De-biasing Facial Albedo Estimation via Visual-Textual Cues

XINGYU REN, Shanghai Jiao Tong University, China
JIANKANG DENG, Imperial College London, United Kingdom
CHAO MA*, Shanghai Jiao Tong University, China
YICHAO YAN, Shanghai Jiao Tong University, China
WENHAN ZHU, Shanghai Jiao Tong University, China
XIAOKANG YANG, Shanghai Jiao Tong University, China

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 an ID2Albedo++ method to directly estimate albedo without constraining illumination. The cornerstone of our approach is the utilization of inherent semantic attributes like race, skin color, and age to restrict the albedo map. We 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 wrinkle. 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 ensure that during the training phase, the generated albedos possess attributes consistent with their input counterparts. 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. Our ID2Albedo++ is trained in a self-supervised way and outperforms state-of-the-art albedo estimation methods in terms of both accuracy and fidelity. It is worth mentioning that our approach has excellent generalizability and fairness, especially on in-the-wild data. Moreover, our framework can be smoothly transitioned to ID2Texture, thereby greatly benefiting pose invariant face recognition.

JAIR Associate Editor: Kai-Wei Chang

JAIR Reference Format:
Xingyu Ren, Jiankang Deng, Chao Ma, Yichao Yan, Wenhan Zhu, and Xiaokang Yang. 2026. De-biasing Facial Albedo Estimation via Visual-Textual Cues. *Journal of Artificial Intelligence Research* 86, Article 23 (July 2026), 30 pages. DOI: 10.1613/jair.1.17875

1 Introduction

3D face reconstruction is one of the fundamental problems in computer vision and computer graphics. Its primary goal involves deriving detailed 3D facial shapes and appearances from 2D images, utilizing either multi-view or single-view images. 3D face reconstruction plays a vital role in numerous vision applications, such as face manipulation (Thies, Zollhofer, et al. 2016), speech-driven facial animation (Thies, Elgharib, et al. 2020), and video

*Corresponding Author.

Authors' Contact Information: Xingyu Ren, ORCID: 0000-0002-8719-878X, rxy_sjtu@sjtu.edu.cn, Shanghai Jiao Tong University, Shanghai, China; Jiankang Deng, ORCID: 0000-0002-3709-6216, j.deng16@imperial.ac.uk, Imperial College London, London, United Kingdom; Chao Ma, ORCID: 0000-0002-8459-2845, chaoma@sjtu.edu.cn, Shanghai Jiao Tong University, Shanghai, China; Yichao Yan, ORCID: 0000-0003-3209-8965, yanyichao@sjtu.edu.cn, Shanghai Jiao Tong University, Shanghai, China; Wenhan Zhu, ORCID: 0000-0001-8781-1110, zhuwenhan823@sjtu.edu.cn, Shanghai Jiao Tong University, Shanghai, China; Xiaokang Yang, ORCID: 0000-0003-4029-3322, xkyang@sjtu.edu.cn, Shanghai Jiao Tong University, Shanghai, China.

DOI: 10.1613/jair.1.17875

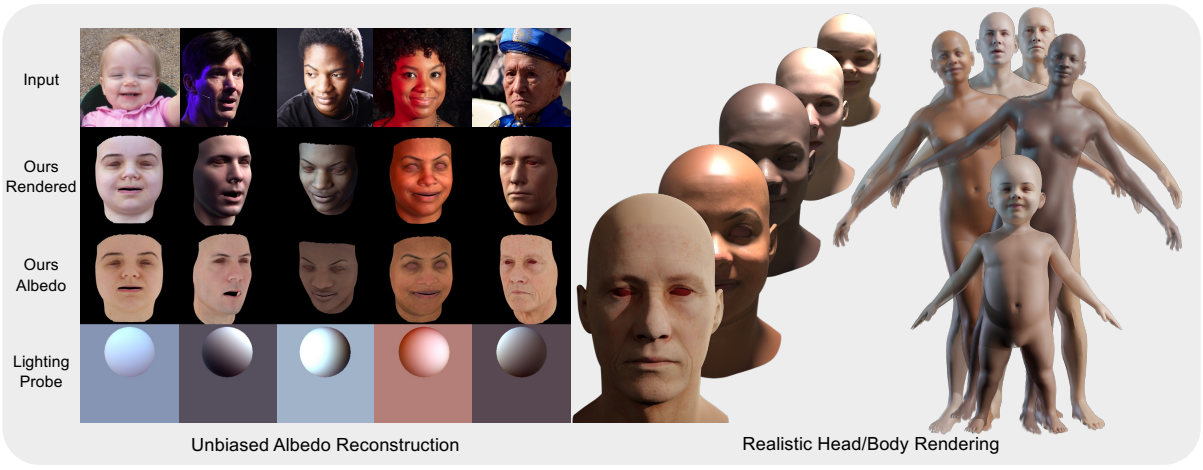

Fig. 1. We introduce ID2Albedo, a high-quality, unbiased albedo reconstruction method. ID2Albedo maps the facial identity features to the latent space of the albedo generator and uses novel visual-textual cues to constrain albedo attributes. Our approach can alleviate the illumination/albedo ambiguity and generate high-fidelity albedo maps for realistic rendering. Images are all from FFHQ (Karras, Laine, and Aila 2019) dataset.

conferencing (T.-C. Wang et al. 2021). Since the pioneering work of 3D Morphable Model (3DMM)(Vetter and Blanz 1998), monocular face reconstruction methods have made remarkable progress due to their high speed and remarkable geometric accuracy. To enable more realistic applications such as avatar creation, interactive AR/VR, etc., fine-grained albedo reconstruction attracts considerable attention (H. Feng et al. 2022).

Facial albedo refers to the intrinsic skin reflectance independent of illumination. Racial bias arises because darker skin tends to be incorrectly interpreted as illumination shadows, causing existing methods to over-lighten or distort the albedo. 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 (Gerig et al. 2018; Paysan et al. 2009; W. A. P. Smith et al. 2020), and 2) introducing additional lighting constraints to reduce ambiguity (Aldrian and W. A. Smith 2012; Y. Deng et al. 2019; Y. Feng et al. 2021). 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 (H. Feng et al. 2022) re-built 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 (H. Feng et al. 2022) 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 (J. Deng, J. Guo, Xue, et al. 2019) and the face attribute analysis model (Kärkkäinen and Joo 2019)? 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 (W. A. P. Smith et al. 2020) fails in reconstructing high-frequency facial details. Given hundred-level training data, high-resolution Generative Adversarial Networks (GANs) (Karras, Laine, Aittala, et al. 2020) 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 (T.-Y. Lin et al. 2017)

produced by a fixed ImageNet model. To reduce flicker artifacts caused by limited real samples, we inject random Gaussian noise into the discriminator feature pyramid. Based on our high-resolution albedo generator, we further utilize the illumination-robust identity features (J. Deng, J. Guo, Xue, et al. 2019) 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 (eg. 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 explore 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 (Radford et al. 2021), to provide semantic cues for individual faces. Specifically, we predefine diverse texts from various perspectives, including races, skin tones, ages, wrinkles, gender, 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:

(1) 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.
(2) We employ visual-textual cues in the face reconstruction framework to overcome the illumination/albedo ambiguity problem by constraining facial semantic attributes.
(3) The proposed method improves the accuracy and fairness of facial albedo estimation, achieving state-of-the-art performance on the FAIR benchmark.

This work is an extension of our conference paper (Ren, J. Deng, Ma, et al. 2023) accepted by CVPR 2023 as the highlight. Compared to the preliminary conference paper, our journal version makes the following extra contributions: 1) a well-trained high-fidelity structure-aligned albedo generator; 2) two additional applications (ie., identity-guided texture completion and realistic head/human creation) based on the proposed method; 3) an extensive analysis, ablation study, and experiments for the aforementioned additions.

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 (Aldrian and W. A. Smith 2012; Blanz, Romdhani, et al. 2002; Blanz and Vetter 1999; Schönborn et al. 2017; Thies, Zollhöfer, et al. 2016) and regression-based approaches (A. Chen et al. 2019; Y. Deng et al. 2019; Y. Feng et al. 2021; Genova et al. 2018; H. Kim et al. 2018; Shang et al. 2020; Tewari et al. 2017). More details are described in (Egger, W. A. P. Smith, et al. 2020; Zollhöfer et al. 2018). 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 refer to the survey (Egger, W. A. P. Smith, et al. 2020) for more information. The widely used Basel Face Model (BFM) (Paysan et al. 2009) 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. (W. A. P. Smith et al. 2020) proposed AlbedoMM by creating an expressive albedo model from varied light-stage data and simultaneously modeled the diffuse and specular albedos to increase the diversity of the appearance space. Based on AlbedoMM (W. A. P. Smith et al. 2020), the recent TRUST (H. Feng et al. 2022) 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. (J. Deng, S. Cheng, et al. 2018) trained a generative adversarial network to reconstruct textures from a single image. Gecer et al. (Gecer, Ploumpis, Kotsia, and S. Zafeiriou 2019; Gecer, Ploumpis, Kotsia, and S. P. Zafeiriou 2021) 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. (Lattas, Moschoglou, Gecer, et al. 2020; Lattas, Moschoglou, Ploumpis, et al. 2021) trained an image-to-image translation network with light-stage data to synthesize diffuse/specular albedo from high-quality textures.

Recent works continue to explore high-fidelity facial albedo estimation through larger datasets and stronger priors. Ren et al. (Ren, J. Deng, Y. Cheng, J. Guo, et al. 2024) introduced ID2Reflectance, an identity-conditioned reflectance reconstruction framework that learns image-space reflectance priors and employs codebook fusion to produce relightable appearance from in-the-wild inputs. While this approach estimates more complete reflectance representations (beyond diffuse albedo), its reliance on identity features limits its robustness when facial appearance deviates from identity embeddings. Hokari et al. (Hokari et al. 2025) proposed a two-shot framework that employs an albedo-conditioned diffusion model to recover facial geometry together with SVBRDF components. Although the method produces high-quality reconstructions from merely two input views, its reliance on multi-view observations restricts its applicability in unconstrained single-image settings. More recently, Ren et al. (Ren, J. Deng, Y. Cheng, W. Zhu, et al. 2025) presented S^3Face, a diffusion-prior-guided method targeting subsurface-scattering-compliant facial reflectance reconstruction, incorporating hemoglobin and melanin cues to better model physiologically faithful skin appearance. However, the method requires multiple reflectance components and introduces additional biological priors, making it computationally heavier compared to our albedo-focused framework.

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 is an ill-posed problem (Ramamoorthi and Hanrahan 2001). Although the appearance prior has constrained the albedo variation, it does not completely eliminate the ambiguity problem. The common idea is to find stronger prior knowledge to constrain both. Hu et al. (Hu et al. 2013) normalized the symmetry of albedo. Aldrian et al. (Aldrian and W. A. Smith 2012) regularized light by imposing a "gray world" constraint that constrains light to be monochromatic, and subsequent works such as (Y. Deng et al. 2019; Y. Feng et al. 2021) used a similar regularization approach for approximate decomposition. Egger et al. (Egger, Schönborn, et al. 2018) took into account the existence of a certain distribution of illumination and directly learned a statistical prior for the SH coefficients. TRUST (H. Feng et al. 2022) extended the range of light estimation by decomposing light into face light and ambient light and used 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 (Radford et al. 2021). CLIP learned a joint embedding space for images and text. StyleCLIP (Patashnik et al. 2021) leveraged pre-trained StyleGAN (Karras, Laine, and Aila 2019; Karras, Laine, Aittala, et al. 2020) for CLIP-guided image modification. VQGAN-CLIP (Esser et al. 2021)

