LOW-FREQUENCY GUIDED SELF-SUPERVISED LEARNING FOR HIGH-FIDELITY 3D FACE RECONSTRUCTION IN THE WILD

Pengrui Wang[1,2] *Chunze Lin*[3] *Bo Xu*[1] *Wujun Che*[1] *Quan Wang*[3]

[1] Institute of Automation, Chinese Academy of Sciences, China
[2] University of Chinese Academy of Sciences, China
[3] Sensetime Research

wangpengrui2015@ia.ac.cn, linchunze@sensetime.com, {xubo, wujun.che}@ia.ac.cn, wangquan@sensetime.com

ABSTRACT

In this paper, we propose a low-frequency guided self-supervised learning method for high-fidelity 3D face reconstruction from an in-the-wild image. Unlike other self-supervised methods only using the color difference between the original image and the estimated image, we add low-frequency albedo information to enhance the self-supervised learning for more realistic albedo while insensitive to the non-skin regions. Specifically, based on a PCA albedo model, we first train a Boosting Network (B-Net) to provide illumination and intact albedo distribution. Then with above information, we learn an image-to-image non-linear Facial Albedo Network (FAN) by self-supervision to produce a high-fidelity albedo. We further propose a Detail Recovering Network (DRN) to recover geometric details such as wrinkles. FAN and DRN permit to reconstruct 3D faces with high-fidelity albedo and geometry details. Finally, experimental results demonstrate the effectiveness of the proposed method.

Index Terms— 3D Face Reconstruction, Self-supervision, Facial albedo, Facial Geometric Details

1. INTRODUCTION

Reconstruction of the 3D digital faces from single images is an important problem at the intersection of computer vision and computer graphics. The increasing availability of facial images makes monocular face reconstruction have numerous practical applications [1] such as face editing, 3D avatar generation, virtual and augmented reality.

Facial skin albedos also called reflectances play an important role in lively digital faces after facial shapes are created. Some methods use textures standing for albedos [2, 3]. This cause the re-rendered faces unrealistic when re-rendering under different illumination and projections because the illumination and geometry details have baked into albedos. Thus, linear low-frequency albedo models, such as PCA models in 3D Morphable Models (3DMM) [4], are commonly used to help to separate the albedo and illumination [5, 6, 7, 8, 9]. However, PCA-based models are usually learned from scan databases of limited size [4] and lie in a low-dimensional linear subspace, which means those models usually imitate low-frequency albedos and have limited expression abilities.

Because of the scarce albedo labels, some deep learning methods use non-linear encoder-decoder albedos models trained by self-supervised learning [10, 11, 12]. Limited by the performance of the models and regularization methods used for invisible parts completion and occlusion insensitivity, they usually unable capture realistic albedos and details. Some optimization-based methods use Generative Adversarial Network (GAN) to model the facial skin [13] or model the facial texture as a convex combination of "style" features extracted from high-resolution face database [14]. These methods do not provide the corresponding fine-scale geometric details and their optimization procedures are usually time-consuming. Recently, supervised methods such as [15] use incomplete texture UV maps to infer complete albedo maps. They use image-to-image models to capture high-fidelity albedos and geometry details from the input textures. However, collecting and annotating (registration) 3D scans of face geometry and albedos are extremely laborious and expensive. Moreover, the image-to-image models in [15] need face region detection methods to remove the occlusion area caused by hands, hair, glasses, etc. Hence, they produce obvious non-skin objects when skin detection is inaccurate.

In this paper, we propose a low-frequency guided self-supervised learning method for inferring complete and high-fidelity facial albedo and geometric details from a single in-the-wild face image. Different from most existing self-supervised learning methods which minimize the difference between rendered reconstruction faces and the input faces, we absorb the advantages of both linear models and non-linear approaches. The linear face models, usually based on PCA models, in contrast to the non-linear deep learning method, can generate complete facial albedo even in occlu-

This work is supported by Beijing Municipal Natural Science Foundation (No. L192005) and National Natural Science Foundation of China (No. 61471359).

