Improving Fairness in Facial Albedo Estimation via Visual-Textual Cues

Xingyu Ren Jiankang Deng* Chao Ma* Yichao Yan Xiaokang Yang
[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
{rxy_sjtu,chaoma,yanyichao,xkyang}@sjtu.edu.cn, jiankangdeng@gmail.com

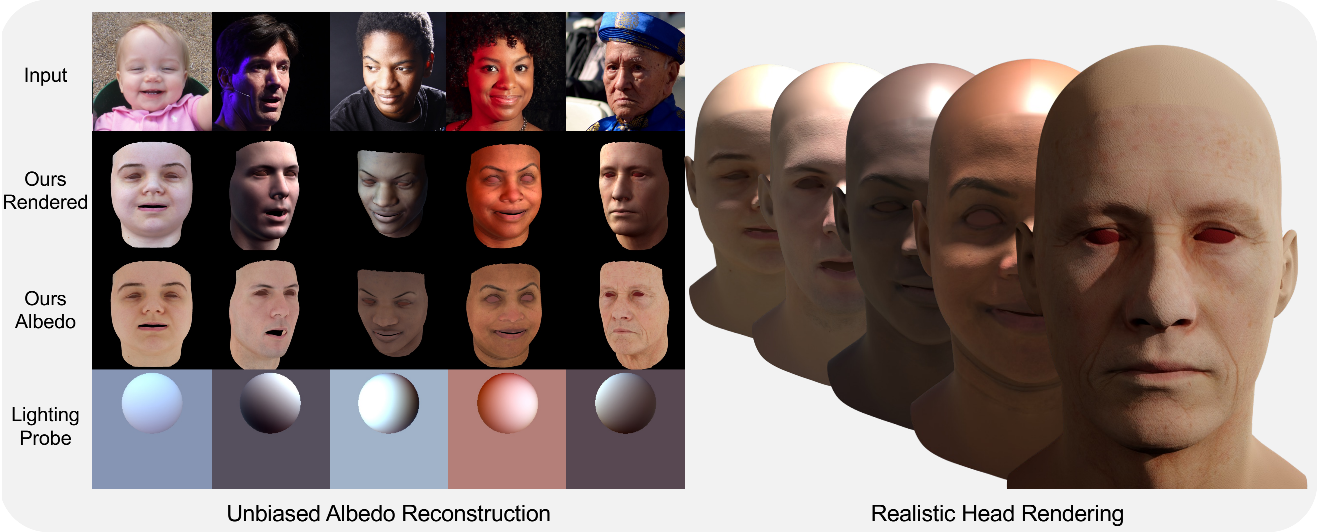

Figure 1. We introduce *ID2Albedo*, a high-quality, unbiased albedo reconstruction method. *ID2Albedo* maps the facial identity features to the latent space of the albedo generator and uses novel visual-textual cues to constrain albedo attributes. Our approach can alleviate the illumination/albedo ambiguity and generate high-fidelity albedo maps for realistic rendering. Images are all from FFHQ [30] dataset.

Abstract

Recent 3D face reconstruction methods have made significant advances in geometry prediction, yet further cosmetic improvements are limited by lagged albedo because inferring albedo from appearance is an ill-posed problem. Although some existing methods consider prior knowledge from illumination to improve albedo estimation, they still produce a light-skin bias due to racially biased albedo models and limited light constraints. In this paper, we reconsider the relationship between albedo and face attributes and propose a ID2Albedo to directly estimate albedo without constraining illumination. Our key insight is that intrinsic semantic attributes such as race, skin color, and age can be used to constrain the albedo map. We first introduce visual-textual cues and design a semantic loss to supervise facial albedo estimation. Specifically, we pre-define text labels such as race, skin color, age, and wrinkles. Then, we employ the text-image model (CLIP) to compute the similarity between the text and the input image, and assign a pseudo-label to each facial image. We constrain generated albedos in the training phase to have the same attributes as the inputs. In addition, we train a high-quality, unbiased facial albedo generator and utilize the semantic loss to learn the mapping from illumination-robust identity features to the albedo latent codes. Finally, our ID2Albedo is trained in a self-supervised way and outperforms state-of-the-art albedo estimation methods in terms of accuracy and fidelity. It is worth mentioning that our approach has excellent generalizability and fairness, especially on in-the-wild data.

1. Introduction

3D face reconstruction is one of the fundamental problems in computer vision and graphics. It aims to estimate realistic 3D face shapes and appearances from 2D images, given only multi-view or single-view images. 3D face reconstruction plays a vital role in numerous vision applications, such as face manipulation [52], speech-driven facial animation [51], and video conferencing [56]. Since the pioneering work of 3D Morphable Model (3DMM) [54], monocular face reconstruction methods have made remark-

[*] Corresponding authors.

able progress due to their high speed and geometric accuracy. To enable more realistic applications such as avatar creation, interactive AR/VR, *etc*., fine-grained albedo reconstruction attracts a lot of attention [16].

Inferring albedo from pixels is an ill-posed problem, and existing methods attempt to achieve approximate results. The primary approaches are 1) creating a texture model to restrict the albedo space [24, 41, 49], and 2) introducing additional lighting constraints to reduce ambiguity [1,12,17]. Despite these constraints, most current albedo reconstruction methods continue to bias light-colored albedos, unfair to people of different ages and races. The main reasons behind the biased albedo estimation include 1) biased albedo models and 2) limited lighting constraints. To address the above issues, TRUST [16] rebuilt a balanced albedo model, estimated the environment light from the scene and used this prior to decrease the ambiguity between light and albedo. Given the difficulty of the illumination estimation for both face and scene, the albedo estimation method proposed in TRUST [16] is still vulnerable under complex scenarios and complicated facial appearance variations.