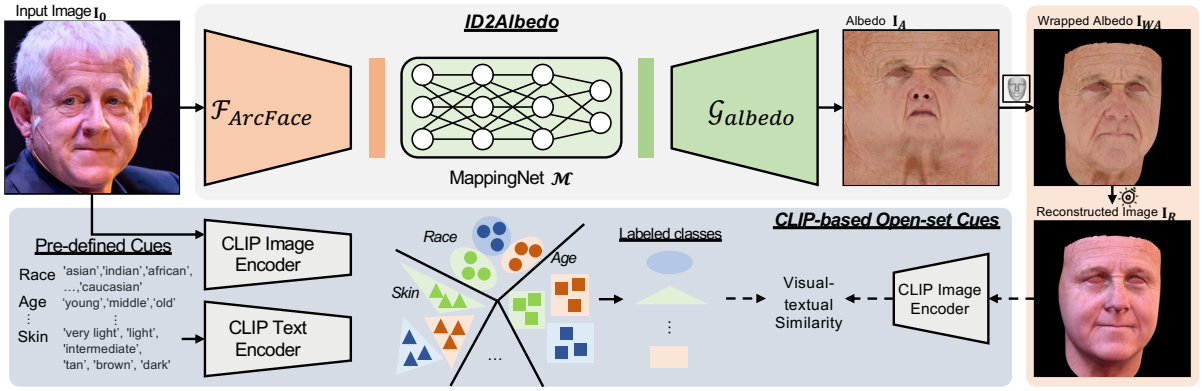

Fig. 2. Overview of the proposed method ID2Albedo. We address realistic albedo estimates by tackling the light/albedo ambiguity using visual-textual cues. Given a facial input image, we first infer facial geometry by an off-the-shelf face model and extract the identity feature from a pre-trained ArcFace 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.

employed CLIP for text-guided image generation. In the stylization domain, Gal et al. (Gal et al. 2022) used CLIP to fine-tune a pre-trained StyleGAN for images. Based on a textual question, Text2Mesh (Michel et al. 2022) predicted color and geometry details for a specified template mesh. In an implicitly differentiable rendering framework, TANGO (Y. Chen et al. 2022) employed CLIP to improve the physical attributes of objects for more realistic stylization. In the area of generation, Sanghi et al. (Sanghi et al. 2022) utilized CLIP for unconditional 3D voxel generation. CLIP-Draw (Frans et al. 2022) produced 2D vector graphics for drawing styles using textual instruction. Jetchev et al. (Jetchev 2021) 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

The primary objective of this research is to reconstruct 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) by feature pyramid augmentation (Sec. 3.2) and design an albedo estimation method based on a pre-trained face recognition model (Sec. 3.3). To reduce the ambiguity of albedo estimation, we explore the semantic facial attribute constraints through visual-textual cues (Sec. 3.4). 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. 3.5).

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 (W. A. P. Smith et al. 2020), a novel light stage 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 (Fitzpatrick 1988), as shown in Fig. 3. Based on the limited

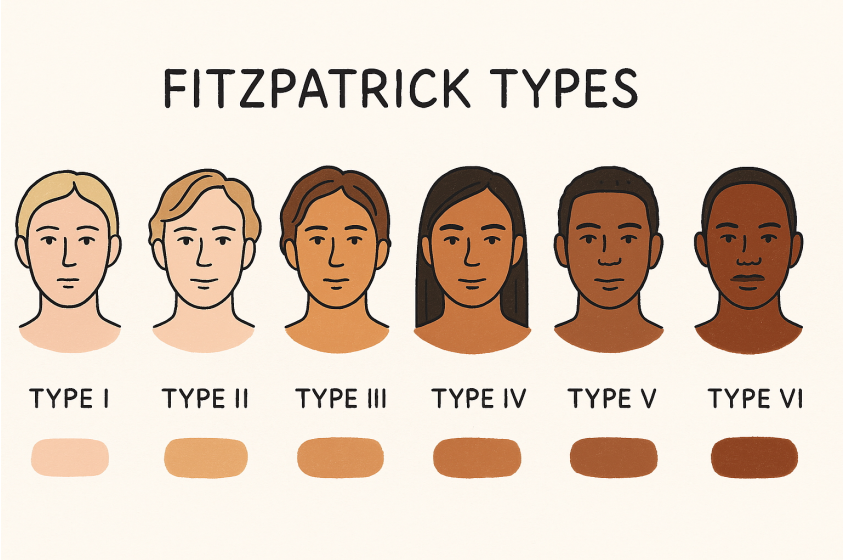

Fig. 3. Fitzpatrick skin type classification. Six skin types are shown from lightest (Type I) to darkest (Type VI), providing reference for our albedo estimation experiments and bias analysis.

albedo training data, a morphable face albedo model (W. A. P. Smith et al. 2020) 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 (eg. wrinkles and pores). Recently, Generative Adversarial Networks (GANs) (Karras, Laine, Aittala, et al. 2020) have shown excellent ability in capturing image details. Specifically, GANs aim to optimize the following minimax objective

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

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

In this paper, we purchase 142 high-quality albedo maps from the 3D Scan Store[1] 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 (ie. 142 facial albedo maps). As we target high-resolution albedo generation (ie. 1024×1024), the generator $\mathscr{G}$ can not be easily compressed. However, the discriminator $\mathscr{D}$, 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 $\mathscr{F}$, extract multi-level feature maps (eg. $512^2, 256^2, 128^2, 64^2$) from both real images x and generated images $\mathscr{G}(z)$, and apply four independent discriminators $\{\mathscr{D}_2, \mathscr{D}_3, \mathscr{D}_4, \mathscr{D}_5\}$ to the feature pyramid (T.-Y. Lin et al. 2017). 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 $\mathscr{F}$. In this way, the parameter number significantly drops from 23.1M (Karras, Laine, Aittala, et al. 2020) to 10.3M. The proposed subspace-based

[1] https://www.3dscanstore.com/

GAN training can thus be formulated as follows,

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

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

3.2 Augment Feature Pyramid via Gaussian Noise

The subspace discriminator proposed in Sec. 3.1 makes it possible to train a high-quality albedo generator from 142 albedo maps. However, we find that the current generator produces flicker and unpleasant artifacts, resulting in poor Frechet Inception Distance values, as shown in Fig. 11. We argue that the cause of the above phenomenon is that the discriminator still tends to overfit. Although the number of trainable parameters has been reduced to alleviate the overfitting of the discriminator, the training procedure of GAN is still very challenging given only more than a hundred real data.

Inspired by the adaptive augmentation strategy proposed in StyleGAN2-ADA (Karras, Aittala, et al. 2020) for limited data, we examine how to add augmentation techniques to our training pipeline to achieve better results. Karras et al. (Karras, Aittala, et al. 2020) employ several augmentations for real-world images, such as geometric transformation, color transformation, image space filtering, additional noise, and cropping, while these cannot be directly used in our training pipeline. This is because the former only needs to maintain consistent image semantics, whereas albedo maps also need to keep consistent image structure and reasonable albedo values. For example, color transformation tricks will corrupt the ground truth albedo values, and geometric transformation tricks (eg. rotation) will mislead the albedo structure. Therefore, we further consider augmentations in the feature subspace.

Bora et al. (Bora et al. 2018) prove that a Generative Adversarial Network can counteract augmentation-induced distortion during training and accurately identify the distribution if the augmentation process is depicted as a reversible transformation of the data space probability distribution. Consider that popular denoising methods (Ho et al. 2020; J. Song et al. 2020; Y. Song et al. 2021) use a reversible noise representation, in which the diffusion process can be seen as a Markov Chain that starts from the original sample and gradually erases its information until reaching a noise level σ^2 after T steps. The core idea is that the magnitude of each noise distortion is small (both forward and backward) so that the noise is always a Gaussian distribution. Therefore, we introduce the forward noise diffusion process into the multi-scale feature subspace of the discriminator.

We extract multi-level feature pyramids $\mathbf{f}$ from both real images x and generated images $\mathscr{G}(z)$, where $\mathbf{f}_r = \mathscr{F}(x)$ and $\mathbf{f}_g = \mathscr{F}(\mathscr{G}(z))$. We gradually adds noise to the feature sample $\mathbf{f} \sim q(\mathbf{f}_0)$ in T steps with pre-defined variance schedule β and variance σ, which can be formulated as:

$$\begin{aligned} q(\mathbf{f}_{1:T} \,|\, \mathbf{f}_0) &= \textstyle\prod_{t=1}^{T} \pi_t q(\mathbf{f}_t \,|\, \mathbf{f}_{t-1}), \\ q(\mathbf{f}_t \,|\, \mathbf{f}_{t-1}) &= \mathcal{N}(\mathbf{f}_t; \sqrt{1-\beta_t}\mathbf{f}_{t-1}, \beta_t \sigma^2 \mathbf{I}), \end{aligned} \tag{3}$$

where the mixture weights π_t are non-negative sum to one, and $\mathbf{f}_t$ at an arbitrary time-step t has a closed form as

$$q(\mathbf{f}_t \,|\, \mathbf{f}_0) = \mathcal{N}(\mathbf{f}_t; \sqrt{\bar{\alpha}_t}\mathbf{f}_0, (1-\bar{\alpha}_t)\sigma^2 \mathbf{I}), \tag{4}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. We mark the noisy sample $\mathbf{f}_t$ as $\mathbf{n}$, and the final noise distribution can be expressed as follows:

$$q(\mathbf{n}_r \mid \mathscr{F}_p(x), t) = \mathscr{N}(\mathbf{n}_r; \sqrt{\bar{\alpha}_t}\mathscr{F}_p(x), (1-\bar{\alpha}_t)\sigma^2\mathbf{I}), \tag{5}$$

$$q(\mathbf{n}_g \mid \mathscr{F}_p(\mathscr{G}(z)), t) = \mathscr{N}(\mathbf{n}_g; \sqrt{\bar{\alpha}_t}\mathscr{F}_p(\mathscr{G}(z)), (1-\bar{\alpha}_t)\sigma^2\mathbf{I}). \tag{6}$$

Next, we use noised feature $\mathbf{n}$ instead of $\mathbf{f}$ for optimization. In this way, we can augment features in the subspace. Our loss function can be updated as follows:

$$\begin{aligned}\min_{\mathscr{G}} \max_{\mathscr{D}} \sum_{p \in P} \Big(& E_{x,t\sim p_\pi, n_r \sim q_{(\mathbf{n}\mid \mathscr{F}_p(x),t)}}[\log \mathscr{D}_p(\mathbf{n}_r)] \\ & + E_{z,t\sim p_\pi, n_g \sim q_{(\mathbf{n}\mid \mathscr{F}_p(\mathscr{G}(z)),t)}}[log(1-\mathscr{D}_p(\mathbf{n}_g))]\Big),\end{aligned} \tag{7}$$

3.3 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 (ie. ArcFace (J. Deng, J. Guo, Xue, et al. 2019)) to predict robust identity features and train a lightweight mapping network $\mathscr{M}$ to fine-tune the identity features, which are finally interpreted by our high-quality albedo decoder $\mathscr{G}$. That is:

$$z = \mathscr{M}(\Phi(\mathbf{I})), \tag{8}$$

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 (Z. Zhu, G. Huang, J. Deng, Ye, J. Huang, X. Chen, J. Zhu, T. Yang, Du, et al. 2022; Z. Zhu, G. Huang, J. Deng, Ye, J. Huang, X. Chen, J. Zhu, T. Yang, Lu, et al. 2021) under the ArcFace loss (J. Deng, J. Guo, Xue, et al. 2019; J. Deng, J. Guo, J. Yang, et al. 2021). 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 $\mathscr{M}$ consists of three MLP layers with leaky ReLU as the activation function and a final linear output layer. After training, the mapping network $\mathscr{M}$ modifies the feature distribution of the original identity features to match the latent space of our albedo generator.

3.4 Albedo Disambiguation by Visual-Textual Cues

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

$$\mathscr{R} = \mathscr{A} \odot \mathscr{S}, \tag{9}$$

where $\mathscr{R}$ stands for the final rendered image, $\mathscr{A}$ and $\mathscr{S}$ 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 (eg. 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 (eg. 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, ie. a pre-trained CLIP (Radford et al. 2021) network, to introduce a flexible attribute constraint for albedo disambiguation. Specifically, we first pre-define multiple face attributes, eg. race, age, skin color, gender, and wrinkle. 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,

$$L_{sem} = \sum_{i=1}^{N} \left\| \mathscr{L}_i - \mathscr{L}_i^* \right\|_2, \tag{10}$$

where $\mathscr{L}_i$ is the predicted attribute similarity, $\mathscr{L}_i^*$ is the pseudo-label of input image attribute, and 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.

