DOI: 10.1111/cgf.14706
Pacific Graphics 2022
N. Umetani, E. Vouga, and C. Wojtan
(Guest Editors)

Volume 41 (2022), Number 7

BareSkinNet: De-makeup and De-lighting via 3D Face Reconstruction

Xingchao Yang and Takafumi Taketomi

CyberAgent, AI Lab, Japan

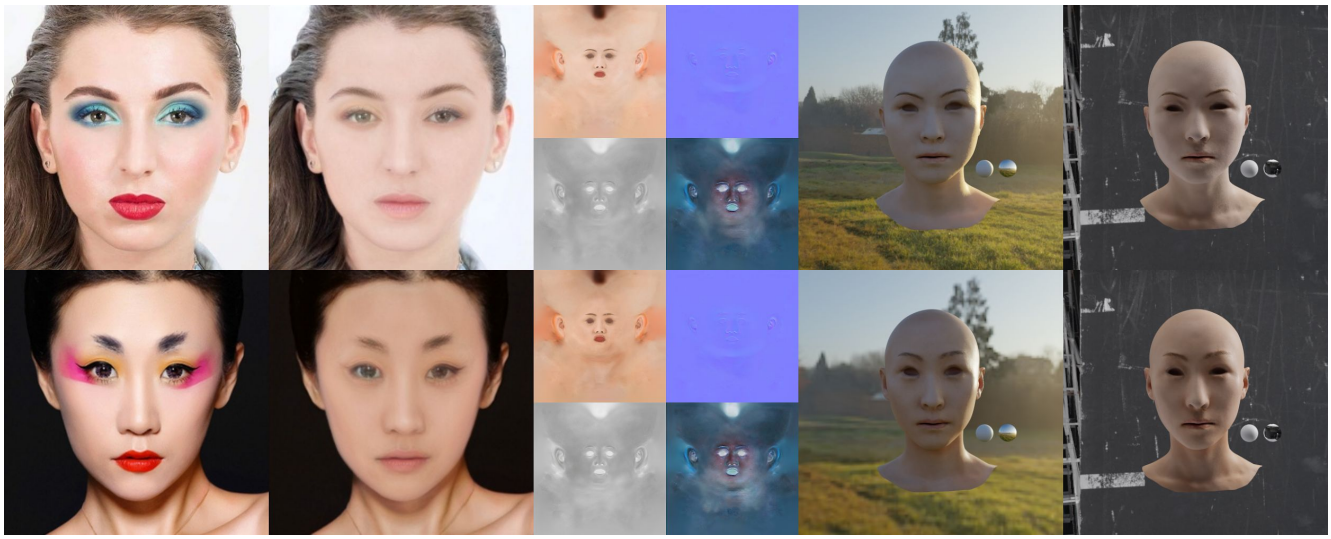

Figure 1: *Our method successfully removes makeup and lighting influences from the input face image to recover a bare skin (de-makeup and de-lighting) face image. The bare skin image facilitates subsequent applications such as the normalized texture (diffuse, normal, roughness, and specular) inference and avatar creation.*

Abstract

We propose BareSkinNet, a novel method that simultaneously removes makeup and lighting influences from the face image. Our method leverages a 3D morphable model and does not require a reference clean face image or a specified light condition. By combining the process of 3D face reconstruction, we can easily obtain 3D geometry and coarse 3D textures. Using this information, we can infer normalized 3D face texture maps (diffuse, normal, roughness, and specular) by an image-translation network. Consequently, reconstructed 3D face textures without undesirable information will significantly benefit subsequent processes, such as re-lighting or re-makeup. In experiments, we show that BareSkinNet outperforms state-of-the-art makeup removal methods. In addition, our method is remarkably helpful in removing makeup to generate consistent high-fidelity texture maps, which makes it extendable to many realistic face generation applications. It can also automatically build graphic assets of face makeup images before and after with corresponding 3D data. This will assist artists in accelerating their work, such as 3D makeup avatar creation.

CCS Concepts

• ***Computing methodologies*** → ***Computer vision; Machine learning; Computer graphics;***

1. Introduction

The realistic 3D avatar generation has become increasingly popular because of its ever-expanding applications in technologies such as virtual and augmented reality, video conferences, and games. With research efforts, a high-quality 3D face model can be acquired through consumer cameras or smartphones [YSN*18, WRV20, BLC*21] and time-consuming specialized hardware [DHT*00, SXZ*20, RGB*20]. 3D face reconstruction from portraits dramatically improves the convenience and speed of the avatar creation.

In general, high-fidelity avatar creation from portraits is achieved using deep-learning-based generative models [GPKZ19, LMG*20, YSN*18]. In these methods, a large-scale high-resolution texture dataset is used to train the networks through supervised learning. However, portrait-based high-fidelity 3D face reconstruction is still difficult when the face image includes complex environment lighting or large expressions. Many face normalization methods have been introduced to overcome these issues [NLW*19, LNK*21]. The input face image is normalized to become as close as possible to be consistent with the high-resolution texture dataset. Although these methods show impressive results on environmental issues, such as pose, lighting, and expression, they do not consider artificial factors such as facial makeup. However, makeup is prevalent in daily life photos. It is necessary to be aware of how to normalize the face image with various kinds of makeup, from light to heavy. Although enlarging the dataset will increase the ability to handle makeup [SRH*11], it is still difficult to address all of them. Additionally, collecting numerous makeup faces is not a realistic task in controlled environments, such as the light stage.

The study aims to generate face images without makeup and lighting influences as shown in Fig. 1. Fig. 2 shows an overview and application of the proposed method. We introduce a de-makeup and de-lighting method that can generate bare skin images by utilizing 3D face reconstruction. Therefore, the corresponding 3D information can be easily accessed. As a result, we can obtain clean 3D face textures without makeup and lighting. The application of high-fidelity texture inference becomes more accurate and accessible.

For the makeup face image input, we propose to build a specialized network called BareSkinNet, which can estimate two types of information. 1) A bare skin image is generated to preserve the high-frequency appearance information of the input face image. The bare skin color is consistent with the high-resolution texture dataset. 2) The low-frequency information is estimated by a 3D face reconstruction process of the 3D morphable model (3DMM).

To remove makeup influences from the input face image, the LADN dataset [GWC*19] is used to train BareSkinNet via weakly supervised learning. The lighting influences are removed by the process of 3D face reconstruction using 3DMM. Our experiments proved that by combining the process of 3D face reconstruction, the effectiveness of makeup removal is also enhanced. We further advanced the capabilities of BareSkinNet in a teacher-student manner. We can employ joint learning of BareSkinNet using a differentiable rendering technique.

In addition, we show an application of high-fidelity texture inference using the results of BareSkinNet. We use the 180 scanned face dataset captured by the high-quality multi-camera scan system to train the high-fidelity texture inference network. Scanned data include facial geometries and the corresponding 4K-resolution diffuse, normal, roughness, and specular texture maps. By combining BareSkinNet and the high-fidelity texture inference network, we can obtain a makeup- and lighting-free 3D face model.

Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to explore how to remove makeup for face normalization. Our method removes both makeup and lighting influences to produce a bare skin image consistent with the texture dataset, which helps the subsequent process of high-fidelity texture inference.
- We propose BareSkinNet, the method jointly training a generator and a process of 3D face reconstruction by leveraging a 3D morphable model. This will be a great convenience because we do not require a reference clean face image or specified light conditions, which solves the problem of non-existent ground truth for in-the-wild de-makeup and de-lighting face images. A teacher-student manner of 3D face reconstruction is introduced to improve performance. We conducted a detailed ablation study to confirm the effectiveness of each component and loss function designs.
- We demonstrate that BareSkinNet can stably produce bare skin images under various makeup and lighting conditions. We show the result of normalized texture inference for 3D avatar creation. The inferred texture maps can be used to create clean avatars for re-lighting and re-makeup processes.

2. Related Work

This study aims to generate a normalized 3D face model from a single image input under various makeup and lighting conditions. Therefore, in this section, we briefly review the existing works on 3D face reconstruction from single image input and face normalization techniques for constructing facial avatars.