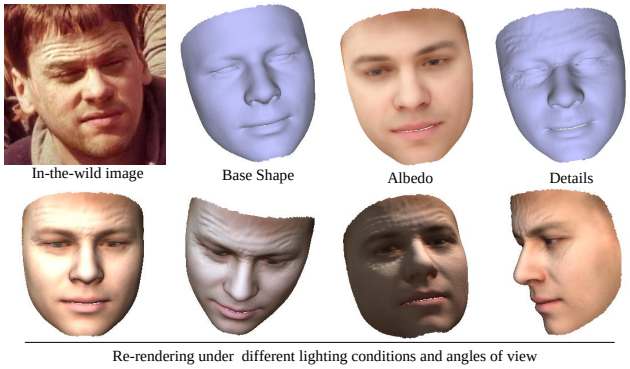

Fig. 1. The proposed method can reconstruct appropriate facial albedo and geometry details from a in the wild image.

sion or large pose situations with invisible facial parts. Although the image-to-image deep networks tend to generate albedo with occlusion (*e.g.* sunglasses) or incomplete regions, they can produce more realistic albedos that resemble the input faces.

To be more specific, in this work, we firstly train a Boosting Network (B-Net) by self-supervision to estimate illumination and coefficients of a PCA albedo model. With these boost information, we train our Facial Albedo Network (FAN) with both self-supervised signals which force the network to generate facial albedos that resemble as much as possible to input faces and the auxiliary low-frequency albedos which guide our FAN learn intact albedos. The FAN trained in this fashion can output complete and high-fidelity facial albedos despite in-the-wild input faces with different degrees of occlusion or pose. Moreover, we further present a Detail Recovering Network (DRN) to estimate the facial shape details. Combining results from FAN and DRN, we can effectively reconstruct high-fidelity 3D faces that can be re-rendered under different lighting conditions and angles of view, as illustrated in Fig. 1. We conduct extensive experiments to demonstrate the effectiveness of the proposed method. Qualitative comparison with state-of-the-art face reconstruction approaches shows that the 3D facial albedo and geometry generated by our method are more faithful to the input image.

2. RELATED WORK

Albedo Inference: PCA models are commonly used to represent facial albedos [16, 4]. Many regression methods using supervised learning rely heavily on 3DMM models which are used to synthesize labels to replace the scarce ground-truth labels [17, 9, 18]. Self-supervised or weak-supervised learning methods need no synthetic data, but 3DMM models are still needed to limit the albedo space [6, 5]. To overcome the low dimension space problem of 3DMM, some approaches add non-linear corrective layers [10] or construct non-linear 3DMM [12]. But they need regularization functions to guarantee legal results. These functions including skin symmetry to complete the facial albedos and skin constancy to estimate albedo by noise suppression lead the model to learn low-frequency albedos. There are also optimization-based approaches to estimate high-fidelity albedos [14, 13]. However, they need high-quality face image databases, and the optimization-based approaches are sensitive to super parameters and time-consuming. Recently, the high-fidelity image-to-image facial albedo inference methods [15] are proposed. But they need facial scans which are expensive to be acquired.
Details Inference: Facial details such as wrinkles and crows feet are hard to be captured on the shape based on 3DMM fitting or regression methods. Many approaches are proposed to solve this problem. Some add additional models like corrective models [19] or trainable corrective networks [10]. Some generate bump maps [20], displacement maps [21] which trained by supervised learning. Besides, they don't take into consideration the details on albedos. There are also 3DMM-unrelated 3D shape representations which have more degree of freedoms [22, 23, 7], but they are less robust and realistic than the methods built upon 3DMM. [24].