Since the facial albedo is a property of individual faces that should be consistent even when the lighting changes, could we design an illumination-robust albedo estimation method like the face recognition model [10] and the face attribute analysis model [28]? In this work, we provide an affirmative answer by proposing a novel *ID2Albedo* method. We first train a high-resolution albedo generator as the current PCA-based albedo model [49] fails in reconstructing high-frequency facial details. Given hundred-level training data, high-resolution Generative Adversarial Networks (GANs) [31] are not easy to train. To this end, we replace the single large discriminator with four smaller discriminators, which are applied to the feature pyramids [36] produced by a fixed ImageNet model. Based on our high-resolution albedo generator, we further utilize the illumination-robust identity features [10] to predict the latent codes to reconstruct albedo maps, ensuring the generalization ability on in-the-wild data.

Given the fact that facial albedo is related to facial attributes (*e.g*. ethnicity, age, and skin color), we consider exploring attribute constraints during albedo estimation. For example, African albedos are primarily dark, while Caucasian albedos are mostly light. However, race alone is insufficient because the albedo of different individuals within a race varies due to age, skin color, and other factors. Therefore, we attempt to use diverse facial attribute priors to constrain the albedo estimation. Considering that few face datasets contain diverse semantic labels and manual annotation is time-consuming, we utilize a recent state-of-the-art visual-textual model, CLIP [42], to provide semantic cues for individual faces. Specifically, we predefine diverse texts from various perspectives, including race, skin tones, age, wrinkles, *etc*., and then compute the corresponding semantic attribute labels by embedded image features. Based on the pseudo attribute labels, we propose a novel semantic loss to compare the attribute differences between the reconstructed face and the original input face. The entire pipeline is self-supervised by a differentiable rendering framework. To verify the effectiveness of the proposed albedo reconstruction approach, we conduct exhaustive evaluations on the FAIR benchmark and real-world images. The results show that our method consistently achieves competitive performance compared to state-of-the-art methods, especially under various lighting conditions.

In summary, our contributions are summarized as follows:

- We first train a high-resolution, expressive, and nonlinear face albedo generator. Then, we construct a powerful face albedo predictor, named ID2Albedo, by utilizing the face identification features from a pre-trained face recognition network.
- We employ visual-textual cues in the face reconstruction framework to overcome the illumination/albedo ambiguity problem by constraining facial semantic attributes.
- The proposed method improves the accuracy and fairness of facial albedo estimation, achieving state-of-the-art performance on the FAIR benchmark.

2. Related Work

Face and head reconstruction from monocular RGB, RGB-D, or multi-view data are well-explored in computer vision and computer graphics, and can be divided into optimization-based [1, 3, 4, 46, 53] and regression-based approaches [5, 12, 17, 23, 32, 48, 50]. More details are described in [14, 61]. Albedo reconstruction is a component of 3D face appearance reconstruction, which is an inverse rendering problem. The following focuses on work related to albedo reconstruction via monocular faces.

Albedo Modeling. Current monocular face reconstruction methods mainly rely on statistical facial models such as 3DMM, which consists of a geometric space for shape reconstruction and an appearance space for albedo reconstruction. Please see [14] for more information. The widely used Basel Face Model (BFM) [41] was developed from about 200 European subjects. However, this imbalanced data can lead to a strongly biased appearance space, failing to rebuild dark skin tones appropriately. Smith *et al*. [49] were concerned about this problem. AlbedoMM created an albedo model from varied light-stage data and simultaneously modeled the diffuse and specular albedo models to increase the diversity of the appearance space. Based on AlbedoMM, the recent TRUST [16] discovered that the present albedo model still has the problem of imbalance between different human races. They made a racial-balanced albedo model, which is more balanced for people of different races and skin tones.

In addition to the PCA-based approaches mentioned above, GAN-based models are prominent. Deng *et al*. [8] trained a generative adversarial network to reconstruct textures from a single image. Gecer *et al*. [21, 22] trained a powerful texture GAN based on 10K texture data, dramatically improving texture realism. However, their reconstructed textures are baked with lighting information, whereas our albedo is the consequence of texture delighting. Lattas *et al*. [34, 35] trained an Image-to-Image Translation network with light-stage data to synthesize diffuse/specular albedo from high-quality textures. However, the training data restricts the generalization capacity when confronted with people of different races. Our approach combines both benefits and achieves a high-quality, racially balanced albedo generator.

Disambiguating Appearance and Lighting. Recovering reliable illumination and albedo from image appearance species is an ill-posed problem [43]. Although the appearance prior has constrained the albedo variation, it does not completely eliminate the ambiguity problem. The usual idea is to find stronger prior knowledge to constrain both. Hu *et al*. [26] normalized the symmetry of albedo. Aldrian *et al*. [1] regularized light by imposing a “gray world” constraint that constrains light to be monochromatic, and subsequent work such as [12, 17] used a similar regularization approach for approximate decomposition. Egger *et al*. [13] took into account the existence of a certain distribution of illumination and directly learned a statistical prior for the SH coefficients. TRUST [16] extends the range of light estimation by decomposing light into face light and ambient light and using ambient light consistency to constrain light estimation. Unlike previous regularization approaches, we introduce an open-world visual-textual model that provides rich semantic attribute labeling for various faces, and then directly constrains the albedo to accomplish a successful decomposition.

Text-Driven Generation and Manipulation. Our method is comparable to image manipulation techniques controlled through text descriptions encoded in CLIP [42]. CLIP learned a joint embedding space for images and text. StyleCLIP [40] leveraged pre-trained StyleGAN [30, 31] for CLIP-guided image modification. VQGAN-CLIP [15] employed CLIP for text-guided image generation. In the stylization domain, Gal *et al*. [20] used CLIP to fine-tune a pre-trained StyleGAN for images. Based on a textual question, Text2Mesh [38] predicted color and geometry details for a specified template mesh. In an implicitly differentiable rendering framework, TANGO [6] employed CLIP to improve the physical attributes of objects for more realistic stylization. In the area of generation, Sanghi *et al*. [45] utilized CLIP for unconditional 3D voxel generation. CLIPDraw [19] produced 2D vector graphics for drawing styles using textual instruction. Jetchev *et al*. [27] optimized the parameters of SMPL mannequins using CLIP to generate digital creatures. Unlike the approaches discussed above, our textual cues are fixed during training and do not require any text input during inference. We regard CLIP as a powerful semantic attribute annotator that allows us to restrict the albedo directly.