3.5 Overall Loss

We first train the albedo generator $\mathscr{G}$ on the 142 high-resolution albedo maps (Sec. 3.1) and then train the proposed 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 (J. Deng, J. Guo, Xue, et al. 2019), project it into the latent space of $\mathscr{G}$ by the mapping network $\mathscr{M}$, and then generate high-quality albedo. Meanwhile, we define an illumination network $\mathscr{F}_{illumination}$ and employ an off-the-shelf shape network (Y. Deng et al. 2019) to predict facial illumination, shape, and camera pose (Details are given in Sec. 4.1). Combining the above predictions, we can warp the albedo to image space and render the face $\mathbf{I_R}$. Apart from the semantic attribute loss (Eq. 10), we also employ the following photometric loss, identity loss, and perceptual loss.
The photometric loss is calculated as

$$L_{photo} = \mathbf{M_{mask}} \cdot \|\mathbf{I} - \mathbf{I_R}\|_1, \tag{11}$$

where $\mathbf{M_{mask}}$ is the face skin mask calculated by the off-the-shelf face parsing model (Y. Lin et al. 2021).
The identity loss is the cosine identity distance between the input image and the rendered face:

$$L_{id} = 1 - \frac{\Phi(\mathbf{I}), \Phi(\mathbf{I_R})}{\|\Phi(\mathbf{I})\|_2 \cdot \|\Phi(\mathbf{I_R})\|_2}, \tag{12}$$

where Φ is the pre-trained ArcFace model.
The perceptual loss (Zhang et al. 2018) is defined as follows:

$$L_{per} = \sum_{l} \|\omega_l \odot (\mathscr{F}_l(\mathbf{I}) - \mathscr{F}_l(\mathbf{I_R}))\|_2^2, \tag{13}$$

where l denotes the different level of a pre-trained VGG model $\mathscr{F}$, and ω_l is the scaling factor.
The overall objective function is then defined by combining the above losses:

$$\mathscr{L}_{total} = \lambda_1 \mathscr{L}_{photo} + \lambda_2 \mathscr{L}_{id} + \lambda_3 \mathscr{L}_{per} + \lambda_4 \mathscr{L}_{sem}, \tag{14}$$

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 fairness and quality. We first give the implementation details (Sec. 4.1) and introduce datasets (Sec. 4.2). We conduct ablation studies on the albedo generator, albedo encoder, and visual-textual cues to validate the efficacy of our method (Sec. 4.6). Then, we participate in the FAIR benchmark and compare our ID2Albedo with the state-of-the-art albedo reconstruction methods (Sec 4.4). Finally, our method is tested on the in-the-wild data to ensure fairness under harsh lighting, various poses, and dark skin tones (Sec. 4.5). Please note that due to the released time of TRUST (H. Feng et al. 2022) pre-trained model, we used the BFM version model in the conference version. In the journal version, we utilize the TRUST BalanceAlb version, which results in better rendering and more accurate visualization results.

4.1 Implementation Details

All our implementations are based on PyTorch (Paszke et al. 2019) and NVIDIA V100 cards. For the albedo generator, we employ adaptive discriminator augmentation as in (Karras, Laine, Aittala, et al. 2020). We use Adam (Kingma and Ba 2015) 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 pipeline, we use the differentiable rasterizer from Pytorch3D (Ravi et al. 2020) for rendering. We freeze the parameters of the geometry estimation network (Y. Deng et al. 2019) and ArcFace (J. Deng, J. Guo, Ververas, et al. 2020; J. Deng, J. Guo, Xue, et al. 2019), and train the illumination network $\mathcal{F}_{illumination}$ and the mapping network $\mathscr{M}$. The pre-trained shape network (Y. Deng et al. 2019) is based on the BFM (Paysan et al. 2009) 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 (He et al. 2016) as initialization and predicts 27 illumination coefficients. The mapping network $\mathscr{M}$ 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 2e-5, and a total number of iterations of 50K. All the training process is on the SFHQ dataset (Contributors 2022), a high-quality synthetic dataset without data privacy concerns.

4.2 Datasets

We train an albedo generator by using diverse albedos bought from 3D Scan Store[2]. Then, we learn a mapping network from the facial identity feature to the latent space of our albedo generator by using the SHFQ (Contributors 2022) dataset. After the training process, we test our ID2Albedo model on the FFHQ (Karras, Laine, and Aila 2019), CFP (Sengupta et al. 2016), and FAIR (H. Feng et al. 2022) datasets for fair comparison.

3D Scan Store Albedos. We buy 142 albedo maps from the 3D scan store, including 50 males and 92 females. These albedos have a balanced distribution regarding races. We transfer raw albedo maps to our modified BFM topology, as shown in Fig. 4. Our albedo generator is then trained by these aligned albedo maps.

SFHQ dataset. This dataset consists of 3 parts, each containing 90,000 curated high-quality 1024×1024 synthetic face images. All synthetic images are created by initiating with the inspiration image. We only choose part 2 for training. In part 2, part of the inspiration images are generated from the Stable Diffusion (Rombach et al. 2022) model using various face portrait prompts that span various ethnicities, ages, expressions, hairstyles, etc., and other images are taken from the Face Synthetics Dataset (Wood et al. 2021). All inspiration images are encoded into StyleGAN2 (Karras, Laine, Aittala, et al. 2020) latent space, performing a minor manipulation that turns each image into a photo-realistic image. These candidate images were then further selected using a semi-automatic process.

[2]https://www.3dscanstore.com/

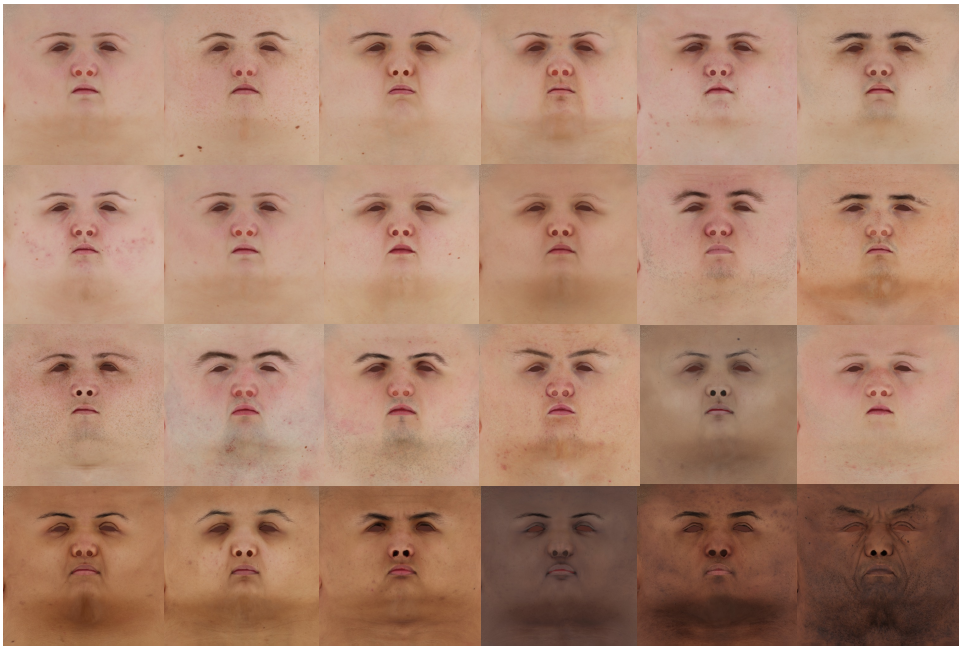

Fig. 4. Visualization of our training data for the albedo generator. Our training data has diversity in race, skin color, and age, while also maintaining high quality.

FFHQ dataset. The FFHQ (Karras, Laine, and Aila 2019) dataset contains 70,000 high-quality facial images with a resolution of 1024 × 1024. It also displays diverse ages, ethnicities, and image backgrounds.
CFP dataset. The CFP (Sengupta et al. 2016) dataset includes front and profile photos of 500 celebrities. For each person, it has ten front images and four profile images. These images are in-the-wild, collected from the Internet.
FAIR dataset. The FAIR Benchmark (H. Feng et al. 2022) is constructed using 206 high-quality 3D head scans. The scans were selected to cover a relatively balanced range of sexes and skin colors. Ages range from 18 to 78 years old. All of the scans were captured under neutral expression and uniform lighting.

4.3 Semantic Validation

To evaluate the feasibility of semantic supervision on albedo maps, we conducted a controlled comparison of CLIP's property predictions. We selected 150 albedo maps from the 3D Scan Store dataset, which provides ground-truth demographic labels and directly renderable images. By relighting images under five distinct lighting conditions, we synthesized images resembling real-world scenes. For each sample, we compared two property sources: (1) CLIP predictions based on reflectance maps, and (2) CLIP predictions based on original RGB images and ground truth labels. As shown in Tab. 1, CLIP achieves slightly lower accuracy on reflectance maps compared to original images—as expected, since reflectance maps lack lighting information and specific high-frequency features—yet reliably distinguishes demographic categories. This confirms that CLIP-based attributes retain semantic meaning on reflectance maps, making them suitable for weakly supervised learning.

Albedo maps preserve the global chromatic structures most indicative of demographic attributes (e.g., pigmentation patterns, melanin distribution, undertones), which dominate CLIP's cross-modal inference. In addition, CLIP's high-level embeddings are designed to be invariant to low-level details and exhibit strong generalization

Table 1. Comparison of CLIP-based attribute prediction accuracy on reflectance maps versus RGB images.

Attribute	CLIP on RGB	CLIP on Albedo
Race	87.3%	81.6%
Skin Tone	84.1%	78.9%
Age Group	82.5%	76.8%
Gender	94.2%	90.4%
Average	**87.0%**	**81.9%**

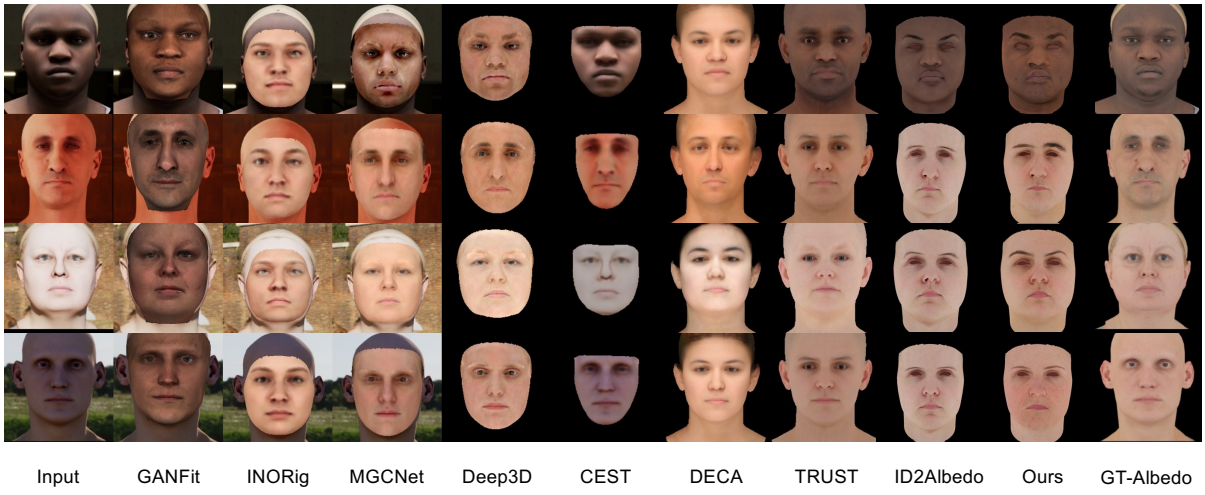

Fig. 5. Comparison on the FAIR benchmark (H. Feng et al. 2022). Please note that we don't have any FAIR benchmark albedo ground truth, so we choose the same input images as TRUST. From left to right: input images, GANFIT (Gecer, Ploumpis, Kotsia, and S. Zafeiriou 2019), INORig (Bai et al. 2021), MGCNet (Shang et al. 2020), Deep3D (Y. Deng et al. 2019), CEST (Wen et al. 2021), DECA (Y. Feng et al. 2021), TRUST (H. Feng et al. 2022), ID2Albedo (Ren, J. Deng, Ma, et al. 2023), ours and ground-truth albedo rendering.

to out-of-distribution appearances. Consequently, even though albedo maps deviate from natural photographs, CLIP can still infer coarse demographic categories—such as race, skin tone, age group, and gender—with reasonable reliability.

4.4 FAIR Benchmark Results

On the FAIR Benchmark (H. Feng et al. 2022), the Individual Typology Angle (ITA) score is recommended to classify skin tones into 6 categories. The ITA score is calculated as follows:

$$\mathrm{ITA}(L^*, b^*) = \frac{180}{\pi} \times \arctan\left(\frac{L^* - 50}{b^*}\right), \tag{15}$$

where L^* and b^* represent the lightness and yellow/blue components of the CIE L*a*b* color space, respectively. The ITA error is the absolute value of the difference between the predicted ITA value and the ground truth ITA value. 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.

