Weakly-Supervised Photo-realistic Texture Generation for 3D Face Reconstruction

Xiangnan Yin, Di Huang, Zehua Fu, Yunhong Wang, Liming Chen

Weakly-Supervised Photo-realistic Texture Generation for 3D Face Reconstruction

Xiangnan YIN,[1] **Di HUANG,**[2] **Zehua FU,**[2] **Yunhong WANG,** [2] **Liming CHEN**[1]

[1] Ecole Centrale de Lyon, France
[2] Beihang University, China
yin.xiangnan@ec-lyon.fr, dhuang@buaa.edu.cn, zehua_fu@163.com, yhwang@buaa.edu.cn, liming.chen@ec-lyon.fr

Abstract

Although much progress has been made recently in 3D face reconstruction, most previous work has been devoted to predicting accurate and fine-grained 3D shapes. In contrast, relatively little work has focused on generating high-fidelity face textures. Compared with the prosperity of photo-realistic 2D face image generation, high-fidelity 3D face texture generation has yet to be studied. In this paper, we proposed a novel UV map generation model that predicts the UV map from a single face image. The model consists of a UV sampler and a UV generator. By selectively sampling the input face image's pixels and adjusting their relative locations, the UV sampler generates an incomplete UV map that could faithfully reconstruct the original face. Missing textures in the incomplete UV map are further full-filled by the UV generator. The training is based on pseudo ground truth blended by the 3DMM texture and the input face texture, thus weakly supervised. To deal with the artifacts in the imperfect pseudo UV map, multiple UV map and face image discriminators are leveraged.

Figure 1: Results of the proposed method. The left column shows the input images. Images on the right are synthesized using the predicted UV-map.

Introduction

3D face reconstruction is an important yet challenging domain in computer vision, aiming to faithfully restore the shape and texture of a face from one or more face images. It has a wide range of applications, such as face recognition, face editing, face animation, and other artistic and entertainment fields. Recently, there has been a surge of interest in single-image based 3D face reconstruction (Deng et al. 2019b; Guo et al. 2020; Richardson et al. 2017; Feng et al. 2018; Tran and Liu 2018). While most previous work has been devoted to predicting more accurate and detailed 3D shapes, not much work has focused on generating photo-realistic face textures. However, studies (Masi et al. 2019; Hassner et al. 2015) have shown that the texture plays a more significant role than that of the shape in face recognition tasks. Thus we can never ignore the importance of the texture in 3D face reconstruction.

Existing 3D face texture generation methods can be broadly classified into three categories: texture model-based, image generation-based, and GAN optimization-based.

Texture model-based Since the 3D Morphable Model (3DMM) (Blanz and Vetter 1999) was proposed, it has been widely used in 3D face reconstruction. The model is a vector basis of the shape and texture learned from a set of 3D face scans. Earlier approaches regress the 3DMM parameters by solving a non-linear optimization problem (Richardson, Sela, and Kimmel 2016; Booth et al. 2017), which is often slow and costly. With the development of Convolutional Neural Networks, recent studies tend to predict the parameters using learning-based methods (Richardson et al. 2017; Guo et al. 2018; Deng et al. 2019b). However, the 3DMM is constructed by a small number of face scans under well-controlled conditions, limiting its diversity to identity, race, age, gender, etc. Besides, due to the linear and low-dimensional nature of the model, it can hardly capture high-frequency details, resulting in blurred textures that are far from satisfactory.

Image generation-based Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) provide a powerful tool for generating photorealistic images. Since its appearance, numerous image generation methods with stunning results have been proposed. Thanks to various large databases and the highly structured geometry of the human face, 2D face image generation is one of the most prosperous areas (Huang et al. 2018; Pumarola et al. 2018; Choi et al. 2018;

Karras, Laine, and Aila 2019; Karras et al. 2017). Influenced by this trend, some recent 3D face reconstruction methods have also leveraged adversarial training to improve the texture quality (Tran and Liu 2018; Deng et al. 2018; Lee and Lee 2020). However, such kind of approaches are highly dependent on large 3D face databases. (Tran and Liu 2018) is trained on a synthesized 3D face database (Zhu et al. 2016b), where originally self-occluded textures are obtained by simple interpolation of visible parts, resulting in imperfect generation. (Deng et al. 2018; Lee and Lee 2020) are trained on a large UV map dataset, which is not publicly available.