2.1. Image-based 3D Face Reconstruction

3DMM has been widely used to reconstruct 3D face models [BV03, THMM17, TZK*17, GZC*18, GCM*18, DYX*19, BLC*21, FFBB21, DBB22]. The 3D face model is reconstructed by fitting the 3DMM to the 2D face image input using facial features, such as face landmarks. Face 3DMM was first proposed by Blanz and Vetter [BV99]. In general, 3DMM is created from a scanned face dataset [PKA*09, CWZ*14, BRZ*16, SSD*20, YZW*20, LBZ*20, LBB*17, WCY*22]. An alternative method is 3DMM built from a face image dataset [Kem13, TL18]. The problem with the 3DMM-based approach is that it is difficult to represent the details of facial appearance using 3DMM. Other methods exist to represent facial shapes in more detail by adding geometric information [JBAT17, GZC*19, KSB11, RSOEK17, SRK17, HCS*18, TZG*18, DMJ*21]. However, it is still difficult to reconstruct a high-fidelity 3D face model using 3DMM fitting. For more detailed discussions on 3DMM, see [EST*20].

Appearance information is an essential factor in achieving high-fidelity 3D face reconstruction. Texture inference methods have

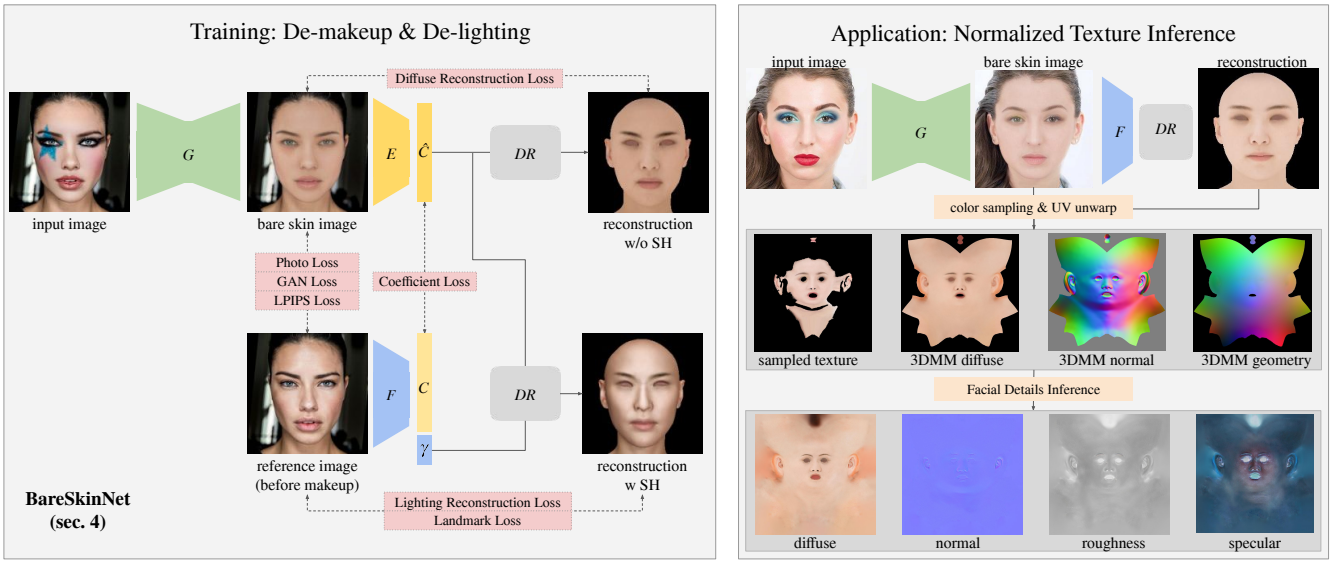

Figure 2: *Overview of the proposed system. BareSkinNet comprises a de-makeup and de-lighting network G, a 3D face reconstruction network E, and a Pre-trained 3D face reconstruction network F. Makeup and lighting influences are removed from the input makeup image. In the inference stage, given an input makeup image, BareSkinNet outputs the bare skin image and the 3DMM coefficients. The diffuse, normal, roughness, and specular maps are inferred from the unwarped UVs of the bare skin image and reconstructed 3DMM.*

been widely proposed for generating high-quality texture maps. UV-GAN [DCX*18] is a UV texture completion method. First, this method generates the face UV map texture by unwarping the input face image using the result of 3D face reconstruction. The generated face UV map has missing regions because of the occlusion. To complete these regions, an image-inpainting technique is applied. This UV map inpainting framework is widely extended [GDZ21, KYT21]. Saito et al. [SWH*17] proposed a photorealistic face texture map generation method. They showed the possibility of a high-resolution face texture map by combining the low-frequency appearance information from 3DMM and the high-frequency appearance information from the input face image. Yamaguchi et al. [YSN*18] employed image completion and image-to-image translation to generate diffuse, specular, and displacement texture maps. They also applied the super-resolution method [LTH*17] to obtain 2K-resolution texture maps. In the training process, they used various lighting conditions to render the synthetic face images to obtain the robustness of the lighting environment change. Using the generated texture maps, they can apply physically based rendering. GANFIT [GPKZ19] uses the advantages of GAN and differentiable rendering to generate a face texture map from a latent vector and optimizes the entire network. They utilized losses based on face landmark detection and face recognition to maintain the identity and fit the face geometry. A similar approach was applied and improved by [LL20]. AvatarMe [LMG*20] is an extension of GANFIT for generating high-fidelity texture maps. AvatarMe trains an image translation network to obtain textures for photorealistic rendering. AvatarMe can generate 4K-resolution diffuse albedo, diffuse normal, specular albedo, and specular normal texture maps. AvatarMe++ [LMP*21] is an improvement of AvatarMe [LMG*20]. In contrast to the other methods, Bao et al. [BLC*21] used an RGB-D selfie video input acquired by a consumer smartphone. They proposed a hybrid method of parametric fitting and CNN-based methods to estimate the reflectance. This method can produce albedo and normal texture maps.

All of the above appearance reconstruction methods directly use information from the input face image. Therefore, the texture generation process fails if the input images are taken under extreme conditions, such as lighting and expressions. To solve this problem, face normalization techniques [CBK*17, NLW*19, LNK*21] have gradually attracted attention, regarded as a preprocess of avatar creation.

2.2. Face Normalization

Many face normalization methods can be used for preprocessing to achieve robust and stable 3D face reconstruction. Face-frontalization methods [HZLH17, YYS*17, YJRF20] can correct face orientation to reduce the number of occluded areas. This process improves the accuracy of the 3D face reconstruction. Lighting and shadow manipulation methods [ZHSJ19, SKCJ18, SBT*19, ZBT*20, HZS*21, RTD*21, PEL*21, WYL*20, RGB*20] that adjust skin color can be further used to generate a stable albedo texture map. Facial attribute editing methods [TEB*20, TER*20, GGU*20, YFD*21, LZG*21] are comprehensive tools that can change pose and lighting. In addition, these methods can restore the input facial expression to a neutral facial expression state.

Nagano et al. [NLW*19] proposed the first face normalization technique for generating an avatar. Perspective distortion, lighting, head pose, and facial expressions in the input image were normalized through a step-by-step process. On the other hand, instead of using the input image directly, Luo et al. [LNK*21] employed

StyleGAN2 [KLA*20] to refine the facial texture. Their method could generate normalized avatar face models under harsh lighting conditions.

Although various works can be used to handle lighting and expression for generating the 3D face model, to the best of our knowledge, no study has investigated the influence of facial makeup on high-fidelity 3D face reconstruction tasks. Recently, GAN-based face image generation can create photorealistic face images. In many cases, the output face image includes facial makeup, especially in female photos. Removing facial makeup is necessary to generate normalized avatar face models. Enlarging the scanned face dataset is not a realistic solution because there are numerous variations in makeup. Makeup transfer is a related research topic that can transfer makeup styles between facial images. BeautyGAN [LQD*18] employed a histogram matching loss to maintain the color distribution between the face regions. LADN [GWC*19] employed a local discriminator to enable a strong makeup style transfer. PSGAN [JLG*19] utilized an attention mechanism to accurately achieve makeup transfer between different head poses and facial expression images. Nguyen et al. [NTH21] proposed a method to transfer makeup patterns precisely by unwarping the face image to a UV representation. SCGAN [DHC*21] employed a style-based encoder to map the makeup style into a disentangled style code, which solves the large spatial misalignment problem. EleGANt [YHXG22] proposed a locally editable makeup transfer method that can achieve more flexible controls.

Unlike the above methods, our method does not require a reference image for makeup removal. In addition, for the application of high-fidelity texture inference, makeup-removed face images are preferably consistent with the high-resolution texture dataset that can achieve consistent normalized texture inference.