Table 2. Comparison to state-of-the-arts on the FAIR benchmark (H. Feng et al. 2022). We utilize the FAIR official metrics, such as average ITA error, bias score (standard deviation), the total score (avg. ITA error + Bias score), mean average error, and average ITA error per skin type in degrees (I: very light, VI: very dark). Our method achieves accurate skin color predictions, especially on very dark skin.

Method	Avg. ITA error ↓	Bias ↓	Score ↓	MAE ↓	ITA error per skin type ↓					
					I	II	III	IV	V	VI
Deep3D (Y. Deng et al. 2019)	22.57	24.44	47.02	27.98	**8.92**	9.08	8.15	10.90	28.48	69.90
GANFIT (Gecer, Ploumpis, Kotsia, and S. Zafeiriou 2019)	62.29	31.81	94.11	63.31	94.80	87.83	76.25	65.05	38.24	11.59
MGCNet (Shang et al. 2020)	21.41	17.58	38.99	25.17	19.98	12.76	8.53	9.21	22.66	55.34
DECA (Y. Feng et al. 2021)	28.74	29.24	57.98	38.17	9.34	11.66	11.58	16.69	39.10	84.06
INORig (Bai et al. 2021)	27.68	28.18	55.86	33.20	23.25	11.88	**4.86**	9.75	35.78	80.54
CEST (Wen et al. 2021)	35.18	12.14	47.32	29.92	50.98	38.77	29.22	23.62	21.92	46.57
TRUST (H. Feng et al. 2022) (BFM)	16.19	15.33	31.52	21.82	12.44	**6.48**	5.69	9.47	16.67	46.37
TRUST (H. Feng et al. 2022) (AlbedoMM)	17.72	15.28	33.00	19.48	15.50	10.48	8.42	**7.86**	**15.96**	48.11
TRUST (H. Feng et al. 2022) (BalancedAlb)	13.87	**2.79**	**16.67**	**18.41**	11.90	11.87	11.20	13.92	16.15	18.21
ID2Albedo (Ren, J. Deng, Ma, et al. 2023)	**12.07**	4.91	16.98	23.33	18.30	9.13	5.83	9.46	19.09	**10.59**
Ours	12.15	4.90	17.05	23.23	18.05	9.24	6.05	8.75	19.34	11.47

Following TRUST (H. Feng et al. 2022), we perform a qualitative and quantitative evaluation on the FAIR benchmark, shown in Fig. 5 and Tab. 2, respectively. In Fig. 5, a common problem with current methods is a strong bias (Bai et al. 2021; Gecer, Ploumpis, Kotsia, and S. Zafeiriou 2019) towards specific skin types, or albedo models that limit the modeling of appropriate skin tone types (Y. Deng et al. 2019; Shang et al. 2020). Both TRUST (H. Feng et al. 2022) and our method perform albedo estimation very well. Thanks to a powerful albedo generator, our method produces more realistic results. Numerically, our method and the conference version (Ren, J. Deng, Ma, et al. 2023) produce similar results for average ITA error and darker skin types, and both are comparable to TRUST in terms of total scores. This is because our improvement is in the albedo generation part, while it is consistent with the conference version in the albedo inference part. It can be observed that a better albedo generator would have a slight impact on the overall performance. In contrast, the rest of the algorithms are biased toward different skin types.

In each specific skin type, our method and the conference version show the same trend. 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 our albedo training data. The situation is the same for type 5 light black skin. TRUST (H. Feng et al. 2022) achieves a minimal bias score because of semi-supervised learning. Overall, our algorithm makes good progress in ITA error and achieves state-of-the-art performance, as shown in Tab. 2.

4.5 Real-World Results

To evaluate the robustness of our approach in real-world images, we qualitatively compare it with other methods on the FFHQ dataset, as shown in Fig. 6 and 7. The results show that our method achieves more realistic results while maintaining fairness.

Furthermore, we compare with TRUST (H. Feng et al. 2022) in various environments and poses on the same subject. TRUST under extreme poses results in irrational estimates, as illustrated in the third row of Fig. 8. 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. 3 demonstrate that adding progressive Gaussian noise to the multiscale feature space improves the performance of the albedo generator. Our method achieves the best scores in all image-level metrics according to the generator indicator.

Fig. 6. Comparisons on in-the-wild images. Input images are all from FFHQ (Karras, Laine, and Aila 2019). From top to bottom: inputs, ours, ID2Albedo (Ren, J. Deng, Ma, et al. 2023) and TRUST (H. Feng et al. 2022) rendered and albedo images, DECA (Y. Feng et al. 2021) and Deep3D (Y. Deng et al. 2019) albedo images. We achieve the most realistic rendered results.

4.6 Ablation Study

Albedo Generator. We first verify the subspace-based GAN. We perform a comparison with the original StyleGAN and several GANs invariants for limited data (eg. StyleGANv2-ADA (Karras, Aittala, et al. 2020), ProjectedGAN (Sauer et al. 2021), InsGAN (C. Yang, Shen, Xu, and B. Zhou 2021) and DynamidGAN (C. Yang, Shen, Xu, D. Zhao, et al. 2022)) on our aligned UV data.

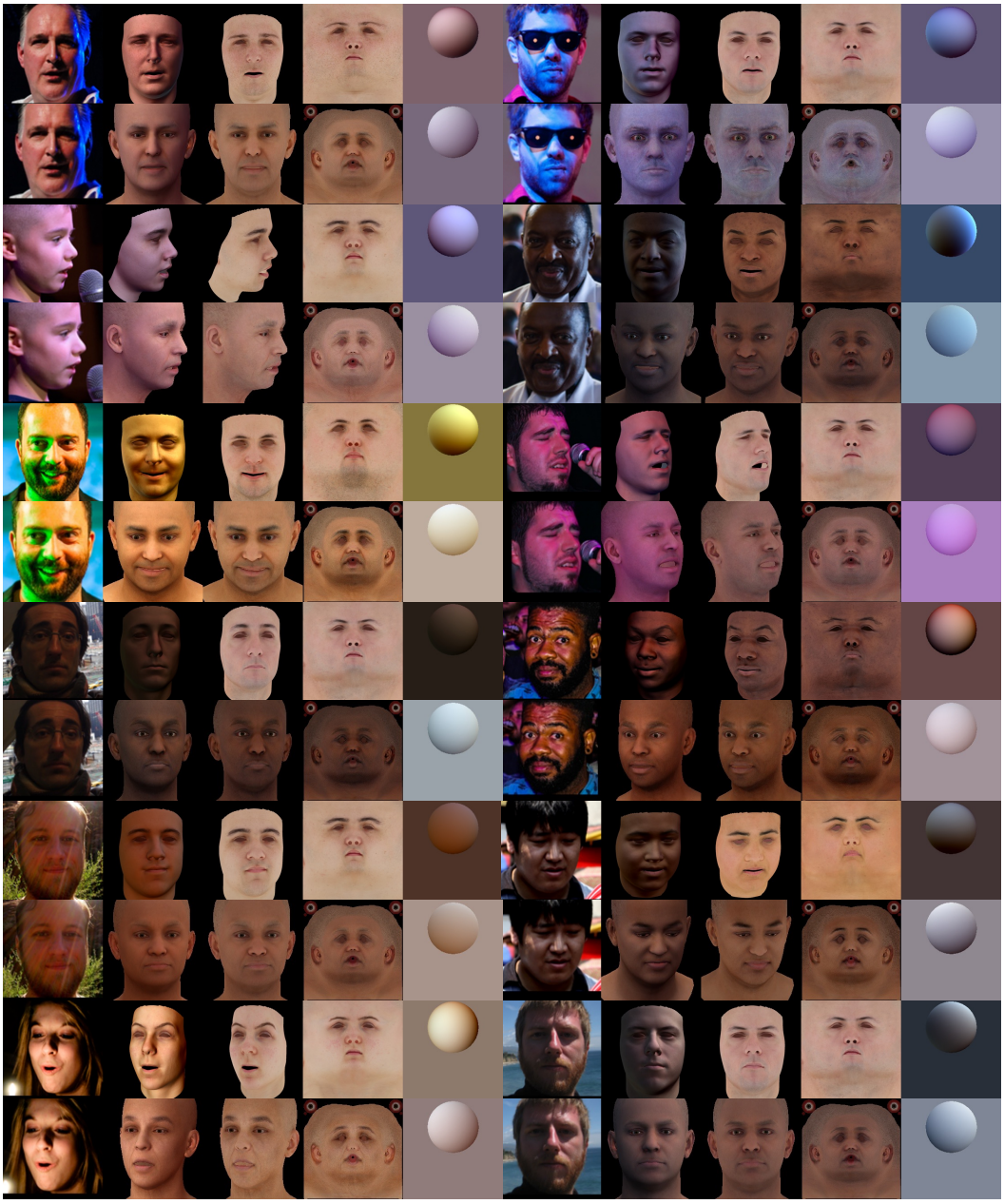

Fig. 7. Comparisons on in-the-wild images. Input images are all from FFHQ (Karras, Laine, and Aila 2019). From left to right: Inputs, rendered images, albedo images, albedo maps, and lighting probes. The odd rows are our results, while the even rows are TRUST (H. Feng et al. 2022)'s results.

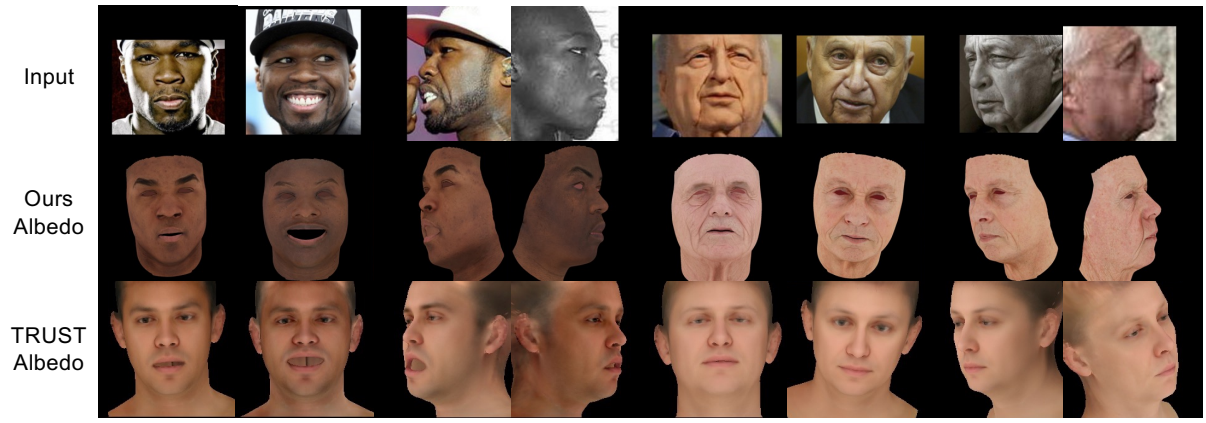

Fig. 8. Comparisons on the real-world images from the same subject. Input images are all from CFP (Sengupta et al. 2016) dataset.

Table 3. Comparisons of our method with other albedo reconstruction methods on FFHQ, eg., Deep3D (Y. Deng et al. 2019), DECA (Y. Feng et al. 2021), and TRUST (H. Feng et al. 2022). Given the absence of GT albedo, we compare the rendered image to the original image.

Methods	M-SSIM↑	LPIPS↓	FID↓	ID↑
Deep3D (Y. Deng et al. 2019)	0.73	0.1933	74.41	0.712
DECA (Y. Feng et al. 2021)	0.61	0.2089	98.13	0.585
TRUST (H. Feng et al. 2022)	0.84	0.1632	62.22	0.785
ID2Albedo	0.87	0.1549	45.56	0.867
Ours	**0.87**	**0.1513**	**42.12**	**0.878**

StyleGAN2-ADA (Karras, Aittala, et al. 2020) proposes an adaptive discriminator augmentation mechanism that significantly stabilizes training in limited data regimes. The subsequent ProjectedGAN (Sauer et al. 2021) considers random projections in the feature space to reduce discriminator overfitting. Additionally, InsGAN (C. Yang, Shen, Xu, and B. Zhou 2021) uses contrast learning to construct instance relations, allowing the discriminator to achieve instance discrimination, which improves the discriminator generation performance. DynamicD (C. Yang, Shen, Xu, D. Zhao, et al. 2022) introduces a dynamic discriminator, which automatically adjusts its size according to the training sample size to maintain a good discriminator capability. However, the above approach still produces a large number of artifacts on limited training data, such as flicker noise and unrobust structures. The visual comparison example is illustrated in Fig. 9. The FID results, as shown in Tab. 4, indicate that our subspace-based GAN can achieve better generation results. Additionally, we visualize latent interpolation results in Fig. 10. Finally, we compared it to ID2Albedo (Ren, J. Deng, Ma, et al. 2023), as shown in Fig. 11. While ID2Albedo (Ren, J. Deng, Ma, et al. 2023) achieves good results, speckles or artifacts may still appear at the edges of the albedo map. By contrast, the improved version can achieve consistently high quality at the edges.