3. Methods

This work aims at reconstructing high-quality, unbiased albedo maps from in-the-wild face images. To this end, we first train a high-resolution face albedo generator (Sec. 3.1) and design an albedo estimation method based on a pre-trained face recognition model (Sec. 3.2). To reduce the ambiguity of albedo estimation, we explore the semantic facial attribute constraints through visual-textual cues (Sec. 3.3). As illustrated in Fig. 2, our method is trained in a self-supervised learning way by combining other losses to achieve good decomposition between the illumination and albedo (Sec. 3.4).

3.1. High-Resolution Albedo Generator

The biggest challenge behind building an expressive face albedo model is the deficiency of large-scale and high-quality albedo maps collected from diverse identities. In AlbedoMM [49], a novel lightstage capture system is proposed for acquiring albedo maps that fully factor out the effects of illumination. However, they have only captured a dataset of 50 individuals (13 females) and their participants range in age from 18 to 67, covering skin types I-V of the Fitzpatrick scale [18]. Based on the limited albedo training data, a morphable face albedo model [49] is built by Principal Component Analysis (PCA). The linear basis in PCA, even though remarkable in representing the basic characteristics of the facial albedo, fails in reconstructing high-frequency facial details (*e.g*. wrinkles and pores). Recently, Generative Adversarial Networks (GANs) [31] have shown excellent ability in capturing image details. Specifically, GANs aim to optimize the following minimax objective

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \Big(E_x[\log \mathcal{D}(x)] + E_z[log(1 - \mathcal{D}(\mathcal{G}(z)))] \Big), \quad (1)$$

where $\mathcal{G}$ is the image generator and $\mathcal{D}$ is the image discriminator. The generator $\mathcal{G}$ maps the latent vectors z sampled from a normal distribution $\mathcal{P}_z$ to the generated images $\mathcal{G}(z)$. The discriminator $\mathcal{D}$ then aims to discriminate real images $x \sim \mathcal{P}_x$ from generated images $\mathcal{G}(z) \sim \mathcal{P}_z$.

In this paper, we purchase 142 high-quality albedo maps from the 3D Scan Store to build our high-resolution albedo

https://www.3dscanstore.com/

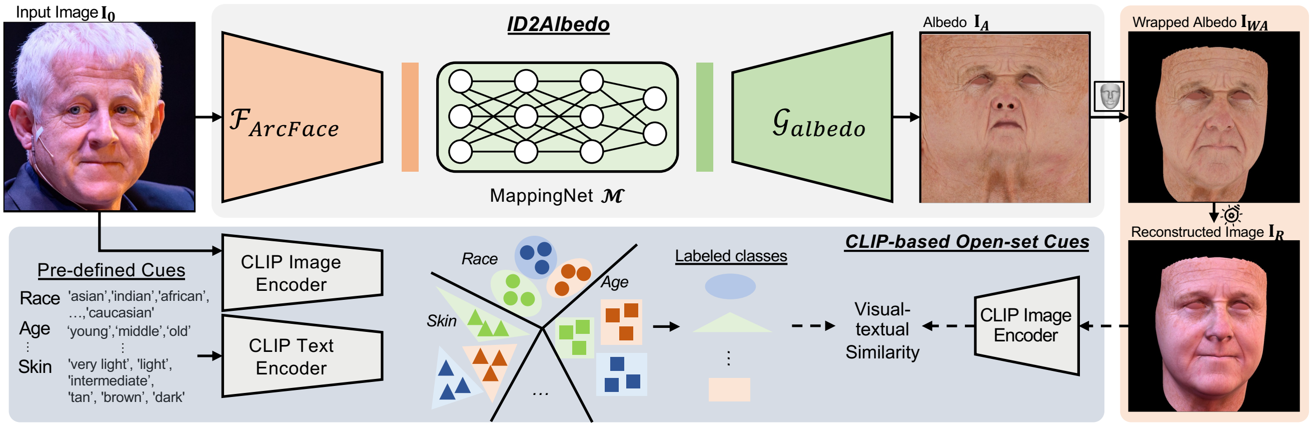

Figure 2. Overview of the proposed method *ID2Albedo*. We address realistic albedo estimates by tackling the light/albedo ambiguity using visual-textual cues. Given a facial input image, we first infer facial geometry by an off-the-shelf face model and extract the identity feature by a pre-trained ArcFace model. We then map the facial identity features to our pre-trained albedo generator latent space to achieve high-quality albedo maps. Combining the face shape, albedo maps are wrapped to image space and rendered using the predicted illumination (SH) coefficients. Besides, we pre-define several facial attribute cues and label each input image by CLIP visual-textual similarity. Finally, our rendered images are supervised by various text-based facial attributes and image-level losses in an end-to-end differentiable way.

model. Even though GANs can effectively model the distribution of a given training dataset, using hundred-level training data is not easy to train an expressive generative model. To this end, we consider compressing the training parameters of the GANs to avoid over-fitting and facilitate model training on the tiny image dataset (*i.e.* 142 facial albedo maps). As we target on high-resolution albedo generation (*i.e.* 1024×1024), the generator $\mathcal{G}$ can not be easily compressed. However, the discriminator $\mathcal{D}$, which takes the input images at the resolution of 1024^2, can be compressed. More specifically, we take advantage of a pre-trained ImageNet model $\mathcal{F}$, extract multi-level feature maps (*e.g.* $512^2, 256^2, 128^2, 64^2$) from both real images x and generated images $\mathcal{G}(z)$, and apply four independent discriminators $\{\mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5\}$ to the feature pyramid [36]. Instead of training one large discriminator on the 1024×1024 images, we simplify the training by introducing four smaller discriminators in a subspace spanned by the fixed ImageNet model $\mathcal{F}$. In this way, the parameter number significantly drops from 23.1M [31] to 10.3M. The proposed subspace-based GAN training can thus be formulated as follows,