3. Data Preparation

We use two datasets. The makeup dataset was used to learn facial de-makeup. The high-resolution texture dataset was utilized to build the 3DMM and train the high-fidelity texture inference network.

3.1. Makeup Dataset

The current publicly available makeup dataset [LQD*18, JLG*19, GWC*19] is not ideal, because the dataset classified as non-makeup also contains many photos with light makeup. To efficiently remove the influence of makeup styles, we used the dataset created in the makeup transfer research LADN [GWC*19], because they have strong contrasts. This dataset contains 334 faces without makeup and 355 faces with makeup. In addition, synthetic makeup images were generated by blending the makeup style and face images without makeup. For more details about the LADN dataset, please refer to [GWC*19]. Owing to the contribution of LADN, we can train our network via weakly supervised learning using before makeup and synthetic makeup image pairs.

3.2. Scanned Face Dataset

To prepare a high-resolution texture dataset, we captured 180 Japanese females with neutral facial expressions using the ESPER LightCage with 55 Sony α7RIII cameras and polarizing lights. A head mesh was reconstructed using the RealityCapture software. Skin specular and diffuse components were separated by the cross-polarization technique using polarizing filters. Normal maps were generated using a photometric stereo technique. Finally, the head mesh and a set of diffuse, normal, roughness, and specular texture maps at 16K-resolution were obtained. However, the raw scan data involved artifacts. We asked the 3DCG artists to clean the scan data. In addition, the reconstructed head mesh was registered with the same topology during the clean-up process.

3.3. 3DMM Construction

We used the scanned head meshes and diffuse texture maps to create a linear PCA-based 3DMM [BV99]. To improve the representation of the expression, we manually made 236 blendshapes from the mean head. The constructed 3DMM contained 65,143 vertices and 130,000 faces. The shape S and appearance A of the 3DMM can be controlled by changing the parameters of identity α, expression β, and appearance δ.

$$\begin{aligned} S &= \bar{S} + B_{id}\alpha + B_{exp}\beta \\ A &= \bar{A} + B_a\delta \end{aligned} \tag{1}$$

where $\bar{S}$ and $\bar{A}$ are the mean shape and appearance, respectively. B_{id}, B_{exp}, and B_a are the identity basis, expression basis, and appearance basis vectors, respectively. In our method, we employ [DYX*19] as the backbone of our neural network to be optimized and regress the 3DMM coefficients $C(\alpha,\beta,\delta,R,t)$. The coefficient vector $C \in \mathbb{R}^{298}$ was composed of the parameters of shape identity $\alpha \in \mathbb{R}^{120}$, expression $\beta \in \mathbb{R}^{120}$, appearance $\delta \in \mathbb{R}^{52}$, rotation $R \in \mathbb{R}^3$, and translation $t \in \mathbb{R}^3$. In addition, spherical harmonics (SH) lighting is parameterized $\gamma \in \mathbb{R}^{27}$ in the case of lighting conditions.

We use two 3D face reconstruction networks F and E. We employ ResNet50 [HZRS16] as the backbone of network F trained with our 3DMM following [DYX*19] to estimate C and SH lighting. The network E is ResNet18 [HZRS16] architecture to estimate $\hat{C}$.

4. BareSkinNet

Given a makeup face image, the de-makeup and de-lighting network (BareSkinNet) generates a bare skin image with a 3D facial geometry and coarse 3D textures.

The training process of BareSkinNet is shown in Fig. 2. The framework comprises two parts. 1) De-makeup and de-lighting network G for removing makeup and lighting influences from the input image to generate a bare skin image (Sec. 4.1). 2) 3D face reconstruction network E and F for estimating the 3DMM coefficients and SH lighting (Sec. 4.2). The coefficients and rendered results are used to improve the capability of G. The loss function for BareSkinNet is represented as follows:

$$L_{BSN} = L_{DD} + L_{FR} \tag{2}$$

L_{DD} is a loss function for the de-makeup and de-lighting network, and L_{FR} is a loss function for the 3D face reconstruction process. The details of each loss are described in the following sections.

4.1. De-makeup and De-lighting network

We employ a U-Net [RFB15] architecture with skip connections for de-makeup and de-lighting network G to maintain the structure of the input face image. Using this architecture, we can remove makeup and lighting influences from the input face image while maintaining the identity information.

As shown in Fig. 2, we used the pairs of before and after makeup images from the LADN makeup dataset [GWC*19]. We remove makeup by minimizing the distance between the bare skin image and the before makeup face image. The loss function for the de-makeup and de-lighting is defined as follows:

$$L_{DD} = w_1 L_{photo} + w_2 L_{GAN} + w_3 L_{LPIPS} \tag{3}$$

L_{photo} is the L1 pixel loss, L_{GAN} is the adversarial loss calculated by PatchGAN [IZZE17] for realistic results, and L_{LPIPS} is the perceptual image patch similarity (LPIPS) metric loss [ZIE*18] preserved meaningful facial features. w_1, w_2, and w_3 are the weights for balancing each term. These losses are calculated from the bare skin image and the corresponding reference image.

The makeup loss proposed in BeautyGAN [LQD*18] has been widely used in makeup transfer tasks. In contrast to the makeup transfer task, we do not need to ensure consistency between the color distributions of different makeup styles. In addition, we remove the lighting influence in our framework by minimizing loss with 3D face reconstruction. For these reasons, we do not employ makeup loss. Note that the purpose of this network is to remove the makeup and lighting influences from the input image. However, the face image before the makeup used as the ground truth already contains lighting information. Therefore, it is difficult to remove the lighting influence using this network completely. We verified this claim in an ablation study (Sec. 7.3). To remove the influence of lighting, we incorporate 3D face reconstruction networks. In the next section, we discuss using the 3D face reconstruction process to achieve de-lighting.

4.2. De-lighting via 3D face reconstruction

To remove the lighting influence, we use 3D face reconstruction networks E and F to estimate the 3DMM coefficients and SH lighting from the bare skin image and reference image.

Since the LADN dataset [GWC*19] only has a small number of subjects, the pre-trained network F will be fixed as a teacher role to help the network E learn the 3D face reconstruction process. We expect that jointly learning the 3D face reconstruction process will improve the capability of network G to remove makeup and lighting influences. In the training stage, we jointly learn networks G and E with fixed F. We employ differentiable rendering to render the 3DMM so that BareSkinNet can be trained in an end-to-end fashion. Therefore, the networks G and E can be optimized simultaneously. In the inference stage, we use the network G to obtain the bare skin image, and then use F to obtain 3DMM coefficients. Because F is trained from a large dataset, which can perform a more accurate 3D reconstruction.

For the differentiable renderer DR, we use Nvdiffrast [LHK*20] to calculate the reconstruction loss and optimize the parameters of the network. γ is estimated from the pre-makeup reference image by F. The consistency loss between E and F results implicitly uses estimated SH lighting. After estimating 3DMM coefficients $\hat{C}$ from E and SH lighting from F, we render the 3DMM with and without lighting. Note that C is not used for 3DMM rendering. The rendered 3DMM image without lighting should be close to the bare skin image, and the rendered 3DMM image with lighting should be close to the reference image. By combining with the de-makeup and de-lighting network G, BareSkinNet can correctly remove the makeup and lighting influences. The following loss function is used for the 3D face reconstruction process to optimize network E.

$$\begin{aligned} L_{FR} = {} & w_{coeff} L_{coeff} + w_{land} L_{land} \\ & + w_{diff} L_{diff} + w_{light} L_{light} + w_{reg} L_{reg} \end{aligned} \tag{4}$$

L_{coeff} is the loss between 3DMM coefficients $\hat{C}$ and C estimated from networks E and F. L_{land} is the reprojection error of facial landmarks between the detected 2D landmarks from the reference image and the projected landmarks from the 3DMM. L_{diff} and L_{light} are the pixel-wise L1 distances between the images and the rendered 3DMM image with and without lighting, respectively. L_{reg} is the regularization term for the coefficients of the 3DMM. This regularization term is commonly used in the 3DMM-based face reconstruction process to avoid an unnatural face output. Concretely, L_{reg} is defined as the distance between the estimated and mean face coefficients. w_{coeff}, w_{land}, w_{diff}, w_{light} and w_{reg} are weights for balancing each term.

By combining the process of 3D face reconstruction, the lighting influence can be removed in the bare skin image, and the overall skin tone of the face region is matched with the 3DMM diffuse. The 3DMM diffuse retains global appearance information at a low frequency in the high-resolution texture dataset.