GAN optimization-based The traditional yet most powerful GANs are trained to synthesize images from noise vectors (Karras et al. 2017; Karras, Laine, and Aila 2019; Brock, Donahue, and Simonyan 2019). To leverage the power of a pre-trained GAN, a series of works are established on inverting the image back to a GAN's latent space using optimization-based approaches (Shen et al. 2020; Ma, Ayaz, and Karaman 2018a,b; Zhu et al. 2016a). Similar methods are used to generate the UV map of a face image (Gecer et al. 2019; Lee et al. 2020). First, they train a generator that converts noise vectors into UV maps. Then they directly optimize the latent code to minimize the reconstruction error between the input face image and the image rendered by the generated UV map. Instead of training a UV map generator, (Gecer, Deng, and Zafeiriou 2020) first rotates the input image in 3D and optimizes the latent code of the pre-trained StyleGAN to fill in the missing textures, then stitches textures of different view angles by alpha blending to form the final UV map. By far, the optimization-based methods can yield the most realistic face UV maps. Nevertheless, they are usually complex and time-consuming, *e.g.*, GANFIT (Gecer et al. 2019) takes 30 seconds to generate the UV map of an input face, while OSTeC (Gecer, Deng, and Zafeiriou 2020) takes up to 5 minutes.

Besides generating a global face texture, we note that a series of pure 2D image generation methods can also synthesize face images of different view angles (Tran, Yin, and Liu 2017; Zhou et al. 2020; Hu et al. 2018). However, the generation consistency is poor due to the absence of global consistency constraints and a priori knowledge of the 3D shape.

In summary, among the current texture generation methods for 3D face reconstruction, those based on texture models cannot yield high-fidelity results due to the model's simplicity; those based on image generation rely heavily on large training dataset; those based on optimization are time-consuming and require a high computational cost.

To this end, we propose a novel image-to-image translation model that converts the input face image into its corresponding UV map. The proposed method is image generation-based, therefore much faster than optimization-based methods. We use the pseudo UV map for training, bypassing the dependency on the real UV map database. Thanks to multiple partial UV discriminators, we can use cropped parts of incomplete UV maps (acquired using the data pre-processing method provided in (Deng et al. 2018)) for training to improve the generation quality. Our contributions are as follows:

- A novel image generation-based UV map prediction framework is proposed. The generated results are comparable to the optimization-based method but much faster.
- With the proposed UV sampler module, the visible face textures can be directly mapped to the UV space, forming an incomplete UV map. No 3D information (shape, occlusion) is required during the inference stage. Therefore, our model can be stitched seamlessly with any 3D shape reconstruction model.
- The training doesn't rely on the real UV map database, and the design of multiple discriminators can compensate well for the imperfect ground truth.
- The proposed method outperforms the state-of-the-art methods, both qualitatively and quantitatively.

Related Work

3D shape reconstruction From earlier optimization-based methods to CNN prediction-based methods, acquiring accurate 3D face shape becomes easier and faster, bringing powerful tools and significant opportunities for face-related tasks. Our training process relies on 3D shape reconstruction of a given face, where numerous 3D shape fitting methods are applicable. In this paper, we adopt an off-the-shelf model (Deng et al. 2019b) as our shape re-constructor, which is the current SOTA 3DMM-based method. The model will predict its corresponding pose and 3DMM shape/texture parameters with a single face image as input.