Besides, our experiments show that using the albedo generator alone did not yield satisfactory results, as demonstrated in Tab. 5. We believe this is because the albedo generator, unlike the PCA-based model, is not as compact, resulting in greater variance when used without proper guidance.

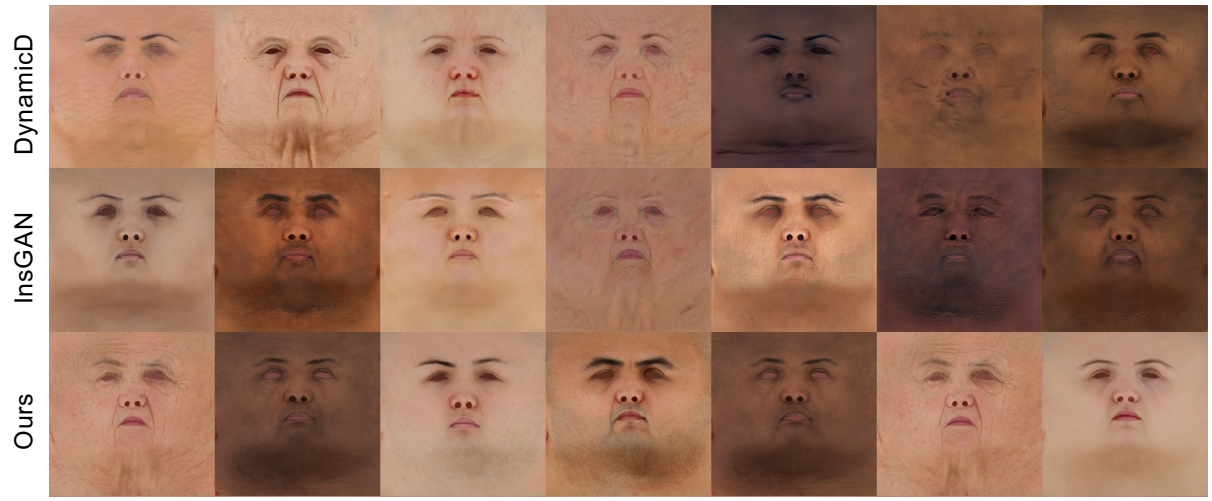

Fig. 9. Visualization of random seeds generation. The figure shows that both DynamicD (C. Yang, Shen, Xu, D. Zhao, et al. 2022) and InsGAN (C. Yang, Shen, Xu, and B. Zhou 2021) produce a considerable amount of noise and artifacts in their results. In contrast, our generator outperforms in terms of image quality, producing clearer and more realistic details in the generated images.

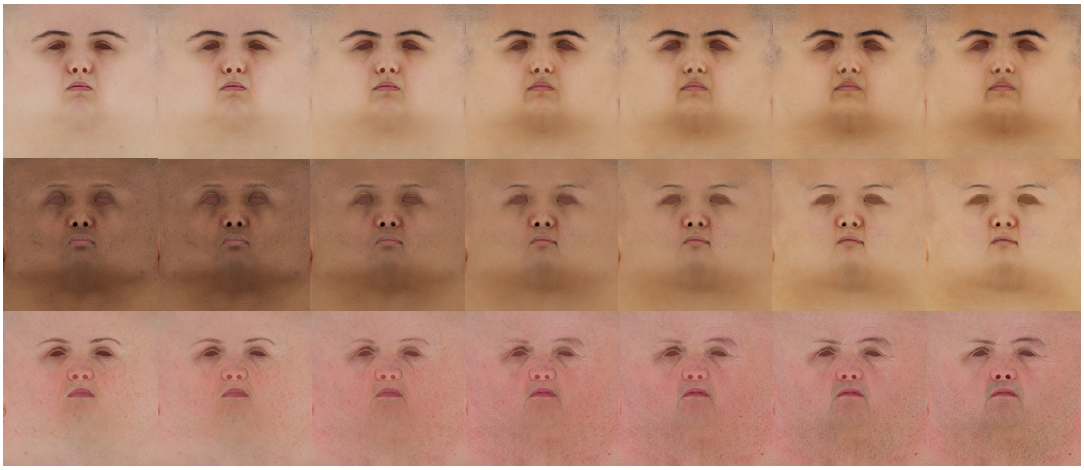

Fig. 10. Visualization of latent codes interpolation. Our generator has been proven to have excellent performance in generating high-quality albedos.

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. 6. We observe that finetuning of partial layers or the entire network results in overfitting of the training data with significantly worse results. In contrast, we start with identity features, which can effectively perform the albedo reconstruction task.

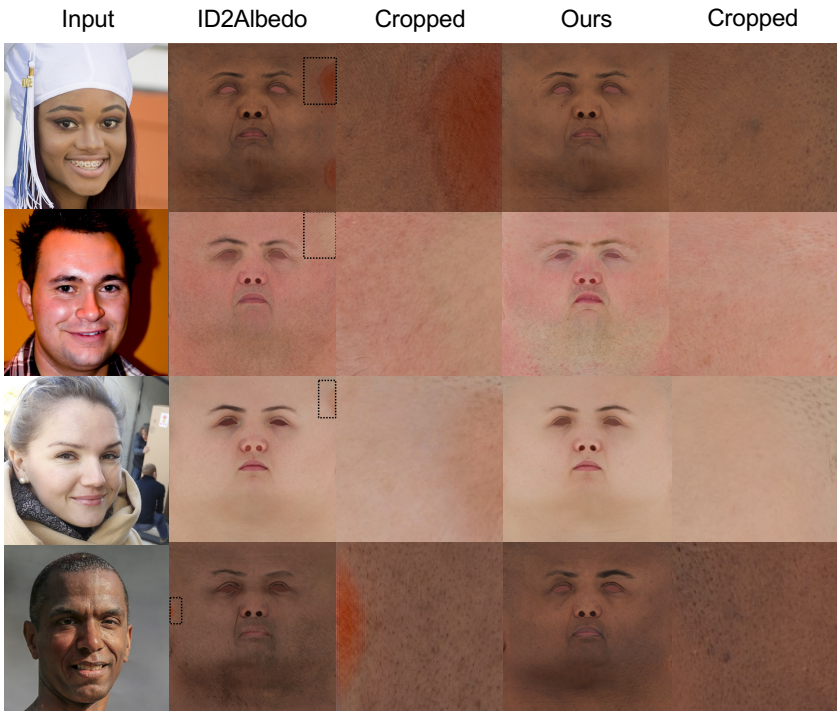

Fig. 11. Visualization comparisons between ours and ID2Albedo (Ren, J. Deng, Ma, et al. 2023). We argue that although ID2Albedo (Ren, J. Deng, Ma, et al. 2023) is able to generate high-quality albedos, it is less stable in the edge region and will occasionally generate artifacts. Our proposed feature augmentation approach can largely alleviate these artifacts.

Table 4. **Ablation study of albedo generator.** Our comparison includes StyleGAN2-ADA (Karras, Aittala, et al. 2020), ProGAN (Sauer et al. 2021), InsGAN (C. Yang, Shen, Xu, and B. Zhou 2021), and DynamidD (C. Yang, Shen, Xu, D. Zhao, et al. 2022).

Methods	**FID ↓**	InsGAN (C. Yang, Shen, Xu, and B. Zhou 2021)	47.6
StyleGAN2 (Karras, Laine, and Aila 2019)	64.7	DynamidD (C. Yang, Shen, Xu, D. Zhao, et al. 2022)	44.5
StyleGAN2-ADA (Karras, Aittala, et al. 2020)	49.3	ID2Albedo (Ren, J. Deng, Ma, et al. 2023)	42.2
ProGAN (Sauer et al. 2021)	52.8	Ours	**40.4**

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 error and bias scores. A broader range of attributes produces better results, as shown in Tab. 7.

Furthermore, we use a manually labeled ethnographic dataset, RFW (M. Wang et al. 2019), to train a supervised race prediction model. The results show that using predicted ethnographic labels underperforms the proposed method because the CLIP model can provide more diverse attribute constraints.

Table 5. Ablation study of different albedo model. We compare our method to the following alternatives: 1) AlbedoMM (W. A. P. Smith et al. 2020) without visual-textual cues, 2) AlbedoMM with cues, and 3) AlbedoGen without cues.

Configs	Avg. ITA error ↓	Bias ↓	Score ↓
AlbedoMM (W. A. P. Smith et al. 2020)	23.25	21.67	44.92
AlbedoMM (W. A. P. Smith et al. 2020)+ cues	18.25	8.53	26.78
AlbedoGen	25.66	23.51	49.17
AlbedoGen+cues (Ours)	**13.46**	**5.86**	**19.32**

Table 6. **Ablation study of albedo encoder.** We show a comparison with various alternative encoder configurations, where *L2*, *L3*, *L4* represents different network stages.

Albedo Encoder	Avg. ITA error ↓	Bias ↓	Score ↓
ResNet-100 (He et al. 2016) (Scratch)	58.46	32.59	91.05
ResNet-100 (He et al. 2016) (ImageNet)	31.63	15.48	47.11
ArcFace (J. Deng, J. Guo, Xue, et al. 2019) (fully trainable)	41.63	19.81	61.44
ArcFace (J. Deng, J. Guo, Xue, et al. 2019) (L2 + L3 + L4)	28.75	11.87	40.62
ArcFace (J. Deng, J. Guo, Xue, et al. 2019) (L3 + L4)	19.52	9.46	28.98
ArcFace (J. Deng, J. Guo, Xue, et al. 2019) (L4)	14.58	6.79	21.37
ArcFace (J. Deng, J. Guo, Xue, et al. 2019) (Frozen)(Ours)	**13.46**	**5.86**	**19.32**

Table 7. **Ablation study of CLIP-based cues.** We compare our method to the following alternatives: 1) no visual-textual cue, 2) training from a labeled race dataset, and 3) using CLIP racial cues.

Configs	Avg. ITA error ↓	Bias ↓	Score ↓
w/o any cues	25.66	23.51	49.17
Predicted race label	18.13	10.46	28.59
Only CLIP race cues	16.21	7.44	23.65
Only CLIP wrinkles cues	21.97	8.92	30.89
Only CLIP age cues	22.46	8.50	30.96
Only CLIP gender cues	22.68	9.10	31.78
Only CLIP skin tone cues	23.43	13.54	36.97
CLIP all cues (ours)	**13.46**	**5.86**	**19.32**

5 Appliciation

In this section, we show two typical ID2Albedo applications to showcase how they can be utilized in various scenarios. We transfer our ID2Albedo method to ID2Texture directly, which in turn allows fast texture completion and boosts pose-invariant face recognition performance. We also extend ID2Albedo from the face to the whole head/body, achieving a realistic head/human reconstruction.

Table 8. Evaluations of occlusion-robust UV completion on the Multi-PIE (Gross et al. 2010) dataset. Note that we manually add occlusions for evaluation.

Methods	w. Occ ID ↑	w.o. Occ ID ↑
UVGAN (J. Deng, S. Cheng, et al. 2018)	0.791	0.838
OSTeC (Gecer, J. Deng, et al. 2021)	0.824	0.879
PBIDR (Ren, Lattas, et al. 2023)	0.873	0.881
ID2Texture(Ours)	**0.882**	**0.886**

5.1 Identity-guided Texture Completion

Considering that recent works (J. Deng, S. Cheng, et al. 2018; Gecer, J. Deng, et al. 2021) have used the visible part as supervision to synthesize the invisible part, we argue that this visible-part-guided synthesis setting is not robust, especially under inputs with external occlusions, extreme lighting conditions, and low-quality inputs. We focus on robust high-resolution texture completion from "in-the-wild" face images using the existing ID2"X" framework.