5. High-fidelity texture inference

This process is considered to be an application. We can apply high-fidelity texture inference methods on top of BareSkinNet. Similar to that reported in the literature [YSN*18, LBZ*20, LMG*20, LMP*21].

The high-fidelity texture inference network takes the bare skin image and the 3D face reconstruction result acquired by BareSkinNet as shown in Fig. 2. The bare skin image, 3DMM diffuse, 3DMM normal, and 3DMM geometry are unwarped to the same UV map. Then, the image-translation framework [WLZ*18] can infer diffuse, normal, roughness, and specular texture maps. The output of the image-translation framework is 1K-resolution. Finally, we use SRGAN [LTH*17] to upscale the 1K-resolution texture maps to 4K-resolution texture maps.

In the training process, to synthesize occlusions depending on the viewing angles, we rendered a scanned face model with diffuse from three different viewpoints: front, left, and right. The 3DMM is then fitted to the rendered image. Following the 3D face reconstruction result, the rendered image is unwarped to the UV map,

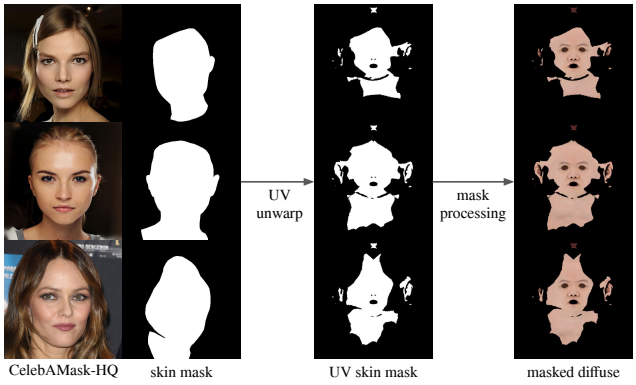

Figure 3: *Data augmentation using 2D skin masks. Skin masks are unwarped using the result of 3D face reconstruction. These unwarped skin masks are used to synthesize the occlusion effect.*

which is the same as the inference process. As a result of this process, we can obtain the pairs of the occluded high-frequency appearance information of face texture, the low-frequency appearance information of 3DMM, geometry, and normal, which correspond to high-resolution diffuse, normal, specular, and roughness texture maps. In addition, to improve the robustness against occlusion due to viewpoint change, we synthesize the occluding effect using 2D face masks as shown in Fig. 3. First, we carefully selected face images and corresponding skin mask images for 500 neutral faces from the celebAMask-HQ dataset [LLWL20], and the 3DMM was fitted to face images. Using the result of 3D face reconstruction, skin mask images were unwarped to the UV map. In each training process, unwarped visible masks were used to enhance the capacity to handle occlusion by multiplying it with the face texture.

6. Implementation detail

We implemented our pipeline with PyTorch. For the differentiable renderer, we use Nvdiffrast [LHK*20]. All training processes were performed on an NVIDIA RTX 2080Ti graphics card. The 3DMM was created using the scanned face dataset. In the de-makeup and de-lighting stage, the face images from the LADN dataset were resized to 256×256. 3D face reconstruction was performed under the same resolution and finally rendered to a 1024×1024 UV map representation. In the high-fidelity texture inference stage, the high-resolution texture contained in the scanned face dataset is resized to 4K-resolution. SRGAN is also trained under $4\times$ super-resolution conditions to upscale the image from 1K-resolution to obtain the final 4K-resolution result.

We set our balancing factors as the following: $w_1 = 100$, $w_2 = 1$, $w_3 = 40$, $w_{coeff} = 1e{-}1$, $w_{land} = 8e{-}2$, $w_{diff} = 100$, $w_{light} = 100$,$w_{reg} = 1e{-}3$. First, The BareSkinNet was trained for 10000 iterations only using L_{DD}. This can warm up the BareSkinNet and generate a stable bare skin image. At this stage, the BareSkinNet is split and only G is used. Then the BareSkinNet was trained in 30000 iterations with the full model. We set a batch size of 4 using the Adam optimizer to train our BareSkineNet. The learning rate of the de-makeup and de-lighting network G was set to $2e{-}5$ and exponential decay rates ($\beta 1$, $\beta 2$) = (0.5, 0.999). The learning rate of the 3D face reconstruction network E was set to $1e{-}4$.

7. Results and Evaluations

To demonstrate the effectiveness of our method, we conducted qualitative and quantitative evaluations. In the qualitative evaluation, we present the results of the bare skin images. In addition, we show the rendering results using the generated high-fidelity texture maps. By comparing it with the existing high-fidelity texture inference method [YSN*18][†] and the normalized avatar synthesis method [LNK*21][‡], we demonstrate the usefulness of our method against makeup. In the quantitative evaluation, we evaluated the stability of the generated texture maps.

7.1. Qualitative Evaluation

De-makeup and de-lighting

First, we compared the makeup removal effects with state-of-the-art methods [LQD*18, JLG*19, DHC*21, YHXG22, GWC*19]. As shown in Fig. 4, our BareSkinNet can remove the makeup and lighting influences without specifying a reference image or a known lighting condition. The state-of-the-art makeup transfer methods could not achieve makeup removal successfully. LADN [GWC*19] could remove makeup to some extent by carefully selecting a non-makeup reference image. However, the results were affected by the reference image and introduced new lighting. We dig deeper into the reasons for BareSkinNet effectiveness. We think the result of 3D face reconstruction can be regarded as a reference image. The diffuse of 3DMM contains ideal conditions without makeup and lighting. Additionally, the 3D face reconstruction process has consistency with the original image. Therefore, in contrast to using another person's reference image, our results preserve the subject's identity.

BareSkinNet can remove makeup and lighting influences correctly. Fig. 5 shows the results of BareSkinNet for the same subject. These results confirm that our method can produce a consistently clean face under different makeup and lighting conditions.

Fig. 6 shows the other results of the de-makeup and de-lighting. The makeup face images were obtained from the CPM-Real dataset [NTH21]. The CPM-Real dataset contains real-world makeup photos. From these results, our method can successfully remove the makeup and lighting influences from the light makeup images. In addition, even for the strong makeup images, the results of de-makeup and de-lighting are of reasonable quality.

Texture inference

Fig. 7 shows the results of the normalized texture inference. We selected five samples of face images from the LADN dataset. The

[†] We used the original implementation and pre-trained model provided by the authors.

[‡] Results were provided by the authors.