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \sum_{p \in P} \Big(E_x[\log \mathcal{D}_p(\mathcal{F}_p(x))] + E_z[log(1 - \mathcal{D}_p(\mathcal{F}_p(\mathcal{G}(z))))] \Big), \tag{2}$$

where p indicates different feature levels and $\mathcal{F}$ is a pre-trained fixed ImageNet model mapping high-resolution images into four-scale feature pyramids. Since albedo maps are pixel-aligned across the whole data, we only employ flip augmentation during training without random cropping or translation. By optimizing Eq. 2, we obtain a high-resolution albedo generator, $\mathcal{G} : \mathbb{R}^{256} \rightarrow \mathbb{R}^{1024 \times 1024 \times 3}$.

3.2. Albedo Estimation via Identity Feature

In this paper, we target on high-resolution albedo estimation from "in-the-wild" face images captured under arbitrary poses, lighting conditions, and even occlusions. To this end, we choose an encode-decoder framework to consistently predict high-quality albedo. Specifically, we use a state-of-the-art face recognition network Φ (*i.e.* ArcFace [10]) to predict robust identity features and train a lightweight mapping network $\mathcal{M}$ to fine-tune the identity features, which are finally interpreted by our high-quality albedo decoder $\mathcal{G}$. That is:

$$z = \mathcal{M}(\Phi(\mathbf{I})), \tag{3}$$

where $\mathbf{I}$ is the input 2D face image and $z \in \mathbb{R}^{256}$ is the latent vector for the proposed albedo decoder.

The identity embedding network Φ is a ResNet-100 model trained on the large-scale WebFace dataset [59, 60] under the ArcFace loss [10, 11]. The pre-trained ArcFace model is able to extract face identity features that are robust to illumination, rotation, and occlusion. Therefore, the proposed albedo estimation can easily handle these face appearance variations in the wild. The lightweight mapping network $\mathcal{M}$ consists of three MLP layers with leaky ReLU as the activation function and a final linear output layer. After training, the mapping network $\mathcal{M}$ modifies the feature distribution of the original identity features to match the latent space of our albedo generator.

3.3. Albedo Disambiguation by Visual-Textual Cues

Assuming that the face is a Lambertian surface, the rendered face image can be computed by

$$\mathcal{R} = \mathcal{A} \odot \mathcal{S}, \tag{4}$$

where $\mathcal{R}$ stands for the final rendered image, $\mathcal{A}$ and $\mathcal{S}$ represent the wrapped face albedo and the shading image, respectively. $\odot$ denotes the hadamard product. When there is a parallel estimation of both albedo and illumination, the ambiguity between albedo and illumination happens. For example, an African face image can be decomposed into both dark skin and bright illumination or light skin and dim illumination.

To alleviate this problem, we explore the face attribute priors (*e.g*. ethnicity, age, skin color, and gender) to reduce ambiguity during albedo estimation. For instance, the facial albedo of an African person is likely to be dark, while the facial albedo of a Caucasian person is likely to be light. Besides, different ages also affect the shade of albedo. However, existing face datasets lack fine-grained facial attribute labels (*e.g*. skin color), and accurate manual annotations can be expensive. In addition, the multi-attribute estimation may involve many independent models, such as the race model, the age model, and the skin color model.

In this paper, we take advantage of the vision-language model, *i.e*. a pre-trained CLIP [42] network, to introduce a flexible attribute constraint for albedo disambiguation. Specifically, we first pre-define multiple face attributes, *e.g*. race, age, skin color, gender, and wrinkles. Then, for each facial attribute, we design a group of query texts. For example, we have "Caucasian", "Asian", "Indian" and "African" for the attribute of race, and "baby", "young", "adult" and "old" for the attribute of age. Afterward, we employ the text encoder of the CLIP model to calculate the feature of these query texts. For any training face image, we can obtain multi-dimensional attribute labels by (1) comparing the image features predicted through the CLIP image encoder with all of these text features, and (2) selecting the maximum cosine similarity score as the corresponding attribute label. During the training phase, we can obtain the attribute predictions of the rendered face in the same way by using the CLIP image encoder. To constrain the attribute of the generated albedo, we employ a semantic attribute loss,

$$L_{sem} = \sum_{i=1}^{N} \|\mathcal{L}_i - \mathcal{L}_i^*\|_2 \, , \tag{5}$$

where $\mathcal{L}_i$ is the predicted attribute similarity, $\mathcal{L}_i^*$ is the pseudo-label of input image attribute, N is the number of attributes we want to constrain. The attribute discrepancy between the input image and the rendered image can be back-propagated through the fixed CLIP image encoder and the differentiable renderer to update the parameters of the albedo estimation network.

3.4. Overall Loss

We first train the albedo generator $\mathcal{G}$ on the 142 high-resolution albedo maps (Sec. 3.1) and then train the ID2Albedo pipeline (Fig. 2) on the in-the-wild 2D face dataset. Given a training image $\mathbf{I}$, we compute the identity feature by the ArcFace model [10], project it into the latent space of $\mathcal{G}$ by the mapping network $\mathcal{M}$, and then generate high-quality albedo. Meanwhile, we define an illumination network $\mathcal{F}_{illumination}$ and employ an off-the-shelf shape network [12] to predicte facial illumination, shape, camera pose(Details in Sec. 4.1). Combining above predictions, we can warp the albedo to image space and render the face $\mathbf{I_R}$. Apart from the semantic attribute loss (Eq. 5), we also employ the following photometric loss, identity loss, and perceptual loss.

The photometric loss is calculated as

$$L_{photo} = \mathbf{M_{mask}} \cdot \|\mathbf{I} - \mathbf{I_R}\|_1, \tag{6}$$

where $\mathbf{M_{mask}}$ is the face skin mask calculated by the off-the-shelf face parsing model [37].