UV map generation There exist mainly two texture representation methods for 3D models, vertex-based and UV map-based. The vertex-based representation is very intuitive, where each vertex has a color, and the interpolation of those colors generates the texture of the 3D surface. However, such representation flattens the texture into a linear vector, destroys the spatial relationship of texture patches, thus prevents it from leveraging powerful CNN-based methods. The UV map-based representation unwraps the 3D texture into a 2D space. Briefly, each 3D vertex's color is mapped to its corresponding location of a 2D image, and adjacent vertices are mapped to adjacent regions so that the positional relationships between vertices are well preserved. (Deng et al. 2018) first sample the color of visible 3D vertices from the input face image, then map them to UV space to get the incomplete UV map, in which the generative model will further complete the missing parts. However, their method is highly dependent on the precise 3D shape and ground truth UV maps. In contrast, our method does not need the UV map data for training or 3D shape for inference. (Tran and Liu 2018) propose a non-linear 3DMM, where the predicted texture takes the UV map-based representation. Nevertheless, their UV map generator's input is a low-dimensional encoding of the input image, resulting in an loss of detail of the predicted UV map. In addition, their model is trained on linear 3DMM synthesized images (Zhu et al. 2016b), where artifacts caused by self-occlusion appear frequently. Unlike (Tran and Liu 2018), our model is trained on real face images, and the coding keeps a large dimension across the forward path, making the generated UV

map photorealistic.

Differentiable renderer To obtain the gradient of the loss function and thus train the network, a differentiable renderer is widely used in 3D face-related algorithms (Richardson et al. 2017; Guo et al. 2018; Tran and Liu 2018; Deng et al. 2019b). Briefly, a renderer is composed of a rasterizer and a shader. The rasterizer applies depth-buffering to select the mesh triangles corresponding to each pixel, and the shader computes the pixel colors as follows:

$$\bar{c} = w_0 c_0 + w_1 c_1 + w_2 c_2 \tag{1}$$

where c_i is the color of the i^{th} vertex of the mesh triangle the pixel resides in, w_i is the barycentric coordinate of the pixel in the triangle. During backward propagation, the gradients are passed from each pixel to the vertices:

$$\frac{\mathrm{d}L}{\mathrm{d}c_i} = \frac{\mathrm{d}L}{\mathrm{d}\bar{c}} \frac{\mathrm{d}\bar{c}}{\mathrm{d}c_i} = \frac{\mathrm{d}L}{\mathrm{d}\bar{c}} w_i \tag{2}$$

where L is the loss function. Since c_i is sampled from the output of the texture generator, *i.e.*, the UV map, the gradients could be further backpropagated. In our project, we adopt the off-the-shelf differentiable renderer of PyTorch3D (Ravi et al. 2020).

Pixel attention sampling To get the UV map of visible parts, UV-GAN (Deng et al. 2018) first fits a 3DMM to the input image, then use the vertices' projected 2D coordinates to sample their corresponding colors, and the incomplete UV map is further generated. However, their method relies on accurate 3D shape fitting and facial landmark detection. Furthermore, such a method does not have a mechanism to deal with face occlusions(hands, hair, eyeglasses, etc.). Inspired by (Yin et al. 2020), we apply a pixel attention sampling (PAS) module to sample the incomplete UV map from the input image directly. Thanks to this module, the inference process is free from 3D shape or facial landmarks. Besides, different from (Yin et al. 2020), where input images require landmark-based pre-alignment due to the arbitrary target poses. The target output, *i.e.*, the UV map, is highly structured, so neither spatial transformation to the input image nor the target pose condition is demanded.

Proposed Method

The goal of our method is to predict the face UV map from a single face image. As illustrated in Figure 3, the proposed model consists of two parts: a UV attention sampling module (UV sampler) and a UV map inpainting module (UV generator). During the inference process, the UV sampler will sample the pixels from the input image to generate an incomplete UV map, and then the UV generator will further complete the semi-finished UV map. We describe the details of each component as follows.