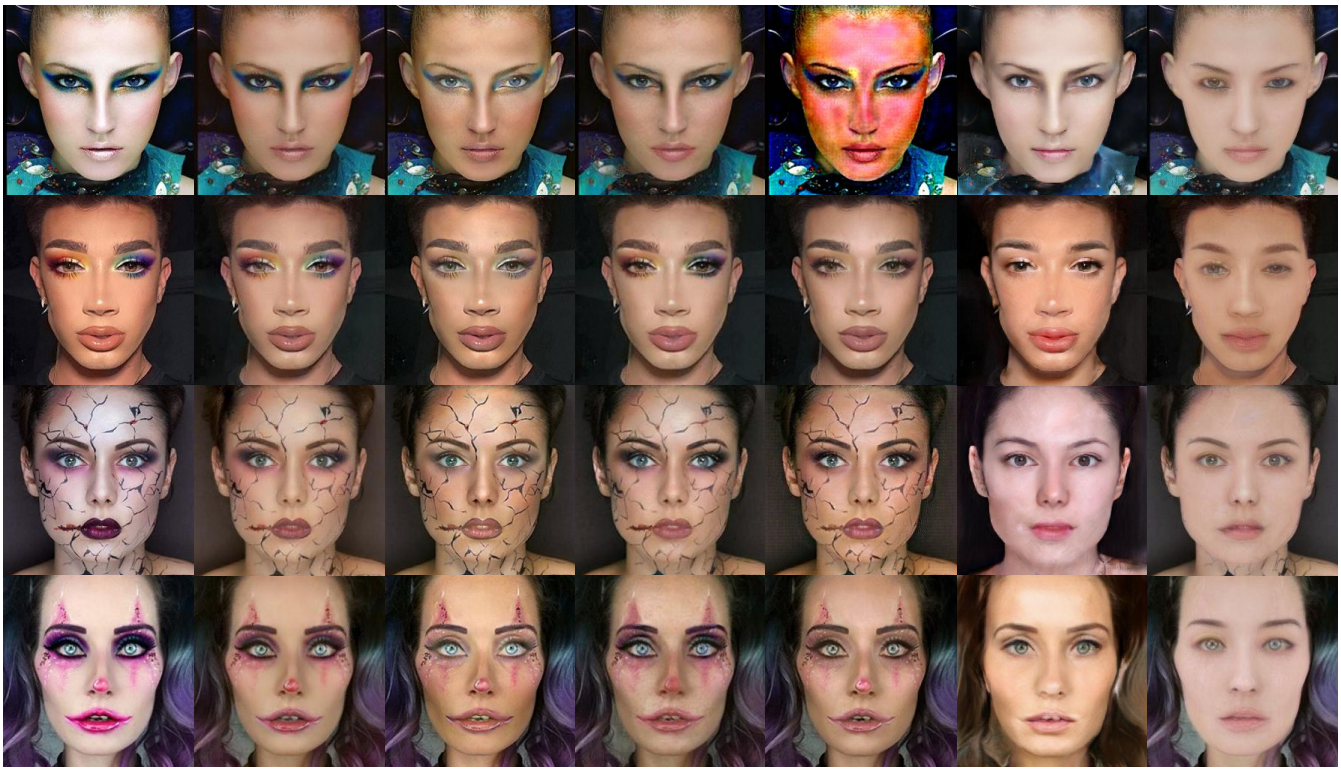

Figure 4: *Comparison with state-of-the-art makeup transfer methods for makeup removal. From left to right, we show (a) input face images; (b) BeautyGAN [LQD*18]; (c) PSGAN [JLG*19]; (d) SCGAN [DHC*21]; (e) EleGANt [YHXG22]; (f) results from the paper of LADN [GWC*19]; (g) our results of bare skin images.*

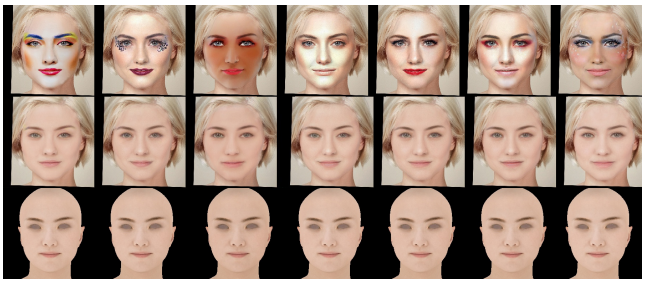

Figure 5: *BareSkinNet for the same subject. The first row is the input images. The second row is the bare skin images. The last row is the reconstructed 3DMM from the bare skin images.*

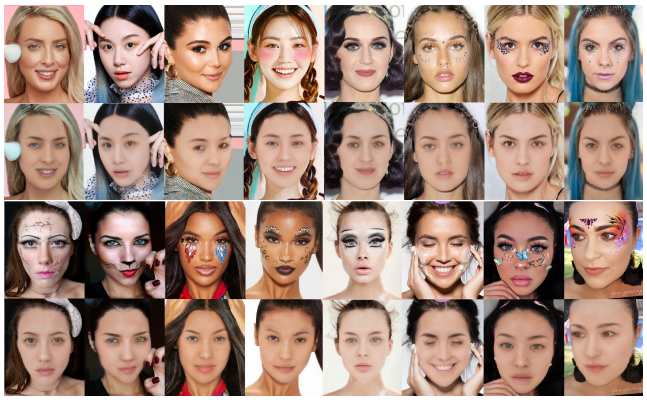

Figure 6: *Results for real-world makeup image from CPM-Real dataset [NTH21]. The first and third rows are the input face images (light makeup and strong makeup). The second and the fourth rows are the bare skin images using BareSkinNet.*

makeup face image was used as an input to BareSkinNet. BareSkinNet outputs a bare skin image and 3D face reconstruction result. The results confirm that BareSkinNet can successfully remove the makeup and lighting influence from the input image. The high-fidelity texture inference network was then executed using the outputs of BareSkinNet. We can confirm that the generated clean texture maps can be used for realistic face rendering.

To demonstrate the effectiveness of BareSkinNet for high-fidelity texture inference, we compared the output texture maps with and without BareSkinNet preprocessing. As input to the texture inference process, with and without of BareSkinNet samples colors from the makeup image and the bare skin image, respec-

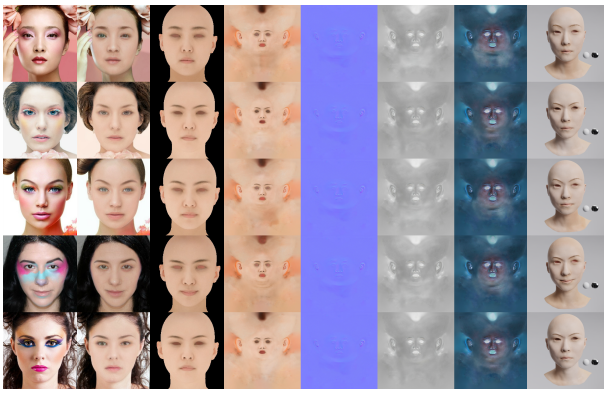

Figure 7: *Example results of normalized texture inference. From left to right, we show (a) makeup images; (b) recovered bare skin images; (c) 3D face reconstruction results; (d) inferred diffuse, normal, roughness and specular texture maps; (e) rendering results using inferred texture maps.*

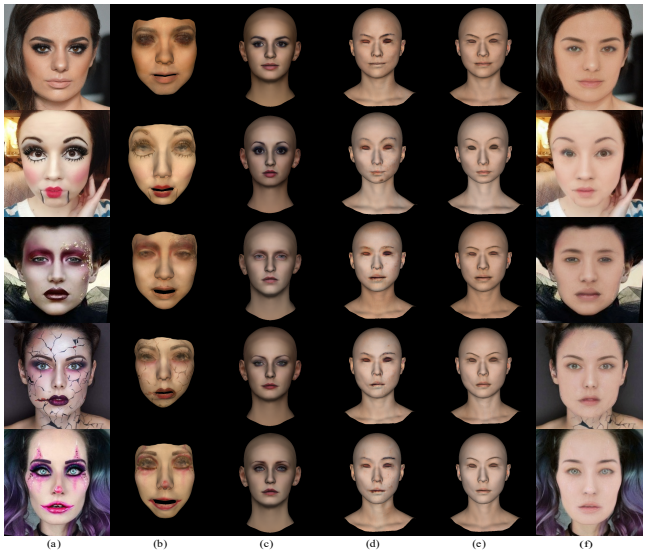

Figure 8: *Qualitative comparison with state-of-the-art methods. From left to right, we present (a) input images; (b) results of Yamaguchi et al. [YSN*18]; (c) Luo et al. [LNK*21]; (d) ours without BareSkinNet; (e) ours; (f) output of BareSkinNet. Result images are rendered using inferred diffuse texture.*

tively. The entire 3D reconstruction process is the same except for the color sampling.

Figs. 8 (d) and (e) show the rendered results using the inferred diffuse texture maps with inputs (a) and (f), respectively. We can confirm artifacts in the generated diffuse texture maps in the case without BareSkinNet. On the other hand, using the BareSkinNet output, the generated diffuse texture maps do not include any artifacts.

We also compared state-of-the-art texture inference methods [YSN*18, LNK*21], as shown in Fig. 8 (b) and (c). For comparison, we only used the diffuse texture map for rendering. In the results of Yamaguchi et al. [YSN*18], the entire makeup texture patterns remained in the inferred diffuse texture maps. Although Luo et al. [LNK*21] could remove some makeup texture patterns, makeup effects remain around the eyebrow, eyes, and mouth. In contrast to these methods, our method successfully removed the makeup patterns for the entire face texture.