3. APPROACH

Our method aims to infer high-quality facial albedos and geometry details from in-the-wild images and train the models without facial scans data. The overview of our approach is shown in Fig. 2 where the sequence numbers (a)-(g) indicate the sequence of operations. Firstly, we introduce the shape estimation method to generate texture maps and normals. By training a non-linear image-to-image network in a self-supervised manner, however, it is unable to infer the invisible and occluded albedo regions from in-the-wild face image. To address this issue, we propose the B-Net which provides the auxiliary information for training facial albedo network. The illumination and complement albedo maps will help to train FAN by self-supervised leaning. To capture the geometry fine-details from the face image, we propose the DRN to infer the details on the geometry shape. The details in geometry rather than reflecting on the albedo make our reconstructed faces more realistic and consistent under different views and illumination as it was shown in Fig. 1.

3.1. Shape Estimation

We estimate shapes by fitting 3DMM, Basel Face Model (BFM) from [25], with facial landmarks. Specifically, after landmark detection, we get the base shape by finding the parameters that minimize the reprojection error on landmarks:

$$\arg\min_{\mathbf{s},\mathbf{R},\mathbf{t},\alpha,\beta} \sum_k \|\mathbf{L}_k - P(\mathrm{l}_k(\alpha,\beta|\mathbf{R}))\|_2 + w_s\|\alpha\|_2 + w_e\|\beta\|_2, \quad (1)$$

where $\mathbf{s}, \mathbf{R} \in \mathrm{SO}(3), \mathbf{t} \in \mathbb{R}^2$ are rigid transformation parameters, α, β are 3DMM's shape and expression coefficients.

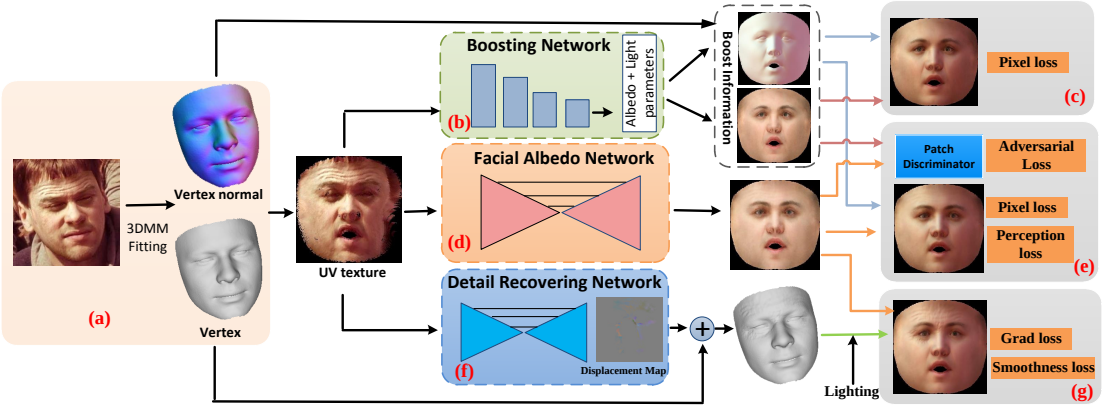

Fig. 2. Detailed overview of our processing pipline. The first step is to estimate the coarse shape from the input image, and then we acquire the texture map and vertex normals which are used for training and inference stages (a). The training process includes (b)-(g). Three networks are trained one-by-one by self-supervision learning. The B-Net is used to boost the training of the image-to-image FAN and DRN. To get albedos and details for rendering, only stages (a) (b) (d) (f) are needed.

The operator $P(\cdot)$ is the rigid transformation followed by the weak perspective projection. $\mathrm{l}_k(\alpha, \beta|R)$ means selecting corresponding vertices from shape generated by α, β. The indexes are decided by rotation $\mathbf{R}$, which considers the sliding landmarks when the input face is non-frontal [26]. To solve the least-squares problem, we first use the iterative strategy to optimize the pose and 3DMM coefficients. Finally, we solve all the parameters altogether using Levenberg-Marquadt algorithm. We extracts the corresponding texture map, denoted as $\mathbf{I_t}(\mathrm{UV})$ after fitting. The texture maps and the base shapes are used in the following sections.