UV Attention Sampling

The UV map is a two-dimensional representation of the global texture of a 3D object. Due to self-occlusion, it is an ill-posed problem to get the UV map from a single image. This section studies how to generate an incomplete UV map that contains only visible textures of the input face image. As

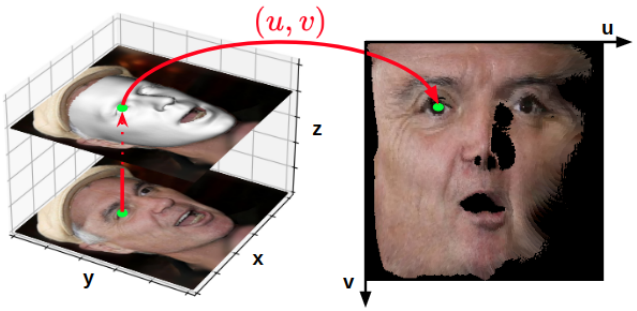

Figure 2: The traditional method for incomplete UV map generation. Which is used for generating the target output of the UV sampler.

a comparison, we recall the traditional method, which consists of four steps: (1) Get the 3D face shape based on the input image. (2) Determine the visible vertices using depth-buffer-based methods. (3) Project these visible vertices onto the image plane and index their colors according to their coordinates. (4) Render the UV map with the colors and the pre-defined UV-coordinates corresponding to each visible vertex. Figure 3 illustrates the above steps. Obviously, such a method is tedious and relies on an accurate 3D shape fitting. Since the UV map contains *only* the texture information of a 3D surface, is it really necessary to fit the exact 3D shape before getting the UV map? We do not think so. In fact, the only purpose of the 3D shape is to establish a one-to-one relationship between the pixel in the 2D face image and the pixel in the UV map, so why not learn such a mapping relationship in a data-driven manner? To achieve such a goal, we designed the UV sampler, a CNN-based model that maps the face image's pixels directly to the UV map.

The model has three parts, *i.e.*, the feature extractor, the segmentation head, and the sampler head. Similar to most generative models, the feature extractor is composed of stacked residual blocks (He et al. 2016). Spectral normalization (Miyato et al. 2018) is applied to each convolution layer to stabilize the training. With this module, 2D feature maps of different scales and a 512-dimensional vector are extracted from the input image. The 2D feature maps are fed into the FPN structured (Lin et al. 2017) segmentation head and output an attention mask m. Besides, the 1D feature vector is fed into the sampler head, a stack of fully connected layers interspersed with ReLU activations. The sampler head's output is reshaped as $S_{att} \in \mathbb{R}^{B\times 256\times 256\times 2}$, which is the attention sampling map, where B is the batch size, 256 is the height/width of the UV map, and the last two channels hold the normalized abscissa and ordinate of the pixel in the input image to sample. Based on S_{att} and m, differentiable sampling (Jaderberg et al. 2015) is applied to the masked input image I, and an incomplete UV map $\widehat{UV}_{spl}$ is finally obtained.

To train the model, we use the above-mentioned traditional method to generate the ground truth (incomplete) UV map, UV_{gt}. A minor improvement is that we multiply the input image by the eroded mask of the face before sam-