The identity loss is the cosine identity distance between the input image and the rendered face:

$$L_{id} = 1 - \frac{\Phi(\mathbf{I}), \Phi(\mathbf{I_R})}{\|\Phi(\mathbf{I})\|_2 \cdot \|\Phi(\mathbf{I_R})\|_2}, \tag{7}$$

where Φ is the pre-trained ArcFace model.

The perceptual loss [58] is defined as follows:

$$L_{per} = \sum_{l} \|\omega_l \odot (\mathcal{F}_l(\mathbf{I}) - \mathcal{F}_l(\mathbf{I_R}))\|_2^2 \, , \tag{8}$$

where l denotes the different level of a pre-trained VGG model $\mathcal{F}$, and ω_l is the scaling factor.

The overall objective function is then defined by combining the above losses:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{photo} + \lambda_2 \mathcal{L}_{id} + \lambda_3 \mathcal{L}_{per} + \lambda_4 \mathcal{L}_{sem}, \tag{9}$$

where the balance hyper-parameters λ_1, λ_2, λ_3, and λ_4 are set as 2.0, 0.2, 1.0, and 0.5, respectively.

4. Experiments

In this section, we evaluate our albedo reconstruction algorithm in terms of unbiasedness and quality. We first give the implementation details (Sec. 4.1). We participate in the FAIR benchmark and compare our ID2Albedo with the state-of-the-art albedo reconstruction methods (Sec 4.2). Then, our method is tested on the in-the-wild data to ensure unbiasedness under harsh lighting, various poses, and dark skin tones (Sec. 4.3). Finally, we conduct ablation studies on the albedo generator, albedo encoder, and visual-textual cues to validate the efficacy of our method (Sec. 4.4).

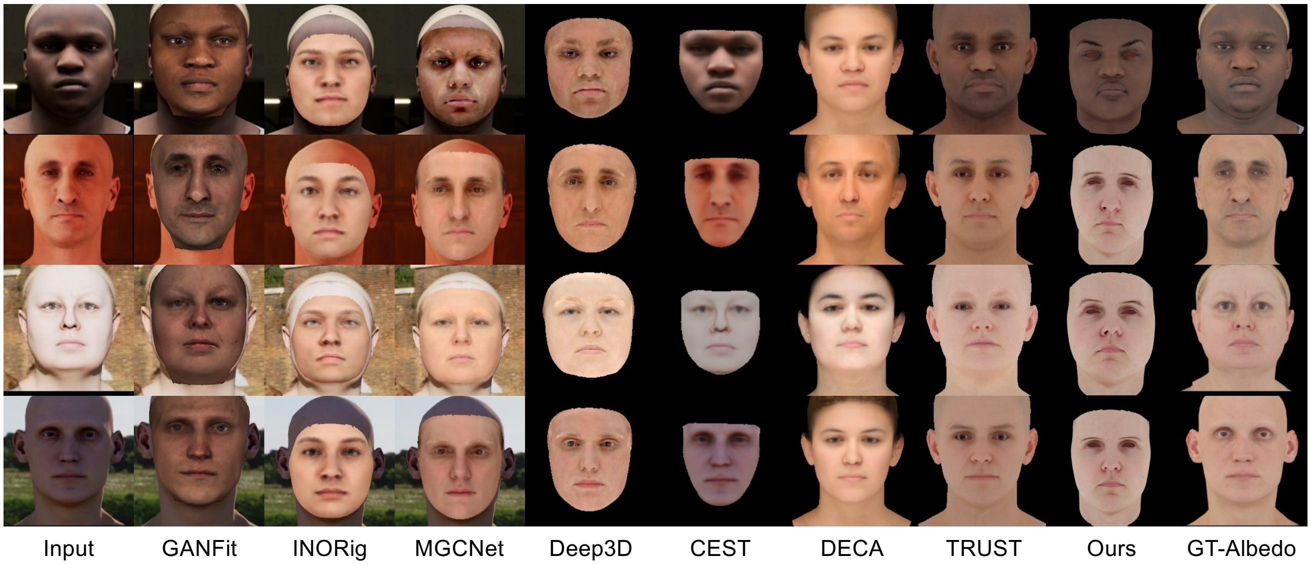

Figure 3. Comparison on the FAIR benchmark [16]. Please note that we don't have any FAIR benchmark albedo ground truth, so we choose the input same as TRUST. From left to right: input image, GANFIT [21], INORig [2], MGCNet [48], Deep3D [12], CEST [57], DECA [17], TRUST [16], ours and ground-truth albedo rendering.

Method	Avg. ITA $\downarrow$	Bias $\downarrow$	Score $\downarrow$	MAE $\downarrow$	ITA per skin type $\downarrow$					
					I	II	III	IV	V	VI
Deep3D [12]	22.57	24.44	47.02	27.98	**8.92**	9.08	8.15	10.90	28.48	69.90
GANFIT [21]	62.29	31.81	94.11	63.31	94.80	87.83	76.25	65.05	38.24	11.59
MGCNet [48]	21.41	17.58	38.99	25.17	19.98	12.76	8.53	9.21	22.66	55.34
DECA [17]	28.74	29.24	57.98	38.17	9.34	11.66	11.58	16.69	39.10	84.06
INORig [2]	27.68	28.18	55.86	33.20	23.25	11.88	**4.86**	9.75	35.78	80.54
CEST [57]	35.18	12.14	47.32	29.92	50.98	38.77	29.22	23.62	21.92	46.57
TRUST [16] (BFM)	16.19	15.33	31.52	21.82	12.44	**6.48**	5.69	9.47	16.67	46.37
TRUST [16] (AlbedoMM)	17.72	15.28	33.00	19.48	15.50	10.48	8.42	**7.86**	**15.96**	48.11
TRUST [16] (BalancedAlb)	13.87	**2.79**	**16.67**	**18.41**	11.90	11.87	11.20	13.92	16.15	18.21
Ours (ID2Albedo)	**12.07**	4.91	16.98	23.33	18.30	9.13	5.83	9.46	19.09	**10.59**

Table 1. Comparison to state-of-the-arts on the FAIR benchmark [16]. We utilize the FAIR official metrics, such as average ITA error, bias score (standard deviation), the total score (avg. ITA+Bias), mean average error, and average ITA score per skin type in degrees (I: very light, VI: very dark). Our method achieves accurate skin color predictions, especially on very dark skin.