Table 1: *Quantitative evaluation of the output texture maps from the facial details inference network. Metrics are root mean square error (RMSE), peak signal-to-ratio (PSNR), and structural similarity index measure (SSIM).*

	w/o BareSkinNet	w/ BareSkinNet
RMSE (Diffuse)	3.910	**3.358**↓
RMSE (Normal)	1.365	**1.244**↓
RMSE (Specular)	3.997	**3.146**↓
RMSE (Roughness)	1.782	**1.242**↓
PSNR (Diffuse)	33.332	**36.167**↑
PSNR (Normal)	45.541	**46.252**↑
PSNR (Specular)	35.188	**37.843**↑
PSNR (Roughness)	43.850	**46.296**↑
SSIM (Diffuse)	0.963	**0.969**↑
SSIM (Normal)	0.976	**0.980**↑
SSIM (Specular)	0.943	**0.961**↑
SSIM (Roughness)	0.981	**0.985**↑

7.2. Quantitative Evaluation

Since it is hard to acquire ground truth data of in-the-wild de-makeup and de-lighting images, it is challenging to evaluate BareSkinNet directly. We evaluated the effectiveness of BareSkinNet by comparing the final outputs of the texture maps from the texture inference network. We randomly selected 70 subject faces with various synthetic 2100 makeup images that were not included in the training set. For comparison, we obtained the final texture maps with and without BareSkinNet. We used the output texture map of the before-makeup image inputs as a reference for computing the error metrics. We then computed the root mean square error (RMSE), peak signal-to-noise ratio (PSNR), and structural similarity index measure (SSIM) as metrics between the generated UV maps. Table 1 lists the results for each score for each texture map. Compared with the texture maps without BareSkinNet, the errors of RMSE were reduced after applying BareSkinNet. In addition, the results confirm that the PSNR and SSIM metrics are improved. From these results, we believe that our BareSkinNet can improve the stability of 3D face reconstruction under various makeup and lighting conditions.

7.3. Ablation Study

To validate the effectiveness of each component in BareSkinNet, we conducted experiments with different submodules and losses. Example results are presented in Fig. 9.

First, we only use de-makeup and de-lighting network G results

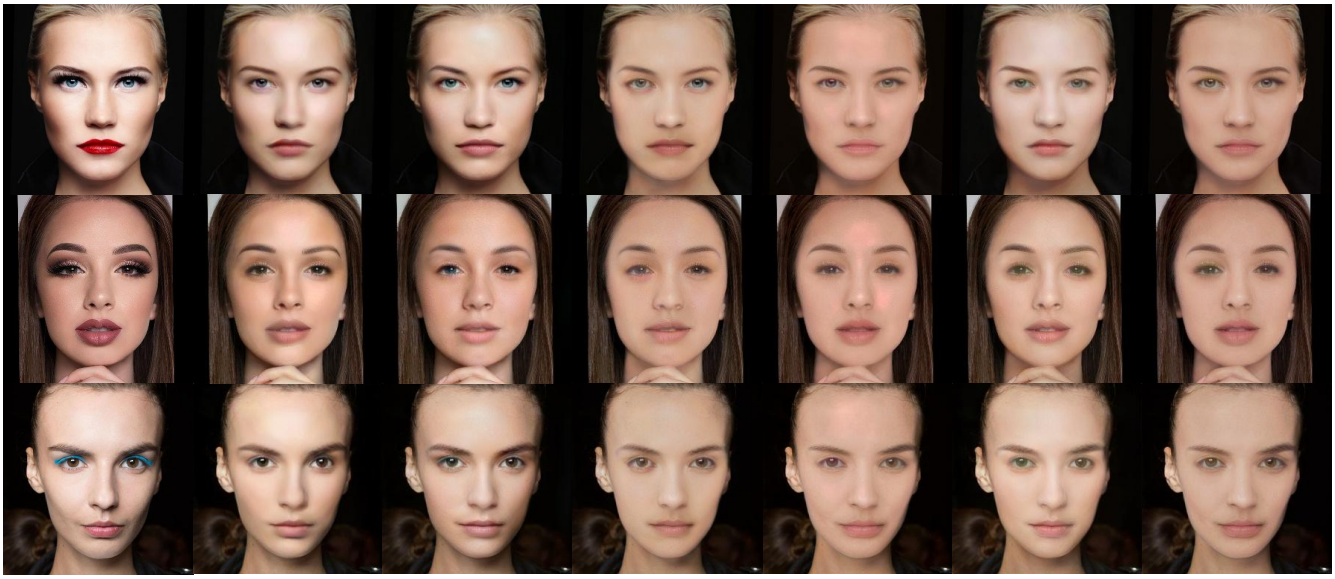

Figure 9: *Results of ablation study. (a) input images; (b) L_{photo} and L_{GAN}; (c) L_{photo}, L_{GAN} and L_{LPIPS}; (d) use F instead of E without fine-tuning; (e) use F instead of E with fine-tuning; (f) our proposed teacher-student framework without L_{light}; (g) our full model.*

in supervised learning. As shown in (b) and (c), makeup effects were removed in most parts. However, some effects remain around the eyes. Utilizing L_{LPIPS} can add some details and make the face clearer.

Next, we added a pre-trained 3D face reconstruction network F. We verified the effectiveness of removing lighting and makeup with and without fine-tuning. As shown in (d) and (e), makeup around the eyes is better removed. Different from (b) and (c), we can confirm that the lighting effect is removed. However, some artifacts appeared in (e) around the illuminated area.

Finally, we went a step further to improve the capability of removing makeup and lighting influences by adding a network E that can be trained. We found that the best performance can be achieved by letting network E learn the diffuse part while the pre-trained network F provides accurate SH lighting estimates. In addition, network E is difficult to train without the support of a teacher-student strategy due to the limitation of the diversity of the makeup dataset. Comparing (f) and (g), by adding L_{light}, the lighting effect was removed significantly.

8. Limitations

Although the results of qualitative and quantitative experiments confirm that our method performed excellently under makeup and lighting conditions, it is challenging to handle largely inclined face poses and extreme facial expressions, especially a face with closed eyes or opened mouth. Fig. 10 shows examples of the failure cases. BareSkinNet cannot produce desirable makeup removal results because BareSkinNet was trained using a front neutral expression face collection dataset.

Also, In the makeup removal process, eyes and hair are affected by the color of 3DMM, which is a limitation. Although hair color and eye color changed, the result of the application in this study is not affected by these factors because these regions are automatically excluded by facial skin area segmentation and the reconstructed 3DMM.

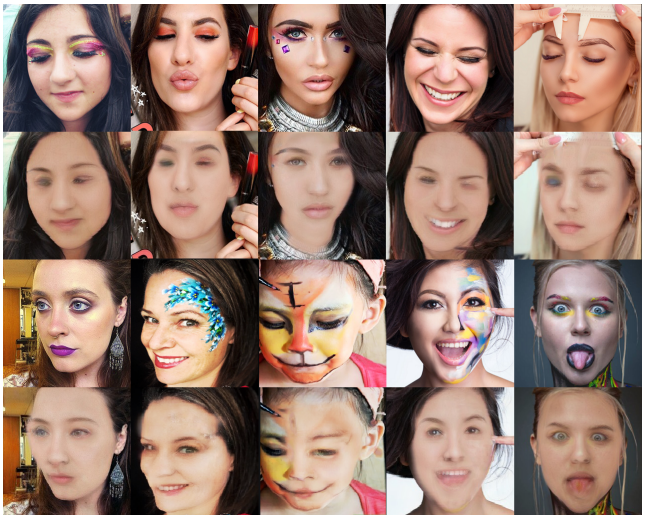

Figure 10: *Examples of the failure cases of de-makeup*

Our scan dataset has limitations in the diversity of faces, leading to similarities in the reconstructed shapes and texture maps. For the BareSkinNet module, we believe the BFM [PKA*09] or FLAME [LBB*17] model can be employed, which is more suitable for non-Asian faces, instead of our original 3DMM. But when considering subsequent applications, we propose that 3DMM origi-

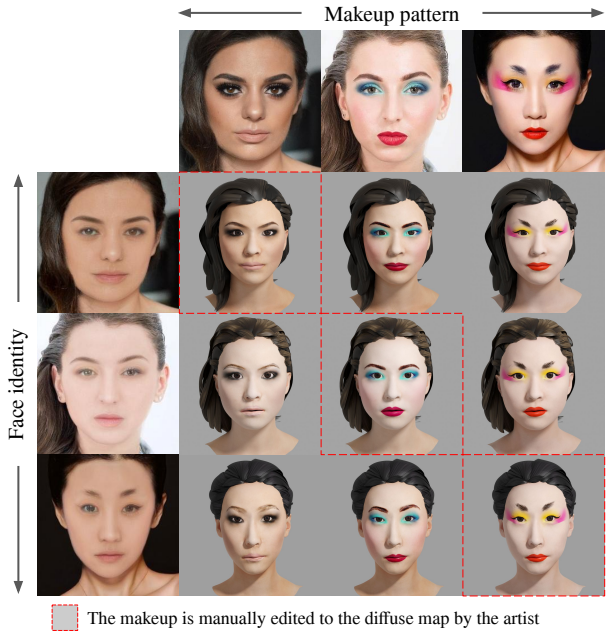

Figure 11: *Examples of makeup transfer. Images with the red square show makeup results by editing diffuse texture maps by the 3DCG artist. Other images show the results of makeup transfer.*

nates from the same scan dataset with high-resolution texture maps because consistency between 3DMM, bare skin image, and texture maps are preferred.