Specifically, we use a state-of-the-art face recognition network Φ (ie. ArcFace (J. Deng, J. Guo, Xue, et al. 2019)) to predict robust identity features and train a lightweight mapping network $\mathscr{M}$ to fine-tune the identity features, which are finally interpreted by our high-quality texture generator $\mathscr{G}$. The identity embedding network Φ is also a ResNet-100 model trained on the large-scale WebFace dataset (Z. Zhu, G. Huang, J. Deng, Ye, J. Huang, X. Chen, J. Zhu, T. Yang, Lu, et al. 2021) under the ArcFace loss (J. Deng, J. Guo, Xue, et al. 2019). The pre-trained ArcFace model is able to extract face identity features that are robust to illumination, rotation, and occlusion. Therefore, the proposed robust texture completion framework (ID2Texture) can easily handle these face appearance variations in the wild. After training, the mapping network $\mathscr{M}$ modifies the feature distribution of the original identity features to match the latent space of our texture generator.

We fit about 12000 high-quality UV textures by using OSTeC (Gecer, J. Deng, et al. 2021) on CelebA Mask HQ dataset, train the texture generator $\mathscr{G}$ on these data, and then train the whole pipeline on our paired face-texture dataset. We employ the above photometric loss, identity loss, and perceptual loss between generated texture $\mathbf{T}$ and pseudo ground truth label $\mathbf{T}_0$. After training, given a facial image $\mathbf{I}$, we compute the identity feature by the ArcFace model (J. Deng, J. Guo, Xue, et al. 2019), project it into the latent space of $\mathscr{G}$ by the mapping network $\mathscr{M}$, and then generate high-quality texture $\mathbf{T}$. This ID2Texture approach can be used for UV texture completion and pose-invariant face recognition.

UV Texture Completion. We compare our ID2Texture model quantitatively with previous UVGAN (J. Deng, S. Cheng, et al. 2018), OSTeC (Gecer, J. Deng, et al. 2021), and PBIDR (Ren, Lattas, et al. 2023) on the UVDB (J. Deng, S. Cheng, et al. 2018) dataset. We randomly select 100 subjects, and half of them have been manually added occlusions as this dataset hardly contains occlusion cases. We employ identity similarity metrics to evaluate the estimated UV maps and the ground truth fairly. Here, we use the off-the-shelf AdaFace (M. Kim et al. 2022) ResNet100 model trained on the MS1MV3 dataset (Y. Guo et al. 2016) to predict the similarity scores. The proposed ID2Texture approach outperforms all baseline methods on both occluded and non-occluded examples, as shown in Tab. 8. We argue that these previous methods predict unknown parts based on the visible parts in the image. In contrast, our ID2Texture directly generates textures from the identity feature, which means it cannot maintain the same color tone as the original image. As seen in Fig. 12, our technique can create high-quality UV texture maps with consistent identity even under occlusion, lighting, and expression fluctuations.

Pose-invariant Face Matching. We evaluate the performance of our work on pose-invariant face recognition in the wild. We choose the widely used dataset CFP (Sengupta et al. 2016), which focuses on extreme pose face

Fig. 12. Comparison of our method with texture completion methods in Multi-PIE (Gross et al. 2010). The occlusions are manually added. Results exhibit that our approach is occlusion-robust and can produce photo-realistic textures.

verification. We follow the OSTeC setting and employ the same ArcFace ResNet-18 network (H. Zhou et al. 2020) trained on MS1MV3 (J. Deng, J. Guo, Xue, et al. 2019; Y. Guo et al. 2016) for face feature embedding. To boost pose-invariant face recognition, frontal-profile comparisons can be changed to frontal-frontal comparisons after frontalization of the profile face or changed to profile-profile comparisons after profiling the frontal face. In UVGAN (J. Deng, S. Cheng, et al. 2018), a set2set comparison strategy is proposed by changing both frontal and profile faces into multi-view image sets. As shown in Tab. 9, the proposed ID2Texute method outperforms the OSTeC under all three test settings.

5.2 Realistic Head/Human Creation

Fig. 13 demonstrates that our realistic albedo can be used for realistic head and human creation. In order to create a photo-realistic full head, we first need to obtain a realistic albedo. Once we have obtained this albedo, we can transfer it to the full head via rigid alignment and non-rigid ICP (Iterative Closest Point) to obtain a textured head that more accurately reflects the individual's appearance. To conduct a low-cost digital human creation, we take our generated albedo and warp it to the full body, allowing us to create a complete digital human that is an accurate representation of the individual. These digital humans can be driven by predefined skeletons, making them ideal for a wide range of virtual reality applications.

Table 9. Verification accuracy(%) comparison on the CFP dataset (Sengupta et al. 2016). The embedding network used here is the ArcFace ResNet-18 Network as in OSTeC (Gecer, J. Deng, et al. 2021).

Methods	Frontal-Frontal ↑	Frontal-Profile ↑
Human	96.24 ± 0.67	94.57 ± 1.10
DR-GAN (Tran et al. 2017)	97.84 ± 0.79	93.41 ± 1.17
DR-GAN+ (Tran et al. 2019)	98.36 ± 0.75	93.89 ± 1.39
PIM (J. Zhao et al. 2018)	99.44±0.36	93.10 ± 1.01
HF-PIM (Cao et al. 2020)	-	94.71 ± 0.83
UVGAN (J. Deng, S. Cheng, et al. 2018)	98.83 ± 0.27	93.09 ± 1.72
+Profile2Frontal	-	93.55 ± 1.67
+Frontal2Profile	-	93.72 ± 1.59
+Set2set	-	94.05 ± 1.73
OSTeC (Gecer, J. Deng, et al. 2021)	99.68 ±0.29	96.14 ± 1.06
+Profile2Frontal	-	97.06 ± 0.74
+Frontal2Profile	-	97.43 ± 0.61
+Set2set	-	97.85 ± 0.57
ID2Texture	99.68 ±0.29	96.14 ± 1.06
+Profile2Frontal	-	97.45 ± 0.62
+Frontal2Profile	-	97.72 ± 0.46
+Set2set	-	98.16 ± 0.36

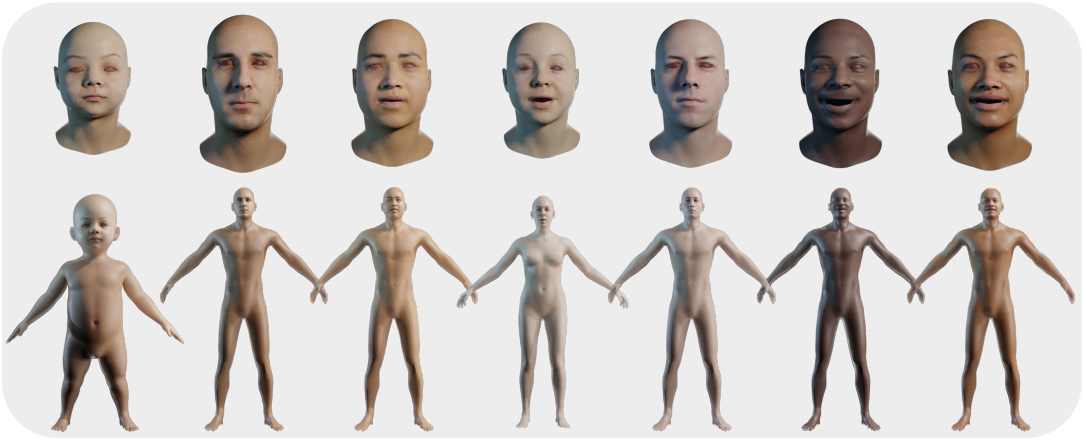

Fig. 13. Realistic head and body creation by the reconstructed albedos.

6 Limitations and Future Work

One concern is that our work focused on skin tones, while facial details may not be correctly recovered, such as eyebrows and freckles. Although we use semantic labels as constraints, there is still a gap between our results and perfect reproduction. In addition, eyes are outside the scope of our approach. Our method currently operates

on individual frames without explicit temporal modeling. Extending ID2Albedo++ with temporal consistency constraints would benefit video-based applications such as video editing or reenactment, which we leave for future work.

7 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 and promoting the metaverse.

Acknowledgments

This work was supported in part by NSFC (62322113, 62376156, 62571314), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and the Fundamental Research Funds for the Central Universities. We also acknowledge Baris Gecer and Alexandros Lattas for the insightful discussion on identity-guided texture completion.

References

O. Aldrian and W. A. Smith. 2012. "Inverse rendering of faces with a 3D morphable model." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35, 5, 1080–1093.

Z. Bai, Z. Cui, X. Liu, and P. Tan. 2021. "Riggable 3D Face Reconstruction via In-Network Optimization." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

V. Blanz, S. Romdhani, and T. Vetter. 2002. "Face identification across different poses and illuminations with a 3D morphable model." In: *FG*.

V. Blanz and T. Vetter. 1999. "A morphable model for the synthesis of 3D faces." In: *SIGGRAPH*.

A. Bora, E. Price, and A. G. Dimakis. 2018. "AmbientGAN: Generative models from lossy measurements." In: *International conference on learning representations*.

J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun. 2020. "Towards high fidelity face frontalization in the wild." *International Journal of Computer Vision*, 128, 1485–1504.

A. Chen, Z. Chen, G. Zhang, K. Mitchell, and J. Yu. 2019. "Photo-realistic facial details synthesis from single image." In: *International Conference on Computer Vision (ICCV)*.

Y. Chen, R. Chen, J. Lei, Y. Zhang, and K. Jia. 2022. "TANGO: Text-driven Photorealistic and Robust 3D Stylization via Lighting Decomposition." In: *Neural Information Processing Systems (NeurIPS)*.

O. Contributors. 2022. *The Synthetic Faces High Quality (SFHQ) Dataset*. https://github.com/SelfishGene/SFHQ-dataset. (2022).

J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou. 2018. "Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. 2020. "Retinaface: Single-shot multi-level face localisation in the wild." In: *CVPR*.

J. Deng, J. Guo, N. Xue, and S. Zafeiriou. 2019. "Arcface: Additive angular margin loss for deep face recognition." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou. 2021. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, 10, 5962–5979.

Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong. 2019. "Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*.

B. Egger, S. Schönborn, A. Schneider, A. Kortylewski, A. Morel-Forster, C. Blumer, and T. Vetter. 2018. "Occlusion-Aware 3D Morphable Models and an Illumination Prior for Face Image Analysis." *IJCV*, 126, 12, 1269–1287.

B. Egger, W. A. P. Smith, et al.. 2020. "3D Morphable Face Models - Past, Present, and Future." *ACM Transactions on Graphics (TOG)*, 39, 5, 157:1–157:38.

P. Esser, R. Rombach, and B. Ommer. 2021. "Taming transformers for high-resolution image synthesis." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

H. Feng, T. Bolkart, J. Tesch, M. J. Black, and V. Abrevaya. 2022. "Towards Racially Unbiased Skin Tone Estimation via Scene Disambiguation." In: *The European Conference on Computer Vision (ECCV)*.

Y. Feng, H. Feng, M. J. Black, and T. Bolkart. 2021. "Learning an animatable detailed 3D face model from in-the-wild images." *ACM Transactions on Graphics (ToG)*, 40, 4, 1–13.

T. B. Fitzpatrick. 1988. "The validity and practicality of sun-reactive skin types I through VI." *Archives of dermatology*, 124, 6, 869–871.

K. Frans, L. Soros, and O. Witkowski. 2022. "Clipdraw: Exploring text-to-drawing synthesis through language-image encoders." *Advances in Neural Information Processing Systems*, 35, 5207–5218.

R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or. 2022. "Stylegan-nada: Clip-guided domain adaptation of image generators." *ACM Transactions on Graphics (TOG)*, 41, 4, 1–13.

B. Gecer, J. Deng, and S. Zafeiriou. 2021. "OSTeC: One-Shot Texture Completion." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou. 2019. "Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

B. Gecer, S. Ploumpis, I. Kotsia, and S. P. Zafeiriou. 2021. "Fast-GANFIT: Generative Adversarial Network for High Fidelity 3D Face Reconstruction." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman. 2018. "Unsupervised training for 3D morphable model regression." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schönborn, and T. Vetter. 2018. "Morphable face models - an open framework." In: *FG*.

R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. 2010. "Multi-pie." *Image and vision computing*, 28, 5, 807–813.

Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. 2016. "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition." In: *The European Conference on Computer Vision (ECCV)*.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. "Deep residual learning for image recognition." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

J. Ho, A. Jain, and P. Abbeel. 2020. "Denoising diffusion probabilistic models." *Advances in neural information processing systems*, 33, 6840–6851.