3.2. Boosting Network

We aim to train our image-to-image albedo model FAN which has more powerful expression abilities than a PCA albedo model, but supervision labels are not easy to be acquired. Without target labels, it's difficult to remove the illumination and complete the invisible and occlusion regions from in-the-wild texture maps. Hence, we propose to train a boosting network (B-Net) in advance which provides the illumination information and a complete albedo map distribution. These will help to train our FAN. B-Net is also trained by self-supervision learning.

As shown in Fig. 2 (b), the B-Net estimates the illumination and albedo parameters. Our B-Net depends on an easily acquired PCA albedo model:

$$\mathbf{R_a} = \mathbf{m_a} + \mathbf{U_a}\mathbf{p_a}, \tag{2}$$

where $\mathbf{m_a}$ and $\mathbf{U_a}$ are the components of the model, $\mathbf{p_a} \in \mathbb{R}^{199}$ is the coefficients to create the albedo. For illumination, we assume Lambertian reflectance and use the second-order spherical harmonics (SH) basis functions to represent the global illumination. The rendered texture color of one channel at vertex v with its normal n and albedo r is:

$$\mathrm{I_{syn}}(\mathbf{n}, \mathbf{r}, \gamma) = \mathbf{r} \cdot \sum_{b=1}^{B^2} \gamma_b H_b(\mathbf{n}), \tag{3}$$

where $B = 3$ and γ is illumination coefficients. We use Resnet-18 as our backbone with small modifications. We set the expansion 2 then the feature dimension after pooling is 1024. The last pooling size is set to 8 to fit our 256×256 texture maps. The output feature then forward a linear model which predicts 226 (199+9*3) parameters including albedo and RGB SH coefficients. The cost function to train B-Net is:

$$E_{\text{B-Net}} = E_{\text{pixel}} + w_{\text{a}} E_{\text{a}}, \tag{4}$$

where $E_{\text{pixel}} = |\mathbf{I_t}(\mathrm{UV}) - \mathbf{I_{syn}}(\mathrm{UV})|$ is the l_1 loss between the synthetic and input textures. $E_{\text{a}} = \|\mathbf{p_a}\|^2$ is the regularization item.

3.3. Facial Albedo Network

We employ a UNet-like structure as our FAN. The FAN is an image-to-image regression network whose outputs have a direct relationship with inputs on each pixel. This is more suitable to estimate albedo map from texture map than the encoder-decoder structures. Following the idea of [15], our FAN makes slight modifications on UNet-8: (1) to let the model size of FAN small, we set "ngf" to 16; (2) to preserve the overall details of input image, we change the stride size from 2 to 1 and the kernel size from 4 to 3 in the first and last convolution layers. We add an extra skip connection block to make sure the spatial dimension of the deepest layer is 1×1; (3) to make our FAN has more abilities to infer the invisible regions, we add flip features after a convolution layer and then forward to the UNet structure. The flip features can naturally

help to infer the albedo from invisible texture regions rather than the symmetry hypothesis [12, 11] which can damage the inference on visible regions.

Without supervisions, we can only resort to the input texture maps to train our FAN. However, the in-the-wild textures which under the affects of diverse illumination, no-skin occlusions and invisible facial regions can't be directly used. So we need the boost information from B-Net including illumination and an important information — albedo distribution in which the albedos are intact which make FAN's self-supervision learning possible. The albedo distribution induces the FAN having the similar albedo distribution which make FAN be robust to no-skin occlusions and complete the invisible regions naturally. This can be realized by adversarial learning. The cost function to train FAN is:

$$E_{\text{FAN}} = \min_{G} \max_{D} E_{\text{GAN}(G,D)} + E_G, \tag{5}$$

where the generator G is FAN in this paper, the discriminator D is Patch-based [27] which helps to concentrate more on distinguishing local textures and

$$E_{\text{GAN}(G,D)} = \mathbb{E}_x[\log D(x)] + \mathbb{E}_z[\log(1 - D(G(z)))], \tag{6}$$