Methods	M-SSIM$\uparrow$	LPIPS$\downarrow$	FID$\downarrow$	ID$\uparrow$
Deep3D [12]	0.73	0.1933	74.41	0.712
DECA [17]	0.61	0.2089	98.13	0.585
TRUST [16]	0.64	0.2112	97.37	0.603
Ours	**0.87**	**0.1549**	**45.56**	**0.867**

Table 2. Comparisons of our method with other albedo reconstruction methods on FFHQ, *e.g.*, Deep3D [12], DECA [17], and TRUST [16]. Given the absence of GT albedo, we compare the rendered image to the original image.

4.1. Implementation Details

All our implementations are based on PyTorch [39] and NVIDIA V100 cards. For the albedo generator, we employ adaptive discriminator augmentation as in [31]. We use Adam [33] as our optimizer with a learning rate of 1e-4, a batch size of 32, and a total number of iterations of 500K.

For *ID2Albedo*, we use the differentiable rasterizer from Pytorch3D [44] for rendering. We freeze the parameters of the geometry estimation network [12] and ArcFace [9, 10], and train the illumination network $\mathcal{F}_{illumination}$ and the mapping network $\mathcal{M}$. The pre-trained shape network [12] is based on the BFM [41] model and regresses the face identity $\alpha \in \mathbb{R}^{80}$, expression $\beta \in \mathbb{R}^{64}$, rotation $r \in \mathbb{R}^3$, translation vector $t \in \mathbb{R}^3$, respectively. We use spherical harmonics (SH) to approximate the illumination model. The illumination encoder uses the pre-trained Resnet-50 [25] as initialization and predicts 27 illumination coefficients. The mapping network $\mathcal{M}$ uses a fully connected architecture with random initialization. The input image size is 224 × 224 and the size of the albedo map is 1024 × 1024. We train it using Adam with a batch size of 8, an initial learning rate

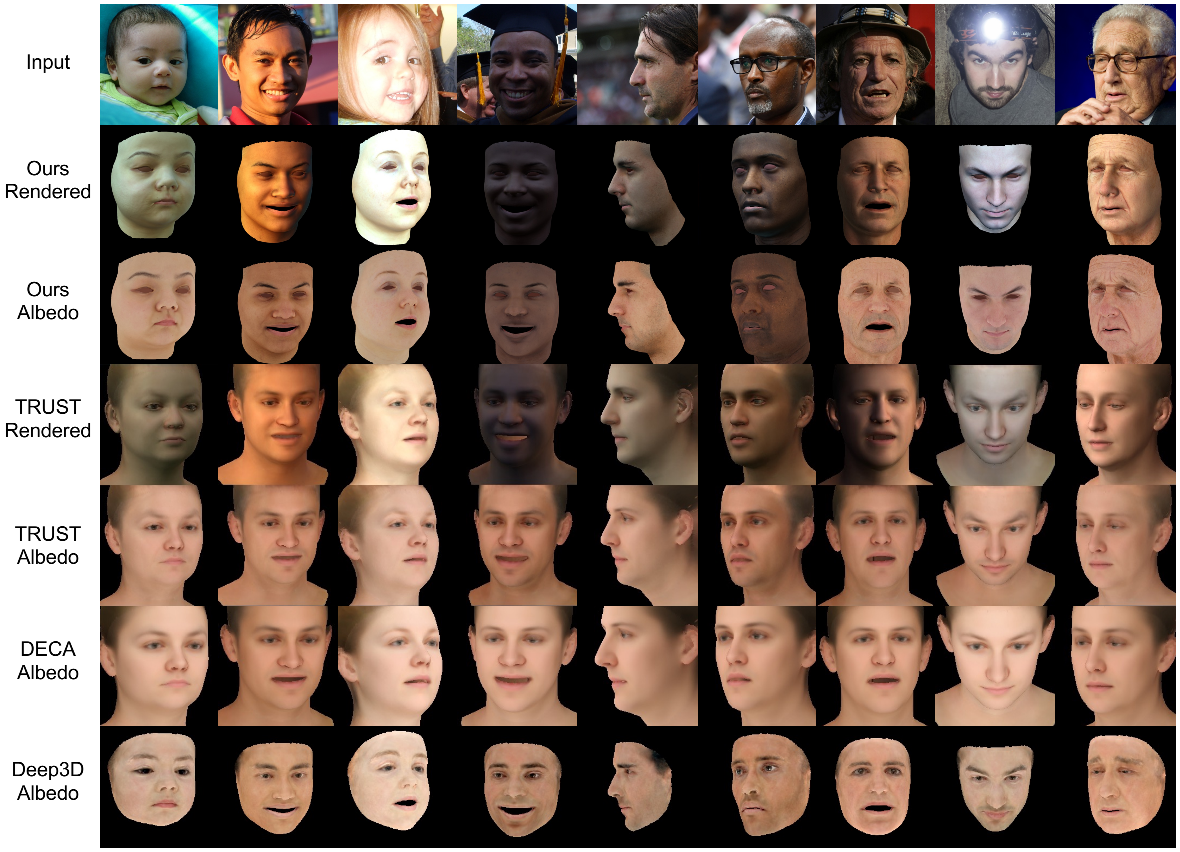

Figure 4. Comparisons on in-the-wild images. Input images are all from FFHQ [30]. From top to bottom: inputs, ours and TRUST [16] rendered and albedo images, DECA [17] and Deep3D [12] albedo images. We achieve the most realistic rendered results.

of 2e-5, and a total number of iterations of 50K. All the training process is on the SFHQ dataset [7], a high-quality synthetic dataset without data privacy concerns.

