

[61] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 2, 4

[62] Thomas Vetter and Volker Blanz. Estimating coloured 3d face models from single images: An example based approach. In *ECCV*, 1998. 3

[63] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, 2018. 7

[64] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 6

[65] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. In *CVPR*, 2021. 3

[66] Chao Xu, Jiangning Zhang, Yue Han, Guanzhong Tian, Xianfang Zeng, Ying Tai, Yabiao Wang, Chengjie Wang, and

Yong Liu. Designing one unified framework for high-fidelity face reenactment and swapping. In *ECCV*, 2022.

[67] Chao Xu, Jiangning Zhang, Miao Hua, Qian He, Zili Yi, and Yong Liu. Region-aware face swapping. In *CVPR*, 2022.

[68] Zhiliang Xu, Hang Zhou, Zhibin Hong, Ziwei Liu, Jiaming Liu, Zhizhi Guo, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Styleswap: Style-based generator empowers robust face swapping. In *ECCV*, 2022. 3

[69] 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. *TOG*, 2018. 2

[70] Longwen Zhang, Qiwei Qiu, Hongyang Lin, Qixuan Zhang, Cheng Shi, Wei Yang, Ye Shi, Sibei Yang, Lan Xu, and Jingyi Yu. Dreamface: Progressive generation of animatable 3d faces under text guidance. In *SIGGRAPH*, 2023. 1, 2

[71] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 4, 5, 7

[72] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, 2022. 2

[73] Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. One shot face swapping on megapixels. In *CVPR*, 2021. 3

[74] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *CVPR*, 2021. 5