where z is texture map like $\mathbf{I_t}(\text{UV})$. The x in Equ. 6 is sampled from the albedo domain of B-Net rather than randomly sampled from the PCA model, thus makes x more like real albedos. The E_G is,

$$E_G = w_{\text{pixel}} E_{\text{pixel}} + w_{\text{percep}} E_{\text{percep}}, \tag{7}$$

where E_{pixel} is the same as it in Equ. 4 except that the albedo comes from FAN. E_{percep} is perception loss,

$$E_{\text{percep}} = \frac{1}{n} \sum_{j}^{n} \frac{\|\mathcal{F}_j(\mathbf{I_t}(\text{UV})) - \mathcal{F}_j(\mathbf{I_{syn}}(\text{UV}))\|_2}{H_j W_j C_j}, \tag{8}$$

where $\mathcal{F}$ is a pretrained network and $\mathcal{F}_j$ means the feature maps at layer j. We choose VGG-16 trained on ImageNet as our $\mathcal{F}$ [28]. The purpose of using E_{percep} is that we hope FAN has attentions on mid-level and high-level features rather than only the raw pixel features.

3.4. Detail Recovering Network

The transient high-frequency details such as wrinkles and crows feet should be part of the mesh shapes rather than on the albedos. Our DRN estimates a displacement UV map to represent details. These details change the vertical normals which impact the synthetic texture according to the Equ. 3. Our D-Net not directly estimates the displacement map, it predicts the moving rate $r_d(\text{UV})$ on the direction of the vertex normal. Thus the displacement is $\Delta\mathbf{d} = r_d\mathbf{n}$ and the final shape becomes $\mathbf{v}' = \mathbf{v} + \Delta\mathbf{d}$. To avoid the overlap of the influence of details and albedos, we train DRN based on FAN.

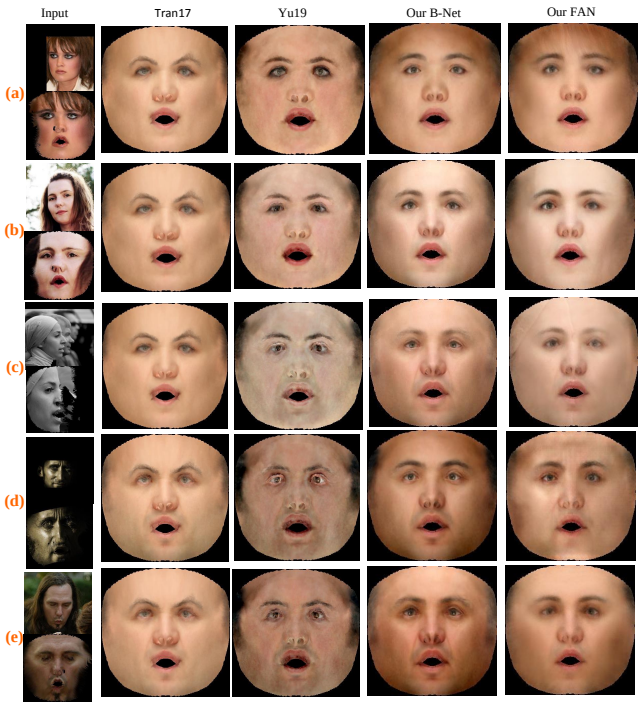

Fig. 3. Comparison with 3DMM-based methods. The second and third columns from works Tuan17 [29] and Yu19 [6].

The structure of DRN is the same as FAN except that the last layer has 1 channel.

To train DRN by self-supervision, the intuitive idea is using pixel loss E_{pixel}. However, we find that the $\mathbf{I_t}(\text{UV})$ and the $\mathbf{I_{syn}}(\text{UV})$ usually have large color gaps which hinders DRN's concentrations on details. So we decide to use pixel's gradients rather than its intensities as the loss function. Finally, the cost function to train DRN is:

$$E_{\text{DRN}} = w_{\text{grad}} E_{\text{grad}} + w_{\text{const}} E_{\text{const}} + w_{\text{smooth}} E_{\text{smooth}}, \tag{9}$$

where $E_{\text{grad}} = |\nabla\mathbf{I_t}(\text{UV}) - \nabla\mathbf{I_{syn}}(\text{UV})|$, $E_{\text{const}} = |\Delta\mathbf{d}|$ and $E_{\text{smooth}} = \sum_{v_i \in V} \sum_{v_j \in \text{Neib}(v_i)} \|\Delta d_i - \Delta d_j\|^2$. The V stands for all the vertices and the $\text{Neib}(v_i)$ means the set of 1-ring neighborhood of vertex i.