A. Hokari, N. Terada, R. Sultana, and I. Shimizu. 2025. "Two-shot Spatially Varying Bidirectional Reflectance Distribution Function Estimation for Lustrous Material by Deep Learning." In: *Proceedings of the 2025 7th International Conference on Image, Video and Signal Processing*, 96–104.

G. Hu, P. Mortazavian, J. Kittler, and W. Christmas. 2013. "A facial symmetry prior for improved illumination fitting of 3D morphable model." In: *ICB*. IEEE.

N. Jetchev. 2021. "Clipmatrix: Text-controlled creation of 3d textured meshes." *arXiv preprint arXiv:2109.12922*.

K. Kärkkäinen and J. Joo. 2019. "Fairface: Face attribute dataset for balanced race, gender, and age." *arXiv:1908.04913*.

T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. 2020. "Training Generative Adversarial Networks with Limited Data." In: *Neural Information Processing Systems (NeurIPS)*.

T. Karras, S. Laine, and T. Aila. 2019. "A style-based generator architecture for generative adversarial networks." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. 2020. "Analyzing and Improving the Image Quality of StyleGAN." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

H. Kim, M. Zollhöfer, A. Tewari, J. Thies, C. Richardt, and C. Theobalt. 2018. "InverseFaceNet: Deep monocular inverse face rendering." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

M. Kim, A. K. Jain, and X. Liu. 2022. "Adaface: Quality adaptive margin for face recognition." In: *CVPR*.

D. P. Kingma and J. Ba. 2015. "Adam: A method for stochastic optimization." In: *The International Conference on Learning Representations (ICLR)*.

A. Lattas, S. Moschoglou, B. Gecer, S. Ploumpis, V. Triantafyllou, A. Ghosh, and S. Zafeiriou. 2020. "AvatarMe: Realistically Renderable 3D Facial Reconstruction." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

A. Lattas, S. Moschoglou, S. Ploumpis, B. Gecer, A. Ghosh, and S. P. Zafeiriou. 2021. "AvatarMe^{++}: Facial Shape and BRDF Inference with Photorealistic Rendering-Aware GANs." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. 2017. "Feature pyramid networks for object detection." In: *CVPR*.

Y. Lin, J. Shen, Y. Wang, and M. Pantic. 2021. "RoI Tanh-polar transformer network for face parsing in the wild." *IVC*.

O. Michel, R. Bar-On, R. Liu, S. Benaim, and R. Hanocka. 2022. "Text2mesh: Text-driven neural stylization for meshes." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

A. Paszke et al.. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In: *Neural Information Processing Systems (NeurIPS)*.

O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski. 2021. "Styleclip: Text-driven manipulation of stylegan imagery." In: *International Conference on Computer Vision (ICCV)*.

P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. 2009. "A 3D face model for pose and illumination invariant face recognition." In: *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*. Ieee, 296–301.

A. Radford et al.. 2021. "Learning transferable visual models from natural language supervision." In: *ICML*. PMLR.

R. Ramamoorthi and P. Hanrahan. 2001. "A signal-processing framework for inverse rendering." In: *SIGGRAPH*. Ed. by L. Pocock, 117–128.

N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari. 2020. *PyTorch3D*. https://github.com/facebookresearch/pytorch3d. (2020).

X. Ren, J. Deng, Y. Cheng, J. Guo, C. Ma, Y. Yan, W. Zhu, and X. Yang. June 2024. "Monocular Identity-Conditioned Facial Reflectance Reconstruction." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. (June 2024), 885–895.

X. Ren, J. Deng, Y. Cheng, W. Zhu, Y. Yan, X. Yang, S. Zafeiriou, and C. Ma. June 2025. "S^3-Face: SSS-Compliant Facial Reflectance Estimation via Diffusion Priors." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. (June 2025), 16051–16060.

X. Ren, J. Deng, C. Ma, Y. Yan, and X. Yang. 2023. "Improving Fairness in Facial Albedo Estimation via Visual-Textual Cues." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

X. Ren, A. Lattas, B. Gecer, J. Deng, C. Ma, and X. Yang. 2023. "Facial Geometric Detail Recovery via Implicit Representation." In: *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–8.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. 2022. "High-Resolution Image Synthesis with Latent Diffusion Models." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

A. Sanghi, H. Chu, J. G. Lambourne, Y. Wang, C.-Y. Cheng, M. Fumero, and K. R. Malekshan. 2022. "Clip-forge: Towards zero-shot text-to-shape generation." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

A. Sauer, K. Chitta, J. Müller, and A. Geiger. 2021. "Projected GANs Converge Faster." In: *Neural Information Processing Systems (NeurIPS)*.

S. Schönborn, B. Egger, A. Morel-Forster, and T. Vetter. 2017. "Markov chain monte carlo for automated face image analysis." *IJCV*, 123, 2, 160–183.

S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. 2016. "Frontal to profile face verification in the wild." In: *WACV*.

J. Shang, T. Shen, S. Li, L. Zhou, M. Zhen, T. Fang, and L. Quan. 2020. "Self-Supervised Monocular 3D Face Reconstruction by Occlusion-Aware Multi-view Geometry Consistency." In: *The European Conference on Computer Vision (ECCV)*.

W. A. P. Smith, A. Seck, H. Dee, B. Tiddeman, J. Tenenbaum, and B. Egger. 2020. "A Morphable Face Albedo Model." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

J. Song, C. Meng, and S. Ermon. 2020. "Denoising diffusion implicit models." *arXiv preprint arXiv:2010.02502*.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. 2021. "Score-Based Generative Modeling through Stochastic Differential Equations." In: *International Conference on Learning Representations*. https://openreview.net/forum?id=PxTIG12RRHS.

A. Tewari, M. Zollöfer, H. Kim, P. Garrido, F. Bernard, P. Perez, and T. Christian. 2017. "MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction." In: *International Conference on Computer Vision (ICCV)*.

J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner. 2020. "Neural voice puppetry: Audio-driven facial reenactment." In: *The European Conference on Computer Vision (ECCV)*.

J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. 2016. "Face2face: Real-time face capture and reenactment of rgb videos." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. 2016. "Face2Face: Real-time face capture and reenactment of RGB videos." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

L. Tran, X. Yin, and X. Liu. 2017. "Disentangled Representation Learning GAN for Pose-Invariant Face Recognition." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2017-January, 1283–1292.

L. Tran, X. Yin, and X. Liu. 2019. "Representation Learning by Rotating Your Faces." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41, 3007–3021. doi:10.1109/TPAMI.2018.2868350.

T. Vetter and V. Blanz. 1998. "Estimating coloured 3D face models from single images: An example based approach." In: *The European Conference on Computer Vision (ECCV)*.

M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang. 2019. "Racial faces in the wild: Reducing racial bias by information maximization adaptation network." In: *International Conference on Computer Vision (ICCV)*.

T.-C. Wang, A. Mallya, and M.-Y. Liu. 2021. "One-shot free-view neural talking-head synthesis for video conferencing." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Y. Wen, W. Liu, B. Raj, and R. Singh. 2021. "Self-Supervised 3D Face Reconstruction via Conditional Estimation." In: *International Conference on Computer Vision (ICCV)*.

E. Wood, T. Baltrušaitis, C. Hewitt, S. Dziadzio, T. J. Cashman, and J. Shotton. 2021. "Fake it till you make it: face analysis in the wild using synthetic data alone." In: *ICCV*.

C. Yang, Y. Shen, Y. Xu, D. Zhao, B. Dai, and B. Zhou. 2022. "Improving GANs with A Dynamic Discriminator." In: *Neural Information Processing Systems (NeurIPS)*.

C. Yang, Y. Shen, Y. Xu, and B. Zhou. 2021. "Data-efficient instance generation from instance discrimination." In: *Neural Information Processing Systems (NeurIPS)*.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. 2018. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

J. Zhao et al.. 2018. "Towards pose invariant face recognition in the wild." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

H. Zhou, J. Liu, Z. Liu, Y. Liu, and X. Wang. 2020. "Rotate-and-render: Unsupervised photorealistic face rotation from single-view images." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, D. Du, et al.. 2022. "Webface260M: A benchmark for million-scale deep face recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, et al.. 2021. "Webface260m: A benchmark unveiling the power of million-scale deep face recognition." In: *CVPR*.

M. Zollhöfer, J. Thies, D. Bradley, P. Garrido, T. Beeler, P. Péerez, M. Stamminger, M. Nießner, and C. Theobalt. 2018. "State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications." *EG*.

A Additional Results

In this appendix, we show more additional visualization results.

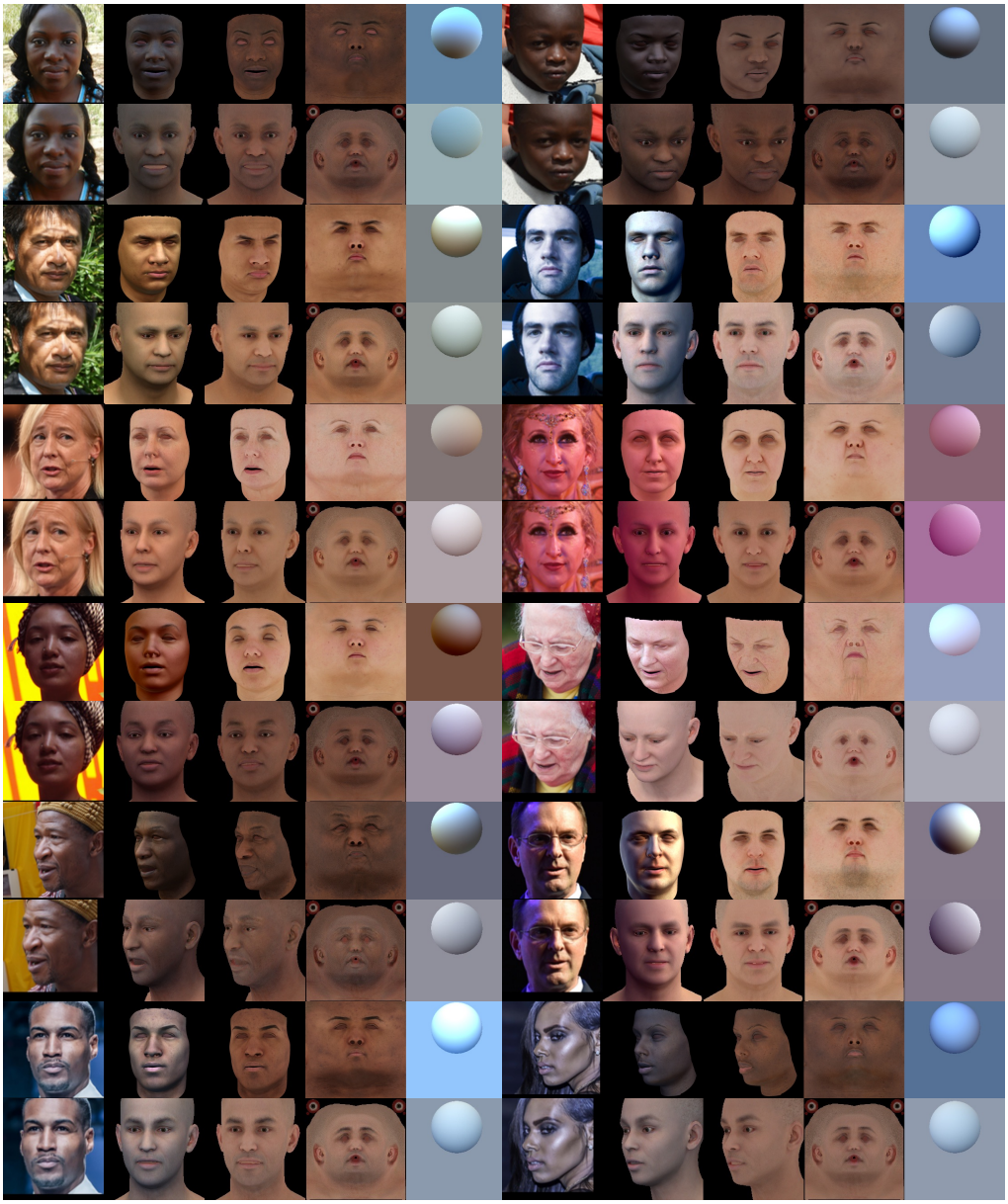

Fig. 14. Comparisons on in-the-wild images. Input images are all from FFHQ (Karras, Laine, and Aila 2019). From left to right: Inputs, rendered images, albedo images, albedo maps, and lighting probes. The odd rows are our results, while the even rows are TRUST (H. Feng et al. 2022)'s results.