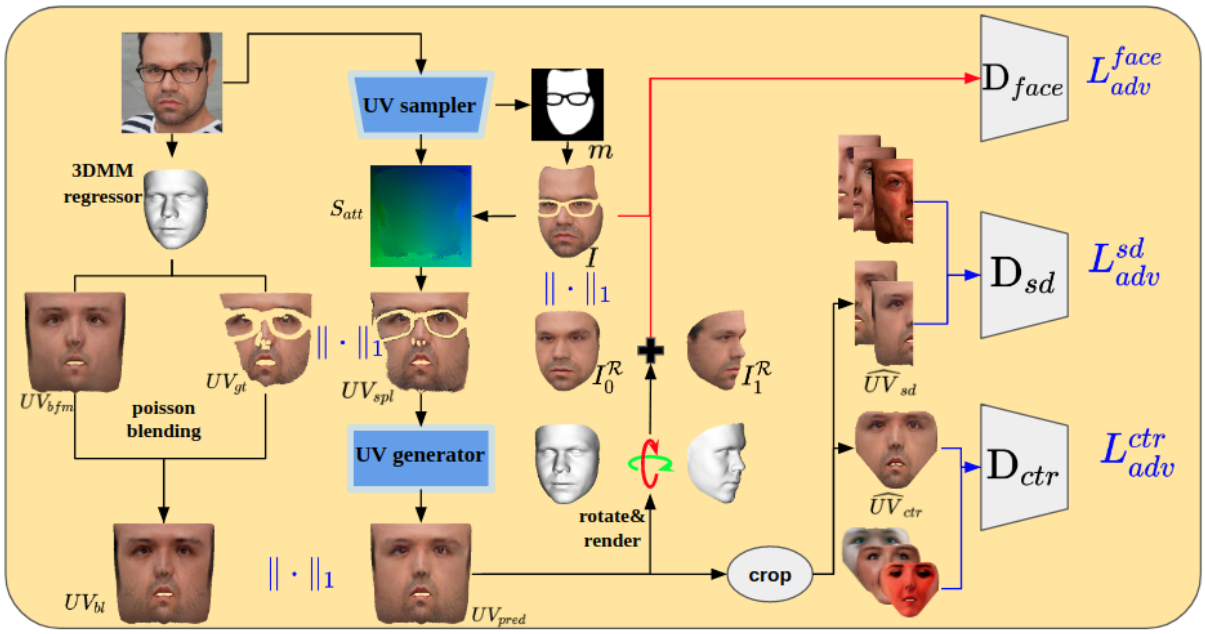

Figure 3: Overview of our approach. (1) Given an input face image, the UV sampler predicts its face mask m and sampling map S_{att}, based on which samples an incomplete UV map UV_{spl} (2) The UV generator will further complete the sampled UV map and output the UV_{pred}. (3) With an off-the-shelf 3DMM regressor, we predict the shape and texture of the input face image, which is used for getting the ground truth of the UV_{spl}: UV_{gt} and the pseudo ground truth of the UV_{pred}: UV_{bl}. (4) The predicted UV map is used to render face images of different poses: $I_0^{\mathcal{R}}$ and $I_1^{\mathcal{R}}$, which are further fed into a face discriminator. (5) UV_{pred} is cropped to the side part $\widehat{UV}_{sd}$ and center part $\widehat{UV}_{ctr}$, fed into their corresponding discriminators.

pling, which avoids incorrectly sampling the occlusion and the background (due to the inaccurate 3D shape) into the UV map. We erode the mask's edge to ensure that the region inside of which must be the face, the generation of UV_{gt} only takes into account the vertices that fall inside the mask. Although this would result in a loss of texture near the edge, it is worth sacrificing the unimportant edges to ensure the accuracy of UV_{gt}.

The training is guided by the following loss function:

$$\mathcal{L}_{spl} = \|\widehat{UV}_{spl} - UV_{gt}\|_1 + \|S_{att} - S_{gt}\|_1 + L_{seg}(m, m_{gt}) + \lambda TV(\widehat{UV}_{spl}) \quad (3)$$

where S_{att} and m are the outputs of the UV sampler, $\widehat{UV}_{spl}$ is the sampled UV map based on them. S_{gt} is the ground truth sampling map, which is obtained by mapping the normalized x,y coordinates of the visible 3D vertices into the UV space, *i.e.*, UV position map (Feng et al. 2018). $L_{seg}(m, m_{gt})$ is the binary cross-entropy loss of the predicted face mask.

$$L_{seg} = -[m_{gt}\log m + (1 - m_{gt})\log(1 - m)] \quad (4)$$

$TV(\widehat{UV}_{spl})$ is the total variation loss (Mahendran and Vedaldi 2015) of the predicted UV map, which is powerful in smoothing the noises of the generated UV map.

$$TV(\widehat{UV}_{spl}) = \sum_{x,y,c=1}^{W-1,H,C} \left|\widehat{UV}_{spl}(x+1,y,c) - \widehat{UV}_{spl}(x,y,c)\right|^2 + \sum_{x,y,c=1}^{W,H-1,C} \left|\widehat{UV}_{spl}(x,y+1,c) - \widehat{UV}_{spl}(x,y,c)\right|^2 \quad (5)$$

Thanks to the UV sampler, an incomplete UV map could be sampled directly from the input image, bypassing a series of complex and expensive steps of traditional methods, including 3D shape fitting, visible vertices determination, UV map rendering, etc.