9. Conclusions and Future Work

We presented BareSkinNet, a framework to remove makeup and lighting influences from the face image input to reconstruct a high-fidelity 3D face model. Using our BareSkinNet as a preprocessing step for 3D face reconstruction, we can obtain consistent results of high-fidelity texture maps for the same subject. Through experiments, we confirmed that our approach could be successfully applied to various makeup image inputs.

Fig. 11 shows an example of makeup editing. The reconstructed face models can be found in Figs. 1 and 8. In this example, we asked a 3DCG artist to put makeup on the reconstructed face models by editing the inferred diffuse texture maps. Note that the 3DCG artist also created eyes and hairs. Created models are marked with the red square in the figure. A makeup layer was extracted by subtracting the reconstructed and edited diffuse texture. The makeup transfer was then achieved by adding the reconstructed diffuse texture maps and extracted makeup layers. The images without the red square show the results of makeup transfer. These results confirm the possibility of creating photo-realistic avatars in various makeup styles with minor effort.

This study focused on de-makeup and de-lighting for 3D face reconstruction. Currently, our system cannot separate makeup style and skin color. In the future, we plan to extend our method to extract the makeup layer and use it for 3D makeup reconstruction and transfer.

In addition, there is a potential to use BareSkinNet for other tasks such as face recognition and face verification. It can also be used to improve the accuracy of 3DMM texture space, for example, the texture creation process of the FLAME [LBB*17] model. We will investigate the availability of our method.

Acknowledgements

We thank Vladlen Erium for providing the 3D scan dataset, as well as Vladislava Mironenko for her help with 3D makeup editing. This work was supported by CyberHuman Productions.

References

[BLC*21] Bao L., Lin X., Chen Y., Zhang H., Wang S., Zhe X., Kang D., Huang H., Jiang X., Wang J., Yu D., Zhang Z.: High-fidelity 3d digital human head creation from rgb-d selfies. *ACM Transactions on Graphics* (2021). 2, 3

[BRZ*16] Booth J., Roussos A., Zafeiriou S., Ponniah A., Dunaway D.: A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), IEEE, pp. 5543–5552. 2

[BV99] Blanz V., Vetter T.: A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques* (1999), ACM, p. 187–194. 2, 4

[BV03] Blanz V., Vetter. T.: Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25*, 9 (2003), 1063–1074. 2

[CBK*17] Cole F., Belanger D., Krishnan D., Sarna A., Mosseri I., Freeman W.: Synthesizing normalized faces from facial identity features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (07 2017), IEEE, pp. 3386–3395. 3

[CWZ*14] Cao C., Weng Y., Zhou S., Tong Y., Zhou K.: Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics 20*, 3 (2014), 413–425. 2

[DBB22] Danecek R., Black M. J., Bolkart T.: EMOCA: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2022). 2

[DCX*18] Deng J., Cheng S., Xue N., Zhou Y., Zafeiriou S.: Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (06 2018), IEEE, pp. 7093–7102. 3

[DHC*21] Deng H., Han C., Cai H., Han G., He S.: Spatially-invariant style-codes controlled makeup transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), IEEE, pp. 6549–6557. 4, 6, 7

[DHT*00] Debevec P., Hawkins T., Tchou C., Duiker H.-P., Sarokin W., Sagar M.: Acquiring the reflectance field of a human face. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques* (2000), ACM, p. 145–156. 2

[DMJ*21] Deng Q., Ma L., Jin A., Bi H., Le B. H., Deng Z.: Plausible 3d face wrinkle generation using variational autoencoders. *IEEE Transactions on Visualization and Computer Graphics* (2021), 1–1. 2

[DYX*19] Deng Y., Yang J., Xu S., Chen D., Jia Y., Tong X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)W* (2019), IEEE. 2, 4

[EST*20] Egger B., Smith W. A. P., Tewari A., Wuhrer S., Zollhoefer M., Beeler T., Bernard F., Bolkart T., Kortylewski A., Romdhani S., Theobalt C., Blanz V., Vetter T.: 3d morphable face models—past, present, and future. *ACM Transactions on Graphics* (2020). 2

[FFBB21] FENG Y., FENG H., BLACK M. J., BOLKART T.: Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics 40*, 4 (jul 2021). 2

[GCM*18] GENOVA K., COLE F., MASCHINOT A., SARNA A., VLASIC D., FREEMAN W. T.: Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), IEEE, pp. 8377–8386. 2

[GDZ21] GECER B., DENG J., ZAFEIRIOU S.: Ostec: One-shot texture completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021), pp. 7628–7638. 3

[GGU*20] GHOSH P., GUPTA P. S., UZIEL R., RANJAN A., BLACK M., BOLKART T.: Gif: Generative interpretable faces. *Proceedings of International Conference on 3D Vision* (2020), 868–878. 3

[GPKZ19] GECER B., PLOUMPIS S., KOTSIA I., ZAFEIRIOU S.: Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), IEEE, pp. 1155–1164. 2, 3

[GWC*19] GU Q., WANG G., CHIU M. T., TAI Y.-W., TANG C.-K.: Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In *Proceedings of the International Conference on Computer Vision (ICCV)* (2019), IEEE, pp. 10481–10490. 2, 4, 5, 6, 7

[GZC*18] GUO Y., ZHANG J., CAI J., JIANG B., ZHENG J.: Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence PP* (05 2018), 1–1. 2

[GZC*19] GUO Y., ZHANG J., CAI J., JIANG B., ZHENG J.: Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 41*, 6 (2019), 1294–1307. 2

[HCS*18] HUYNH L., CHEN W., SAITO S., XING J., NAGANO K., JONES A., DEBEVEC P., LI H.: Mesoscopic facial geometry inference using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), IEEE, pp. 8407–8416. 2

[HZLH17] HUANG R., ZHANG S., LI T., HE R.: Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the International Conference on Computer Vision (ICCV)* (2017), IEEE, pp. 2439–2448. 3

[HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778. 4

[HZS*21] HOU A., ZHANG Z., SARKIS M., BI N., TONG Y., LIU X.: Towards high fidelity face relighting with realistic shadows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), IEEE, pp. 14719–14728. 3

[IZZE17] ISOLA P., ZHU J.-Y., ZHOU T., EFROS A. A.: Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017). 5

[JBAT17] JACKSON A. S., BULAT A., ARGYRIOU V., TZIMIROPOULOS G.: Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the International Conference on Computer Vision (ICCV)* (2017), IEEE, pp. 1031–1039. 2

[JLG*19] JIANG W., LIU S., GAO C., CAO J., HE R., FENG J., YAN S.: Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), IEEE, pp. 5194–5202. 4, 6, 7

[Kem13] KEMELMACHER I.: Internet based morphable model. In *Proceedings of the International Conference on Computer Vision (ICCV)* (2013), IEEE, pp. 3256–3263. 2

[KLA*20] KARRAS T., LAINE S., AITTALA M., HELLSTEN J., LEHTINEN J., AILA T.: Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), IEEE, pp. 8110–8119. 4

[KSB11] KEMELMACHER-SHLIZERMAN I., BASRI R.: 3d face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*, 2 (2011), 394–405. 2

[KYT21] KIM J., YANG J., TONG X.: Learning high-fidelity face texture completion without complete face texture. In *Proceedings of the International Conference on Computer Vision (ICCV)* (2021), IEEE, pp. 13970–13979. 3

[LBB*17] LI T., BOLKART T., BLACK M. J., LI H., ROMERO J.: Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics 36*, 6 (2017). 2, 9, 10

[LBZ*20] LI R., BLADIN K., ZHAO Y., CHINARA C., INGRAHAM O., XIANG P., REN X., PRASAD P., KISHORE B., XING J., LI H.: Learning formation of physically-based face attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), IEEE, pp. 3410–3419. 2, 5

[LHK*20] LAINE S., HELLSTEN J., KARRAS T., SEOL Y., LEHTINEN J., AILA T.: Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics 39*, 6 (2020). 5, 6

[LL20] LEE G.-H., LEE S.-W.: Uncertainty-aware mesh decoder for high fidelity 3d face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 6099–6108. 3

[LLWL20] LEE C.-H., LIU Z., WU L., LUO P.: Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), IEEE. 6