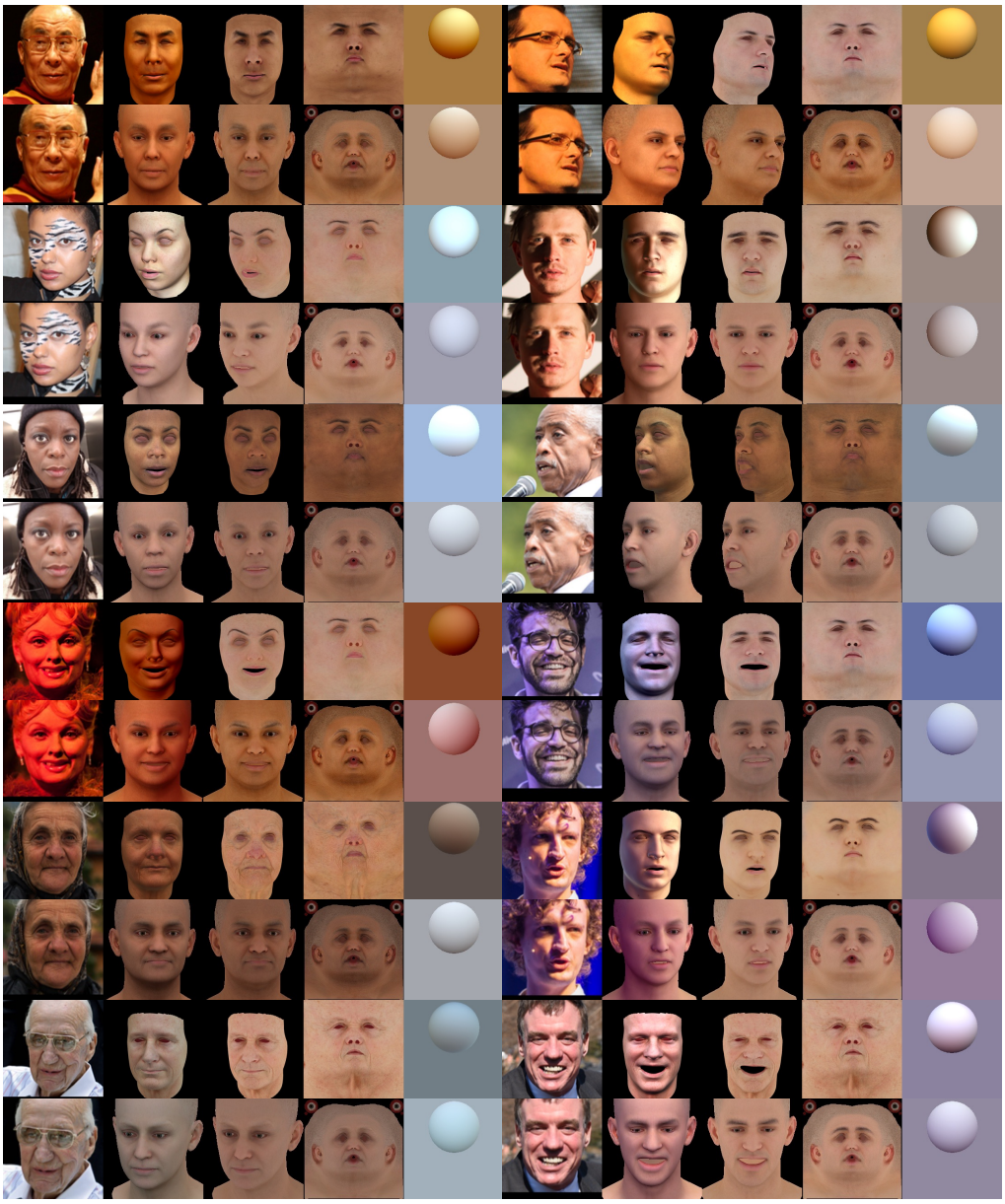

Fig. 15. Comparisons on in-the-wild images. Input images are all from FFHQ (Karras, Laine, and Aila 2019). From left to right: Inputs, rendered images, albedo images, albedo maps, and lighting probes. The odd rows are our results, while the even rows are TRUST (H. Feng et al. 2022)'s results.

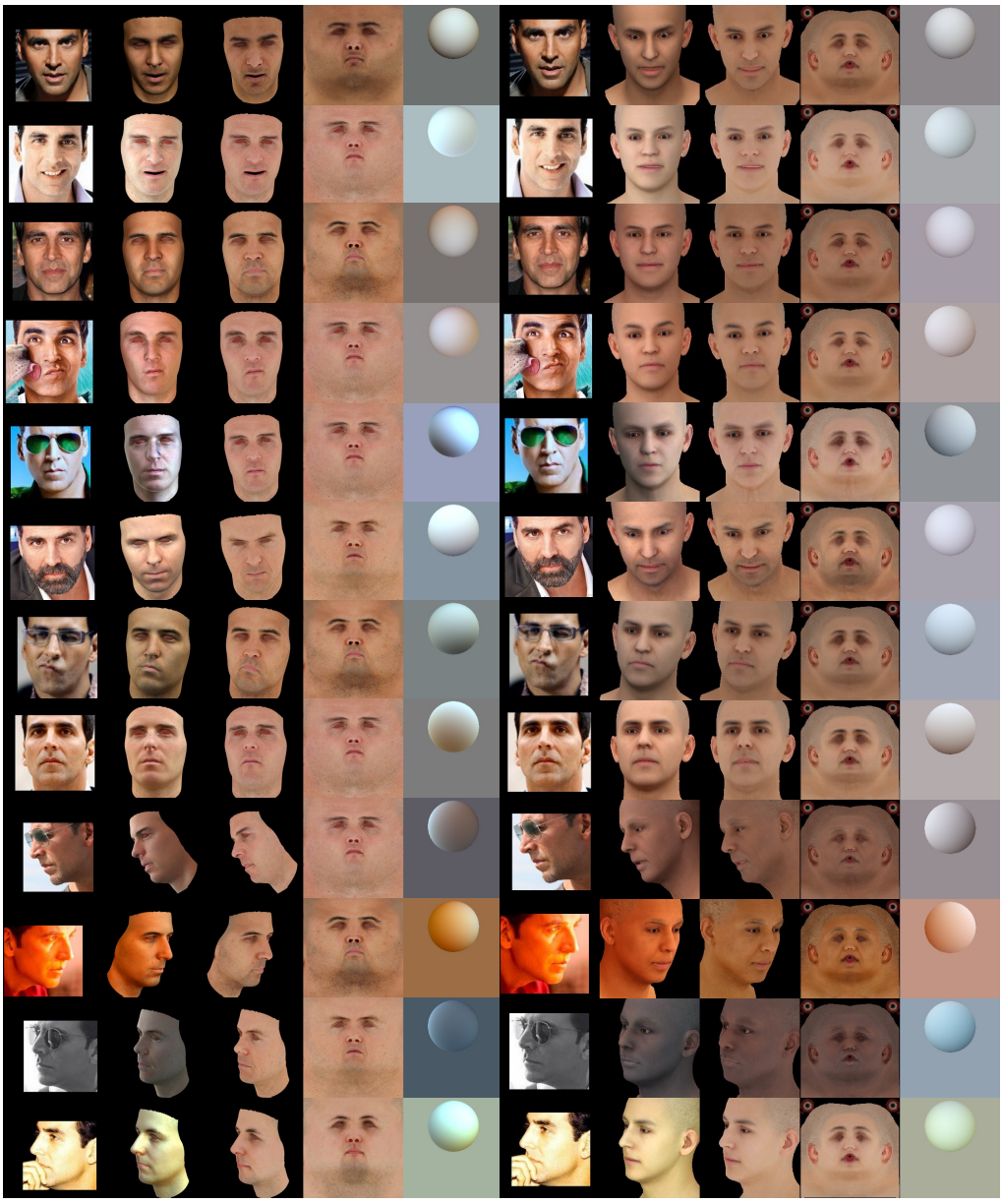

Fig. 16. Comparisons on the real-world same subject images. Input images are all from CFP (Sengupta et al. 2016) dataset. From left to right: Inputs, rendered images, albedo images, albedo maps, and lighting probes. The left column is ours, while the right column is obtained through TRUST (H. Feng et al. 2022).

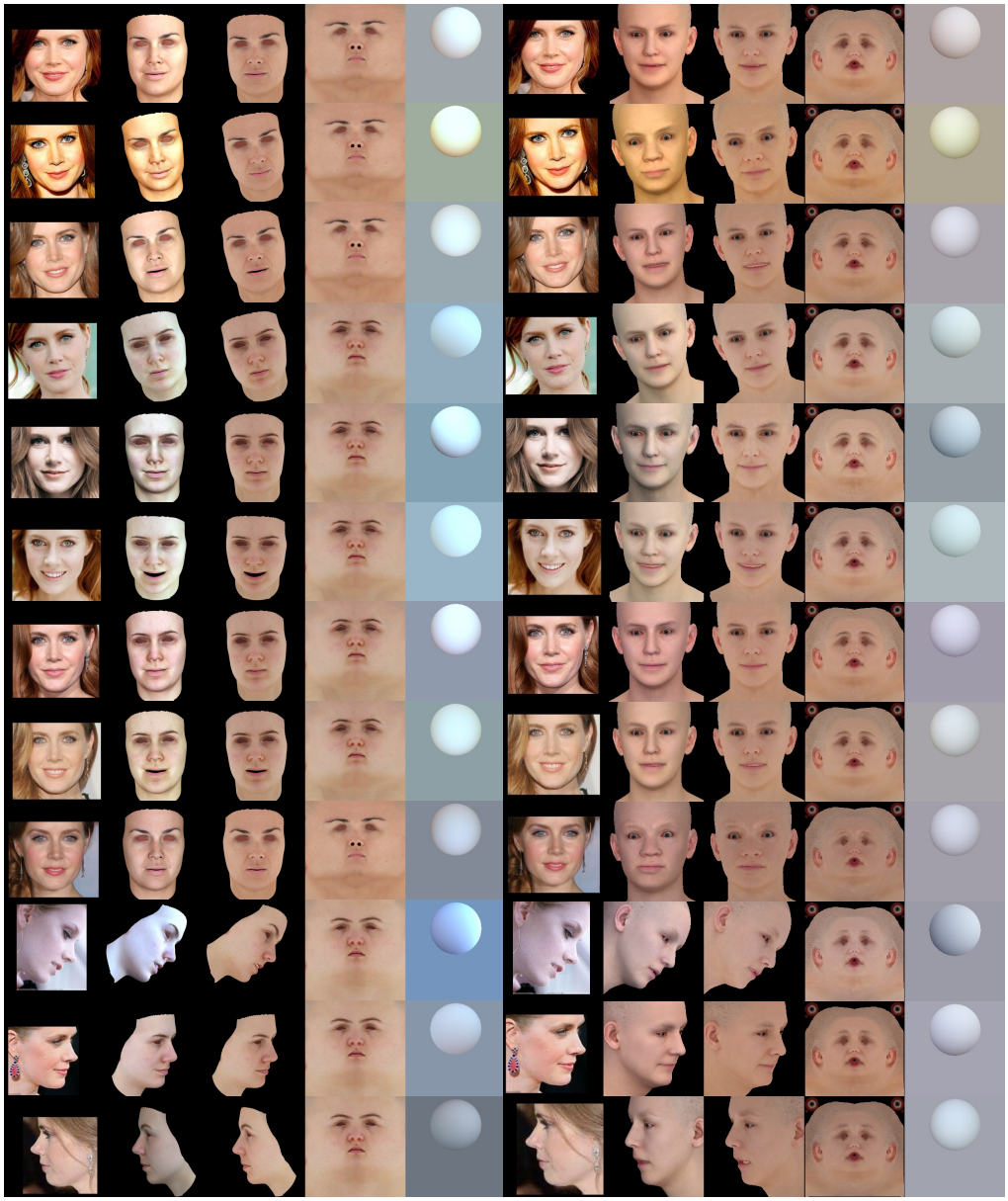

Fig. 17. Comparisons on the real-world same subject images. Input images are all from CFP (Sengupta et al. 2016) dataset. From left to right: Inputs, rendered images, albedo images, albedo maps, and lighting probes. The left column is ours, while the right column is obtained through TRUST (H. Feng et al. 2022).

Received 18 December 2024; accepted 20 April 2026