UV Map Inpainting

With the UV sampler described above, we can sample an incomplete UV map from a face image. The next task is to fill the missing parts with textures consistent with the sampled parts. This is an image inpainting problem, which has been extensively studied. However, most image inpainting methods are trained on paired images, meaning the ground truth image is uniquely determined. In contrast, in our case, the ground truth is not available. This section studies how to train a UV map inpainting model without the supervision of the ground truth. Briefly, our approach is to generate a pseudo ground truth UV map to assist the training. Then,

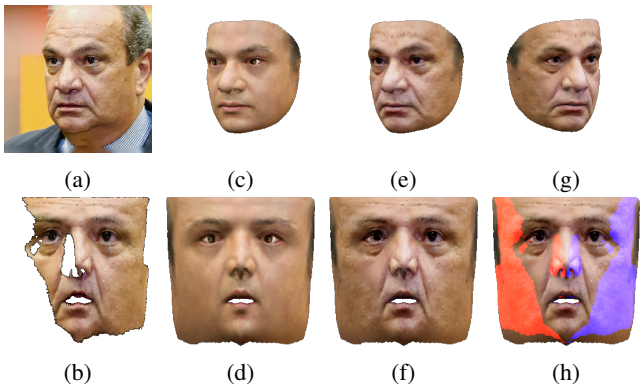

Figure 4: **(a)** The input image. **(b)** UV_{gt}. **(c)** The face reconstructed from UV_{bfm}. **(d)** UV_{bfm}. **(e)** The face reconstructed from UV_{bl}. **(f)** UV_{bl}. **(g)** The face in (e) under different view angle. **(h)** The texture marked in blue is used to fill the missing texture in its symmetric area (marked in red).

we work with multiple discriminators to make the generated images as photorealistic as possible.

Pseudo UV Map Generation Generating the pseudo UV map consists of three steps: 1) incomplete ground truth UV map generation, 2) 3DMM texture fitting, 3) seamless image blending. The first step has been described in detail in the previous section. For the second step, we use directly the BFM (Paysan et al. 2009) texture parameter predicted by (Deng et al. 2019b). The UV map representation of the reconstructed BFM texture is denoted as UV_{bfm}. Obviously, due to the linear, low-dimensional nature of the BFM model, UV_{bfm} is far from reality, as can be seen in Figure 4d. Therefore, we move to the third step: seamless image blending.

Seamless image blending With Poisson image editing (Pérez, Gangnet, and Blake 2003), we can seamlessly blend the results of the first two steps together. We also leverage the texture of the visible region to fill its missing symmetric region. That is, we do two times Poisson Blending, the first time blends the UV_{gt} to the UV_{bfm}, the second time blends the flipped UV_{gt} to its symmetric missing parts, as illustrated in Figure 4h. The final blending result is denoted as UV_{bl}, as in Figure 4f, both Figure 4e and Figure 4g are generated from it, which is far more photorealistic than the BFM reconstruction result in Figure 4c.

Multiple discriminators The training of the UV generator follows an adversarial paradigm; therefore, a large amount of data from the target domain is essential. However, the pseudo UV map U_{bl} generated above is not very reliable. Its quality depends on the accuracy of UV_{bfm}, the texture area of UV_{gt}, and the accuracy of the 3D shape. We only use the pseudo UV map to calculate the reconstruction loss, which is a rough guide to the generator's output. Although the complete UV map data is not available, we might as well collect a bunch of partial UV maps using the traditional method, *i.e.*, for UV maps generated from frontal face images, the central region, denoted as UV_{ctr}, is accurate, and for UV maps generated from profile face images, the visible half side, UV_{sd} is precise. Note that the partial UV maps collected in this way are not paired with UV_{pred}, so they are only used for adversarial loss, thus indirectly force the UV_{pred} lying in the real domain.