4.2. FAIR Benchmark Results

FAIR Benchmark [16] is constructed using 206 high-quality 3D head scans, and the Individual Typology Angle (ITA) score is recommended to classify skin tones into 6 categories. The ITA score is calculated as follows:

$$\mathrm{ITA}(L^*, b^*) = \frac{180}{\pi} \times \arctan(\frac{L^* - 50}{b^*}), \tag{10}$$

where L^* and b^* represent the lightness and yellow/blue components of the CIE L*a*b* color space, respectively. Furthermore, the bias score computes the standard deviation of the per-group ITA error, and the total score is the average of the top two scores.

Following TRUST [16], we perform a qualitative and quantitative evaluation on the FAIR benchmark, shown in Fig. 3 and Tab. 1, respectively. In Fig. 3, a common problem with current methods is a strong bias [2,21] towards specific skin types, or albedo models that limit the modeling of appropriate skin tone types [12, 48]. Both TRUST [16] and our method perform albedo estimation very well. Thanks to a powerful albedo generator, our method produces more realistic results. Numerically, our algorithm obtains the best results in ITA average score and very dark skin types, and is almost equal to TRUST in the overall score. In contrast, the rest of the algorithms are biased toward different skin types. Our method has a slightly higher error in type 1 and type 5 skin for different skin types since training data hardly includes the white skin tones. The network prefers to interpret the very light albedo as white skin tones in type 2 due to the uneven distribution of skin tones in training data. The same is true for type 5 light black skin. TRUST [16] achieves a minimal bias score because of semi-supervised learning. Overall, our algorithm makes good progress in ITA and achieves state-of-the-art, shown in Tab. 1.

4.3. Real-World Results

To evaluate our approach's robustness in real-world images, we qualitatively compare it with other methods on the FFHQ dataset, as shown in Fig. 4. The results show that our albedo achieves more realistic results while maintaining fairness. Furthermore, we compare with TRUST in various environments and poses on the same subject. TRUST under

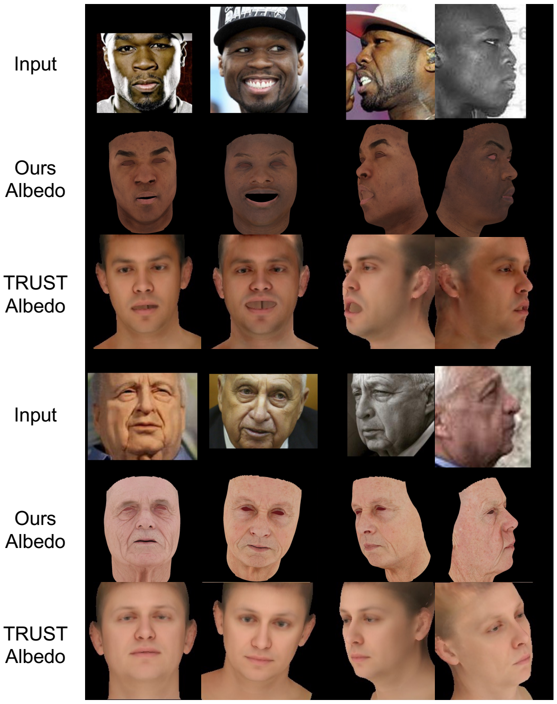

Figure 5. Comparisons on the real-world images from the same subject. Input images are all from CFP [47] dataset.

extreme poses results in irrational estimates, as illustrated in the third row of Fig. 5. While our method consistently generates unbiased, realistic albedo based on light-independent identity features, even in grayscale maps. We also analyze quantitative results in FFHQ. The results in Tab. 2 indicate that our method achieves the best scores in all the image-level metrics.

4.4. Ablation Study

Albedo Generator. We first verify the subspace-based GAN. We perform a comparison with origin StyleGAN [30] and StyleGANv2-ADA [29] on our aligned UV data. The FID results, shown in Tab. 3, indicate our subspace-based GAN can achieve better generation results.

ArcFace Encoder. Predicting albedo maps from real-world images relies on robust illumination-independent facial features. We train this module under different configurations to assess its utility, and the results are shown in Tab. 4. We observe that finetuning of partial layers or the entire pipeline results in large overfitting of the training data with significantly worse results. In contrast, we start with identity features, which can effectively perform the albedo reconstruction task.

Visual-Textual Cues. We investigate the benefits of CLIP-based visual-textual cues. We observe that ignoring visual-textual cues results in high skin color bias, whereas incorporating ethnographic cues results in significantly lower ITA and bias scores. A broader range of attributes produces better results, shown in Tab. 5.

Methods	StyleGAN [30]	StyleGANv2-ADA [29]	Ours
FID↓	64.7	49.3	42.2

Table 3. **Ablation study of albedo generator.** Our comparison includes StyleGAN [30] and StyleGANv2-ADA [29].

Albedo Encoder	Avg. ITA ↓	Bias ↓	Score ↓
ResNet-100 [25] (Scratch)	58.46	32.59	91.05
ResNet-100 [25] (ImageNet)	31.63	15.48	47.11
ArcFace [10] (fully trainable)	41.63	19.81	61.44
ArcFace [10] (L2 + L3 + L4)	28.75	11.87	40.62
ArcFace [10] (L3 + L4)	19.52	9.46	28.98
ArcFace [10] (L4)	14.58	6.79	21.37
ArcFace [10] (Frozen)(Ours)	13.46	5.86	19.32

Table 4. **Ablation study of albedo encoder.** We show a comparison with various alternative encoder configs, where $L2, L3, L4$ represents different network stages.

Configs	Avg. ITA ↓	Bias ↓	Score ↓
w/o any cues	25.66	23.51	49.17
Manual labeled races	18.13	10.46	28.59
CLIP [42] cues (only races)	16.21	7.44	23.65
CLIP cues all (ours)	13.46	5.86	19.32

Table 5. **Ablation study of CLIP-based cues.** We compare our method to the following alternatives: 1) no visual-textual cue, 2) using a labeled race dataset, and 3) using CLIP racial cues.