4. EXPERIMENTS

4.1. Experimental Setup

Dataset. We utilize CelebA-HQ dataset [30] and CoarseDataset [31] to train our B-Net and FAN. We train our DRN using FineDataset [31]. The popular AFLW2000 dataset [25] and several in-the-wild face images are used for evaluation. For CoarseDataset and FineDataset, we select about 30K images which are nearly the size of celebA-HQ from each dataset.

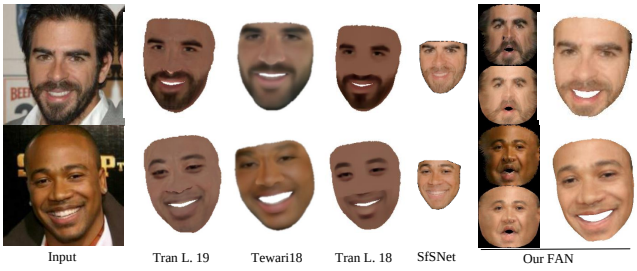

Fig. 4. Comparison with nonlinear albedo inference approaches. From left to right are results from Tran L.19 [11], Tewari18 [10], Tran L.18 [12], SfSNet [18] and our method.

Settings. We train our networks using Adam optimizer with learning rate 2e-5. Our empirical setting of the loss weights is $w_a = 1, w_{\text{pixel}} = 0.005, w_{\text{percep}} = 0.3, w_{\text{grad}} = 0.5, w_{\text{const}} = 1$ and $w_{\text{smooth}} = 20$.

4.2. Facial Albedo Generation

Albedo completion analysis. The faces in in-the-wild images may be occluded by different types of objects, under large pose and exposed to different illumination. Given such an unconstrained image, the goal of our FAN is to produce a complete and high-realistic albedo, being robust to the unconstrained environment. To examine the effectiveness of the proposed approach, we compare the facial albedos generated by our FAN, our B-Net, Tuan17 [29] and Yu19 [6]. The results of Tuan17 [29] and Yu19 [6] are generated from their open-source program. We use non-linear iterative nearest point registration to get the same mesh topology, thus we have the same UV maps. The comparison results are illustrated in Fig. 3. We can see that Tuan17 tends to produce mean albedo without clear distinction, despite large variation of the input in-the-wild images. Yu19 is sensitive to the illumination variations which results in albedos with strange effects, seen in Fig. 3 (c-e). Note that Yu19 is trained with skin region detection, which makes it less sensitive to occlusions. While by combining low-frequency and non-linear information, our FAN can merge the occlusions more naturally in the albedos, hence, can obtain better results. However, B-Net is easily affected by the occlusions, seen in Fig. 3 (a). The comparison shows that our FAN is robust to occlusions, large poses, and illumination, and can generate more appropriate albedo even without the help of skin detection.

High-fidelity albedo. The non-linear property of our FAN permits to produce high-fidelity albedos, so we compare it with some state-of-the-art non-linear approaches. Our FAN in this experiment is retrained to generate albedos more suitable for nearly positive faces with fewer occlusions. We just set $w_{\text{pixel}} = 0.008, w_{\text{percep}} = 0.7$. In Fig. 4 we can see that our method generates lively albedos. The results of Tewari18 [10] and Tran L. 18 [12] are blurry. Tran L. 19 [11] enhances the details of eyes and mouth, but the results seem artificial. The SfSNet [18] only estimates the visible face regions on the input image instead of the albedo map and generates low-resolution albedo, which is too coarse for high-quality albedo map. In contrast, our UV map has a resolution of 256×256 which is higher than all the above methods. The experiments demonstrate that our FAN can effectively take the input texture into account, and output the albedo that resembles the input face. Meanwhile, our FAN can complete the invisible regions and merge the occlusions naturally into albedos.