We design two partial UV map discriminators, one for the half side, the other for the center region. Together with the *masked face* discriminator, the system has three discriminators in total, as shown in Figure 3. The training is guided by the following losses.

Adversarial loss Given an output of the UV sampler, UV_{spl}, the generator will predict a global UV map, UV_{pred}. Three UV patches can be cropped from UV_{pred}, namely $\widehat{UV}_{ctr}$, $\widehat{UV}_{left}$, $\widehat{UV}_{right}$. Due to UV map's symmetry, the latter two can be put together and denoted as $\widehat{UV}_{sd}$. With the UV_{pred} and the 3D shape/pose parameters predicted by the model of (Deng et al. 2019b), a reconstructed face image $I_0^{\mathcal{R}}$ could be rendered. By changing the pose parameter, we can get a face image in a different view angle, denoted as $I_1^{\mathcal{R}}$. So far, we have three types of fake data: $\widehat{UV}_{sd}$, $\widehat{UV}_{ctr}$, and $I_{0,1}^{\mathcal{R}}$, each of which corresponds to real data represented as UV_{sd}, UV_{ctr}, and I^m, where I^m is the input face image with occlusions/background masked.

The adversarial loss is thus formulated as:

$$\mathcal{L}_{adv} = \mathbb{E}_x[\log D(x)] + \mathbb{E}_{\hat{x}}[\log(1 - D(\hat{x}))] \tag{6}$$

where

$$\begin{aligned}(x, \hat{x}, D) \in \{&(UV_{ctr}, \widehat{UV}_{ctr}, D_{ctr}),\\ &(UV_{sd}, \widehat{UV}_{sd}, D_{sd})\\ &(\{I^m, I_{0,1}^{\mathcal{R}}\} \odot m_{gt}, D_{face})\}\end{aligned}$$

Reconstruction loss The reconstruction loss consists of two terms, the UV reconstruction loss and the face reconstruction loss.

$$L_{rec} = \|UV_{pred} - UV_{bl}\|_1 + \|I_0^{\mathcal{R}} \odot m_{gt} - I^m\|_1 \tag{7}$$

Symmetry loss Since the UV map of the face is left-right symmetrical, we design the symmetry loss to help the model learn this property.

$$L_{sym} = \|UV_{pred} - FlipLR(UV_{pred})\|_1 \tag{8}$$

Identity loss Since the pose is arbitrary, the ground truth of $I_1^{\mathcal{R}}$ is not available. Thus we use the pre-trained FaceNet (Schroff, Kalenichenko, and Philbin 2015) to extract the identity feature of $I_1^{\mathcal{R}}$ and I^m, and minimize their L_1 distance.

$$L_{id} = \|\mathcal{F}(I_1^{\mathcal{R}}) - \mathcal{F}(I^m)\|_1 \tag{9}$$

TV loss TV loss of Equation 5 is also applied to UV_{pred}. The total loss function is as follows:

$$L = L_{rec} + \lambda_1 L_{adv} + \lambda_2 L_{sym} + \lambda_3 L_{id} + \lambda_4 TV \tag{10}$$

Experiments

The proposed method can faithfully convert the input face image to its corresponding UV map. To demonstrate the conversion ability, we qualitatively compare the 3D reconstruction results with the current state-of-the-art methods, both 2D-based and 3D-based. A quantitative evaluation is also presented.

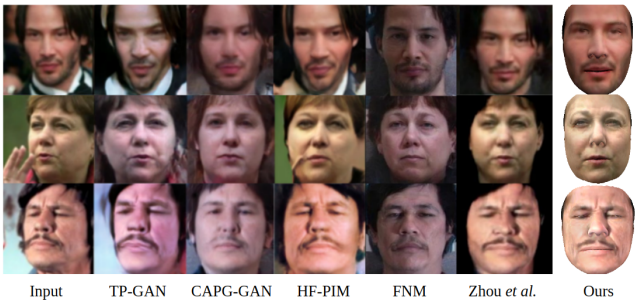

Figure 5: Frontalization results comparing with 2D-based face pose editing methods. Zoom-in for a better view.

