Improving Fairness in Facial Albedo Estimation via Visual-Textual Cues

Xingyu Ren  Jiankang Deng*  Chao Ma*  Yichao Yan  Xiaokang Yang
1MoE 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-
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 define 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 current albedo model still has the problem of imbalance be-
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 de-lighting. 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. Style-CLIP [46] 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. CLIP-Draw [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

\[
\min_{\mathcal{G}} \max_{\mathcal{D}} \mathbb{E}_x [\log D(x)] + \mathbb{E}_z [\log(1 - D(\mathcal{G}(z)))],
\]  

(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 \( P_z \) to the generated images \( \mathcal{G}(z) \). The discriminator \( \mathcal{D} \) then aims to discriminate real images \( x \sim P_x \) from generated images \( \mathcal{G}(z) \sim P_z \).

In this paper, we purchase 142 high-quality albedo maps from the 3D Scan Store to build our high-resolution albedo...
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 $G$ can not be easily compressed. However, the discriminator $D$, 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 $F$, extract multi-level feature maps (e.g. 512$^2$, 256$^2$, 128$^2$, 64$^2$) from both real images $x$ and generated images $G(z)$, and apply four independent discriminators $\{D_2, D_3, D_4, D_5\}$ 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 $F$. 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,

$$\min_{G} \max_{D} \sum_{p \in P} \left( E_{x} [\log D_{p}(F_{p}(x))] + E_{z} [\log (1 - D_{p}(F_{p}(G(z))))] \right),$$

where $p$ indicates different feature levels and $F$ 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, $G : \mathbb{R}^{256} \rightarrow \mathbb{R}^{1024 \times 1024 \times 3}$.

### 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 $\Phi$ (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 $G$. That is:

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

where $I$ 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 $\Phi$ 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

\[ R = A \odot S, \]

where \( R \) stands for the final rendered image, \( A \) and \( S \) 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,

\[ L_{\text{sem}} = \sum_{i=1}^{N} \| L_i - L_i^* \|_2, \]

where \( L_i \) is the predicted attribute similarity, \( L_i^* \) 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 \( G \) on the 142 high-resolution albedo maps (Sec. 3.1) and then train the ID2Albedo pipeline (Fig. 2) on the in-the-wild 2D dataset. Given a training image \( I \), we compute the identity feature by the ArcFace model [10], project it into the latent space of \( G \) by the mapping network \( M \), and then generate high-quality albedo. Meanwhile, we define an illumination network \( F_{\text{illumination}} \) and employ an off-the-shelf shape network [12] to predict 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 \( I_R \).

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

\[ L_{\text{photo}} = M_{\text{mask}} \cdot \| I - I_R \|_1, \]

where \( M_{\text{mask}} \) 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:

\[ L_{\text{id}} = 1 - \frac{\Phi(I), \Phi(I_R)}{\| \Phi(I) \|_2 \cdot \| \Phi(I_R) \|_2}, \]

where \( \Phi \) is the pre-trained ArcFace model.

The perceptual loss [58] is defined as follows:

\[ L_{\text{per}} = \sum_l \| \omega_l \odot (F_l(I) - F_l(I_R)) \|_2^2, \]

where \( l \) denotes the different level of a pre-trained VGG model \( F \), and \( \omega_l \) is the scaling factor.

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

\[ L_{\text{total}} = \lambda_1 L_{\text{photo}} + \lambda_2 L_{\text{id}} + \lambda_3 L_{\text{per}} + \lambda_4 L_{\text{sem}}, \]

where the balance hyper-parameters \( \lambda_1, \lambda_2, \lambda_3, \) and \( \lambda_4 \) 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).
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg. ITA ↓</th>
<th>Bias ↓</th>
<th>Score ↓</th>
<th>MAE ↓</th>
<th>ITA per skin type ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>III</td>
<td>IV</td>
<td>V</td>
</tr>
<tr>
<td>Deep3D [12]</td>
<td>22.57</td>
<td>24.44</td>
<td>47.02</td>
<td>27.98</td>
<td>8.92</td>
</tr>
<tr>
<td>GANFIT [21]</td>
<td>62.29</td>
<td>31.81</td>
<td>94.11</td>
<td>63.31</td>
<td>94.80</td>
</tr>
<tr>
<td>DECA [17]</td>
<td>28.74</td>
<td>29.24</td>
<td>57.98</td>
<td>38.17</td>
<td>9.34</td>
</tr>
<tr>
<td>INORig [2]</td>
<td>27.68</td>
<td>28.18</td>
<td>55.86</td>
<td>33.20</td>
<td>23.25</td>
</tr>
<tr>
<td>CEST [57]</td>
<td>35.18</td>
<td>12.14</td>
<td>47.32</td>
<td>29.92</td>
<td>50.98</td>
</tr>
<tr>
<td>TRUST [16] (AlbedoMM)</td>
<td>17.72</td>
<td>15.28</td>
<td>33.00</td>
<td>19.48</td>
<td>15.50</td>
</tr>
<tr>
<td>TRUST [16] (BalancedAlb)</td>
<td>13.87</td>
<td>2.79</td>
<td>16.67</td>
<td>18.41</td>
<td>11.90</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>M-SSIM↑</th>
<th>LPIPS↓</th>
<th>FID↓</th>
<th>ID↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep3D [12]</td>
<td>0.73</td>
<td>0.1933</td>
<td>74.41</td>
<td>0.712</td>
</tr>
<tr>
<td>DECA [17]</td>
<td>0.61</td>
<td>0.2089</td>
<td>98.13</td>
<td>0.585</td>
</tr>
<tr>
<td>TRUST [16]</td>
<td>0.64</td>
<td>0.2112</td>
<td>97.37</td>
<td>0.603</td>
</tr>
<tr>
<td>Ours</td>
<td>0.87</td>
<td>0.1549</td>
<td>45.56</td>
<td>0.867</td>
</tr>
</tbody>
</table>

Table 4. Comparison 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 \( F_{\text{illumination}} \) and the mapping network \( \mathcal{M} \). The pre-trained shape network [12] is based on the BFM [41] model and regresses the face identity \( \alpha \in \mathbb{R}^{10} \), expression \( \beta \in \mathbb{R}^{44} \), rotation \( r \in \mathbb{R}^{4} \), 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 of 1e-3, 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 \( F_{\text{illumination}} \) and the mapping network \( \mathcal{M} \). The pre-trained shape network [12] is based on the BFM [41] model and regresses the face identity \( \alpha \in \mathbb{R}^{10} \), expression \( \beta \in \mathbb{R}^{44} \), rotation \( r \in \mathbb{R}^{4} \), 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 of 1e-3, a batch size of 32, and a total number of iterations of 500K.
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:

\[ \text{ITA}(L^*, b^*) = \frac{180}{\pi} \times \arctan\left(\frac{L^* - 50}{b^*}\right), \]

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.
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.

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


[37] Yiming Lin, Jie Shen, Yujiang Wang, and Majed Pantic. Roi tahn-polar transformer network for face parsing in the wild.