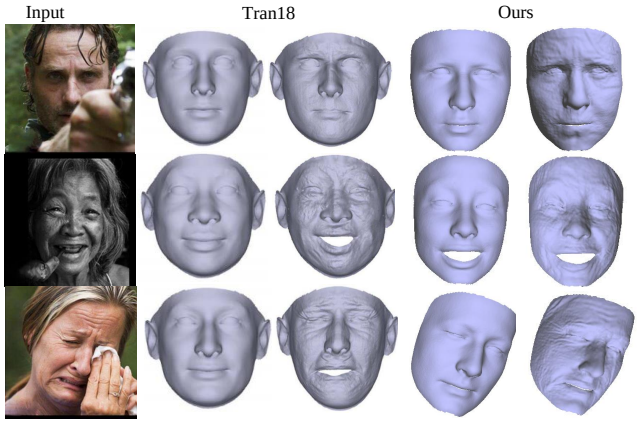

Fig. 5. Comparison with supervised detail estimation approach Tran18 [20].

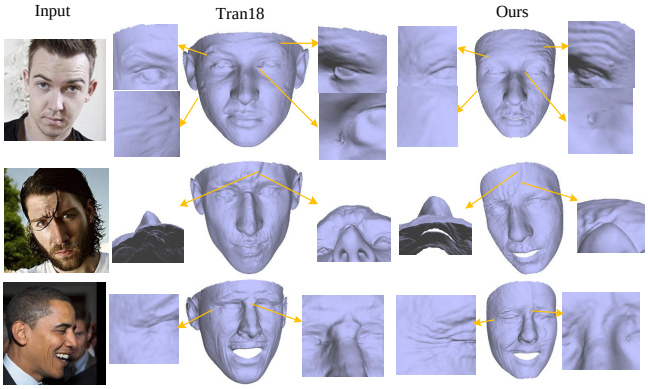

Fig. 6. Detailed Comparison with Tran18 [20].

4.3. Geometric Details Recovering

We evaluate our DRN on the face images with remarkable wrinkles to examine its effectiveness. In Fig. 5, we compare our estimation with Tran18 [20]. From the results, we see that our self-supervised method generates wrinkle details on the base meshes making it more realistic. The results are comparable to Tran18 [20] whose details are estimated by an image-to-image network trained in a supervised manner. In Fig. 6, we examine the reconstructed results at finer scale. It is worth

noting that our method has no operator about "seeing through occlusions" proposed by Tran18 [20], so we just focus on the visible regions. From Fig. 6, we see that our method have advantages in estimating the details in prominent areas (the top right patches of the first row and the left patches of the third row) and is less sensitive to shadings (the patches of the second row) which can cause wrong mesh shapes. This is because our method is trained through the pixel gradient differences which concentrate more on color variations rather than the color itself. Moreover, in contrast to Tran18, our DRN does not affect the original meshes as shown on the bottom left patches of the first row and the right patches of the third row. The bottom right patch of Tran18's mesh on the first row even has holes. This is because our DRN only makes the vertices move along the original vertex normal, which avoids the interleaving of vertices as much as possible.

5. CONCLUSION

In this paper, we propose a low-frequency guided self-supervised learning method for high-fidelity facial albedo generation and geometric details estimation. We exploit both the image-to-image model and the low-frequency information from the linear facial model to train our facial albedo network without annotation. Besides, the detail recovering network estimates facial details such as wrinkles that are not presented in albedo. By merging facial albedo and detailed shape, we can effectively reconstruct 3D faces from in-the-wild images.

6. REFERENCES

[1] Michael Zollhöfer, Justus Thies, et al., "State of the art on monocular 3d face reconstruction, tracking, and applications," in *Computer Graphics Forum*, 2018.