Furthermore, we attempt to use a hand-labeled ethnographic dataset, RFW [55], as direct labeling training. The results show that using manual ethnographic labels does result in an improvement, but the limited amount of data bounds further progress.

5. Conclusions

In this work, we propose an unbiased facial albedo reconstruction method based on the observation that intrinsic semantic attributes such as race, skin color, and age can constrain the albedo map. Our model estimates the albedo map directly from robust identity features rather than indirectly by predicting illumination. To achieve direct estimation, we define novel visual-textual cues as facial attributes to guide the albedo maps regression. The experiments demonstrate the proposed method achieves competitive performance on the FAIR benchmark and has excellent generalizability and fairness on real-world images. Our method can be used for high-quality reconstruction and rendering, which opens up new avenues for creating avatars faster and promoting the metaverse.

Acknowledgements We acknowledge Zijian Li and Weishuang Zhang for their help with realistic rendering. This work was supported in part by NSFC (62201342), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and the Fundamental Research Funds for the Central Universities.

References

[1] Oswald Aldrian and William AP Smith. Inverse rendering of faces with a 3D morphable model. *IEEE TPAMI*, 35(5):1080–1093, 2012. 2, 3

[2] Ziqian Bai, Zhaopeng Cui, Xiaoming Liu, and Ping Tan. Riggable 3D face reconstruction via in-network optimization. In *CVPR*, 2021. 6, 7

[3] Volker Blanz, Sami Romdhani, and Thomas Vetter. Face identification across different poses and illuminations with a 3D morphable model. In *FG*, 2002. 2

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

[5] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *ICCV*, 2019. 2

[6] Yongwei Chen, Rui Chen, Jiabao Lei, Yabin Zhang, and Kui Jia. Tango: Text-driven photorealistic and robust 3d stylization via lighting decomposition. In *NeurIPS*, 2022. 3

[7] Opensource Contributors. The synthetic faces high quality (sfhq) dataset. `https://github.com/SelfishGene/SFHQ-dataset`, 2022. 7

[8] 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. 3

[9] 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

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 2, 4, 5, 6, 8

[11] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, 2021. 4

[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. 2, 3, 5, 6, 7

[13] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3D morphable models and an illumination prior for face image analysis. *IJCV*, 126(12):1269–1287, 2018. 3

[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. *ACM TOG*, 39(5):157:1–157:38, 2020. 2

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

[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. 2, 3, 6, 7

[17] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM TOG*, 2021. 2, 3, 6, 7

[18] Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988. 3

[19] Kevin Frans, Lisa B Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *arXiv preprint arXiv:2106.14843*, 2021. 3

[20] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM TOG*, 41(4):1–13, 2022. 3

[21] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *CVPR*, 2019. 3, 6, 7

[22] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos P Zafeiriou. Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. *IEEE TPAMI*, 2021. 3

[23] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3D morphable model regression. In *CVPR*, 2018. 2

[24] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models - an open framework. In *FG*, 2018. 2

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 8

[26] Guosheng Hu, Pouria Mortazavian, Josef Kittler, and William Christmas. A facial symmetry prior for improved illumination fitting of 3d morphable model. In *ICB*. IEEE, 2013. 3

[27] Nikolay Jetchev. Clipmatrix: Text-controlled creation of 3d textured meshes. *arXiv preprint arXiv:2109.12922*, 2021. 3

[28] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv:1908.04913*, 2019. 2

[29] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020. 8

[30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 3, 7, 8

[31] 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, 4, 6

[32] Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt. Inverse-FaceNet: Deep monocular inverse face rendering. In *CVPR*, 2018. 2

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

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

[35] 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 TPAMI*, 2021. 3

[36] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 2, 4

[37] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. Roi tanh-polar transformer network for face parsing in the wild. *IVC*, 2021. 5

[38] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *CVPR*, 2022. 3

[39] 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

[40] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 3

[41] 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. Ieee, 2009. 2, 6

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021. 2, 3, 5, 8

[43] Ravi Ramamoorthi and Pat Hanrahan. A signal-processing framework for inverse rendering. In Lynn Pocock, editor, *SIGGRAPH*, pages 117–128, 2001. 3

[44] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Pytorch3d. https://github.com/facebookresearch/pytorch3d, 2020. 6

[45] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *CVPR*, 2022. 3

[46] Sandro Schönborn, Bernhard Egger, Andreas Morel-Forster, and Thomas Vetter. Markov chain monte carlo for automated face image analysis. *IJCV*, 123(2):160–183, 2017. 2

[47] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *WACV*, 2016. 8

[48] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3D face reconstruction by occlusion-aware multi-view geometry consistency. In *ECCV*, 2020. 2, 6, 7

[49] William A. P. Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua Tenenbaum, and Bernhard Egger. A morphable face albedo model. In *CVPR*, 2020. 2, 3

[50] Ayush Tewari, Michael Zollöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian. MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV*, 2017. 2

[51] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *ECCV*, 2020. 1

[52] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016. 1

[53] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2Face: Real-time face capture and reenactment of RGB videos. In *CVPR*, 2016. 2

[54] Thomas Vetter and Volker Blanz. Estimating coloured 3d face models from single images: An example based approach. In *ECCV*, 1998. 1

[55] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *ICCV*, 2019. 8

[56] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021. 1

[57] Yandong Wen, Weiyang Liu, Bhiksha Raj, and Rita Singh. Self-supervised 3D face reconstruction via conditional estimation. In *ICCV*, 2021. 6

[58] 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. 5

[59] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Dalong Du, Jiwen Lu, et al. Webface260m: A benchmark for million-scale deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 4

[60] 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. 4

[61] Michael Zollhöfer, Justus Thies, Darek Bradley, Pablo Garrido, Thabo Beeler, Patrick Péerez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. *EG*, 2018. 2