[LMG*20] LATTAS A., MOSCHOGLOU S., GECER B., PLOUMPIS S., TRIANTAFYLLOU V., GHOSH A., ZAFEIRIOU S.: Avatarme: Realistically renderable 3d facial reconstruction "in-the-wild". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), IEEE, pp. 760–769. 2, 3, 5

[LMP*21] LATTAS A., MOSCHOGLOU S., PLOUMPIS S., GECER B., GHOSH A., ZAFEIRIOU S. P.: Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–1. 3, 5

[LNK*21] LUO H., NAGANO K., KUNG H.-W., GOLDWHITE M., XU Q., WANG Z., WEI L., HU L., LI H.: Normalized avatar synthesis using stylegan and perceptual refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), IEEE, pp. 11662–11672. 2, 3, 6, 8

[LQD*18] LI T., QIAN R., DONG C., LIU S., YAN Q., ZHU W., LIN L.: Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *Proceedings of ACM International Conference on Multimedia* (2018), ACM, p. 645–653. 4, 5, 6, 7

[LTH*17] LEDIG C., THEIS L., HUSZÁR F., CABALLERO J., CUNNINGHAM A., ACOSTA A., AITKEN A., TEJANI A., TOTZ J., WANG Z., SHI W.: Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), IEEE, pp. 4681–4690. 3, 5

[LZG*21] LIN J., ZHANG R., GANZ F., HAN S., ZHU J.-Y.: Anycost gans for interactive image synthesis and editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), IEEE, pp. 14986–14996. 3

[NLW*19] NAGANO K., LUO H., WANG Z., SEO J., XING J., HU L., WEI L., LI H.: Deep face normalization. *ACM Transactions on Graphics 38*, 6 (2019). 2, 3

[NTH21] NGUYEN T., TRAN A., HOAI M.: Lipstick ain't enough: Beyond color matching for in-the-wild makeup transfer. In *Proceedings*

of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021), IEEE, pp. 13305–13314. 4, 6, 7

[PEL*21] PANDEY R., ESCOLANO S. O., LEGENDRE C., HÄNE C., BOUAZIZ S., RHEMANN C., DEBEVEC P., FANELLO S.: Total relighting: Learning to relight portraits for background replacement. *ACM Transactions on Graphics 40*, 4 (2021). 3

[PKA*09] PAYSAN P., KNOTHE R., AMBERG B., ROMDHANI S., VETTER T.: A 3d face model for pose and illumination invariant face recognition. In *Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance* (2009), IEEE, pp. 296–301. 2, 9

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of Medical Image Computing and Computer-Assisted Intervention* (2015), pp. 234–241. 5

[RGB*20] RIVIERE J., GOTARDO P., BRADLEY D., GHOSH A., BEELER T.: Single-shot high-quality facial geometry and skin appearance capture. *ACM Trans. Graph. 39*, 4 (2020). 2, 3

[RSOEK17] RICHARDSON E., SELA M., OR-EL R., KIMMEL R.: Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), IEEE, pp. 1259–1268. 2

[RTD*21] R M. B., TEWARI A., DIB A., WEYRICH T., BICKEL B., SEIDEL H.-P., PFISTER H., MATUSIK W., CHEVALLIER L., ELGHARIB M., THEOBALT C.: Photoapp: Photorealistic appearance editing of head portraits, 2021. 3

[SBT*19] SUN T., BARRON J. T., TSAI Y.-T., XU Z., YU X., FYFFE G., RHEMANN C., BUSCH J., DEBEVEC P., RAMAMOORTHI R.: Single image portrait relighting. *ACM Transactions on Graphics 38*, 4 (2019). 3

[SKCJ18] SENGUPTA S., KANAZAWA A., CASTILLO C. D., JACOBS D. W.: Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018). 3

[SRH*11] SCHERBAUM K., RITSCHEL T., HULLIN M., THORMÄHLEN T., BLANZ V., SEIDEL H.-P.: Computer-suggested facial makeup. *Comp. Graph. Forum (Proc. Eurographics 2011) 30*, 2 (2011). 2

[SRK17] SELA M., RICHARDSON E., KIMMEL R.: Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the International Conference on Computer Vision (ICCV)* (2017), IEEE, pp. 1576–1585. 2

[SSD*20] SMITH W. A. P., SECK A., DEE H., TIDDEMAN B., TENENBAUM J., EGGER B.: A morphable face albedo model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 5011–5020. 2

[SWH*17] SAITO S., WEI L., HU L., NAGANO K., LI H.: Photorealistic facial texture inference using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), IEEE, pp. 5144–5153. 3

[SXZ*20] SUN T., XU Z., ZHANG X., FANELLO S., RHEMANN C., DEBEVEC P., TSAI Y.-T., BARRON J. T., RAMAMOORTHI R.: Light stage super-resolution: Continuous high-frequency relighting. *ACM Transactions on Graphics 39*, 6 (2020). 2

[TEB*20] TEWARI A., ELGHARIB M., BHARAJ G., BERNARD F., SEIDEL H.-P., PÉREZ P., ZOLLHÖFER M., THEOBALT C.: Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), IEEE, pp. 6142–6151. 3

[TER*20] TEWARI A., ELGHARIB M., R. M. B., BERNARD F., SEIDEL H.-P., PÉREZ P., ZOLLHÖFER M., THEOBALT C.: Pie: Portrait image embedding for semantic control. *ACM Transactions on Graphics 39*, 6 (2020). 3

[THMM17] TRAN A. T., HASSNER T., MASI I., MEDIONI G.: Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), IEEE, pp. 5163–5172. 2

[TL18] TRAN L., LIU X.: Nonlinear 3d face morphable model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), IEEE, pp. 7346–7355. 2

[TZG*18] TEWARI A., ZOLLHÖFER M., GARRIDO P., BERNARD F., KIM H., PÉREZ P., THEOBALT C.: Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), IEEE, pp. 2549–2559. 2

[TZK*17] TEWARI A., ZOLLHÖFER M., KIM H., GARRIDO P., BERNARD F., PÉREZ P., THEOBALT C.: Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the International Conference on Computer Vision (ICCV)* (2017), IEEE, pp. 3715–3724. 2

[WCY*22] WANG L., CHEN Z., YU T., MA C., LI L., LIU Y.: Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2022). 2

[WLZ*18] WANG T.-C., LIU M.-Y., ZHU J.-Y., TAO A., KAUTZ J., CATANZARO B.: High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), IEEE, pp. 8798–8807. 5

[WRV20] WU S., RUPPRECHT C., VEDALDI A.: Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the International Conference on Computer Vision (ICCV)* (2020), IEEE, pp. 1–10. 2

[WYL*20] WANG Z., YU X., LU M., WANG Q., QIAN C., XU F.: Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Transactions on Graphics 39*, 6 (2020). 3

[YFD*21] YANG G., FEI N., DING M., LIU G., LU Z., XIANG T.: L2m-gan: Learning to manipulate latent space semantics for facial attribute editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), IEEE, pp. 2951–2960. 3

[YHXG22] YANG C., HE W., XU Y., GAO Y.: Elegant: Exquisite and locally editable gan for makeup transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2022). 4, 6, 7

[YJRF20] YIN Y., JIANG S., ROBINSON J. P., FU Y.: Dual-attention gan for large-pose face frontalization. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition* (2020), IEEE, pp. 249–256. 3

[YSN*18] YAMAGUCHI S., SAITO S., NAGANO K., ZHAO Y., CHEN W., OLSZEWSKI K., MORISHIMA S., LI H.: High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics 37*, 4 (2018). 2, 3, 5, 6, 8

[YYS*17] YIN X., YU X., SOHN K., LIU X., CHANDRAKER M.: Towards large-pose face frontalization in the wild. In *Proceedings of the International Conference on Computer Vision (ICCV)* (2017), IEEE, pp. 3990–3999. 3

[YZW*20] YANG H., ZHU H., WANG Y., HUANG M., SHEN Q., YANG R., CAO X.: Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), IEEE, pp. 601–610. 2

[ZBT*20] ZHANG X., BARRON J. T., TSAI Y.-T., PANDEY R., ZHANG X., NG R., JACOBS D. E.: Portrait shadow manipulation. *ACM Transactions on Graphics 39*, 4 (2020). 3

[ZHSJ19] ZHOU H., HADAP S., SUNKAVALLI K., JACOBS D.: Deep single-image portrait relighting. In *Proceedings of the International Conference on Computer Vision (ICCV)* (2019), IEEE, pp. 7193–7201. 3

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), IEEE, pp. 586–595. 5