[2] James Booth, Epameinondas Antonakos, et al., "3d face morphable models" in-the-wild"," in *CVPR*, 2017.

[3] Luan Tran and Xiaoming Liu, "Nonlinear 3d face morphable model," in *CVPR*, 2018.

[4] Volker Blanz, Thomas Vetter, et al., "A morphable model for the synthesis of 3d faces.," in *Siggraph*, 1999.

[5] Ayush Tewari, Michael Zollhofer, et al., "Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction," in *ICCV*, 2017.

[6] Yu Deng, Jiaolong Yang, et al., "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in *CVPRW*, 2019.

[7] Matan Sela, Elad Richardson, et al., "Unrestricted facial geometry reconstruction using image-to-image translation," in *ICCV*, 2017.

[8] Elad Richardson, Matan Sela, et al., "Learning detailed face reconstruction from a single image," in *CVPR*, 2017.

[9] Hyeongwoo Kim, Michael Zollhöfer, et al., "Inversefacenet: Deep monocular inverse face rendering," in *CVPR*, 2018.

[10] Ayush Tewari, Michael Zollhöfer, et al., "Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz," in *CVPR*, 2018.

[11] Luan Tran, Feng Liu, and Xiaoming Liu, "Towards high-fidelity nonlinear 3d face morphable model," in *CVPR*, 2019.

[12] Luan Tran and Xiaoming Liu, "On learning 3d face morphable model from in-the-wild images," *TPAMI*, 2019.

[13] Baris Gecer, Stylianos Ploumpis, et al., "Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction," in *CVPR*, 2019.

[14] Shunsuke Saito, Lingyu Wei, et al., "Photorealistic facial texture inference using deep neural networks," in *CVPR*, 2017.

[15] Shuco Yamaguchi, Shunsuke Saito, et al., "High-fidelity facial reflectance and geometry inference from an unconstrained image," *TOG*, 2018.

[16] James Booth, Anastasios Roussos, et al., "A 3d morphable model learnt from 10,000 faces," in *CVPR*, 2016.

[17] Elad Richardson, Matan Sela, and Ron Kimmel, "3d face reconstruction by learning from synthetic data," in *3DV*, 2016.

[18] Soumyadip Sengupta, Angjoo Kanazawa, et al., "Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'," in *CVPR*, 2018.

[19] Pablo Garrido, Michael Zollhöfer, et al., "Reconstruction of personalized 3d face rigs from monocular video," *TOG*, 2016.

[20] Anh Tuan Tran, Tal Hassner, et al., "Extreme 3d face reconstruction: Seeing through occlusions.," in *CVPR*, 2018.

[21] Chen Cao, Derek Bradley, et al., "Real-time high-fidelity facial performance capture," *ToG*, 2015.

[22] Aaron S Jackson, Adrian Bulat, et al., "Large pose 3d face reconstruction from a single image via direct volumetric cnn regression," in *ICCV*, 2017.

[23] Yao Feng, Fan Wu, et al., "Joint 3d face reconstruction and dense alignment with position map regression network," in *ECCV*, 2018.

[24] Yajing Chen, Fanzi Wu, et al., "Self-supervised learning of detailed 3d face reconstruction," 2019.

[25] Xiangyu Zhu, Zhen Lei, et al., "Face alignment across large poses: A 3d solution," in *CVPR*, 2016.

[26] Luo Jiang, Juyong Zhang, et al., "3d face reconstruction with geometry details from a single image," *IEEE Transactions on Image Processing*, 2018.

[27] Phillip Isola, Jun-Yan Zhu, et al., "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[28] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016.

[29] Anh Tuan Tran, Tal Hassner, et al., "Regressing robust and discriminative 3d morphable models with a very deep neural network," in *CVPR*, 2017.

[30] Tero Karras, Timo Aila, et al., "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[31] Yudong Guo, Jianfei Cai, et al., "Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images," *IPAMI*, 2018.