Implementation details

Our training is based on two datasets:, CelebA-HQ (Karras et al. 2017), and FFHQ (Karras, Laine, and Aila 2019). Face images are pre-aligned with landmarks detected by (Bulat and Tzimiropoulos 2017). The input image size is 256×256, and the predicted UV map is the same size as the input. As for the ground-truth face mask, we first train a stand-alone face segmentation model, using the attribute mask of the CelebA-HQ and our manually labeled occlusions (eyeglasses, hands, etc.). Then we use this model to detect the face masks of the training data, and use them as the ground truth. We set the learning rate to $1e^{-4}$ and use Adam(Kingma and Ba 2014) optimizer with betas of [0.5, 0.999], the batch size is set to 6. We first pre-train the UV sampler until it outputs an incomplete UV map that can perfectly reconstruct the input image, which takes about 100K steps. Then we train the UV generator for 150K steps with the UV sampler's weights fixed. The training of the whole model takes about 100 hours on two Titan X Pascal graphics cards. Since most of the face images in the training sets are frontal, making the model not robust to the large view angles, to solve this problem, one trick we adopt is to rotate and render the input faces with their corresponding shapes and pseudo UV maps, then train the model to reconstruct the original face images.

Qualitative results

We use the predicted UV maps to render 3D shapes. By changing the pose parameters, images of different view angles are generated. For the qualitative evaluation, as a usual convention, we take the same inputs as others and paste the generated results after them. Figure 5 compares our frontalization results with 2D-based face pose editing methods, including TP-GAN (Huang et al. 2017), CAPG-GAN (Hu et al. 2018), HF-PIM (Cao et al. 2018), FNM (Qian, Deng, and Hu 2019) and Zhou*el al.*(Zhou et al. 2020). As shown in Figure 5, TP-GAN doesn't convert the pose well, and the third face image it generates is obviously left-skewed. Furthermore, the images generated by TP-GAN, CAPG-GAN, and FNM have large color deviations with the input images due to the influence of Multi-PIE (Gross et al. 2010) data in the training set. Besides our method, only HF-PIM and Zhou *et al.* maintain a consistent texture style with the input image. However, due to the lack of a priori knowledge of the 3D shape, HF-PIM cannot preserve the face shape well while editing the face pose. In addition, the second image generated by it preserves the finger on the mouth corner, showing that it cannot handle the face occlusions well. Our method achieves similar performance to the current state-of-the-art, Zhou *et al.*, we both use an off-the-shelf shape regressor. However, their method is based on face image generation, which means that we need to re-infer the missing texture each time we change the view angle. Another limitation of face image generation-based method is that the training settings greatly limit their pose editing freedom. The method of Zhou *et al.* cannot well generate face images of a large yaw angles; TP-GAN, FNM, and HF-PIM can only generate face images in frontal view.

A further qualitative comparison of our method and two representative 3D reconstruction-based methods are demonstrated in Figure 6. The method proposed by Deng *et al.* (Deng et al. 2019b) is based on 3DMM parameter regression, and GANFIT (Gecer et al. 2019) is based on latent-code optimization of a pre-trained GAN model. As can be seen, our results are more visually pleasant: large amounts of details are well preserved, including freckles, wrinkles, and expressions. Due to the model's low-dimensional nature, it's difficult for 3DMM-based methods to restore the input image's details faithfully. As can be seen in the 3rd-row of Figure 6, freckles and wrinkles are not well reconstructed. The results of GANFIT do contain richer details, but the resulting textures' styles are very homogeneous and differ considerably from their corresponding input images. We believe this is due to the lack of diversity in their training data, as the face UV map datasets are not easily accessible.

Training Data	Method	ACC(%)	AUC(%)
CASIA(baseline)	Zhou et al.	98.77	99.90
CASIA+rot	Zhou et al.	98.95	**99.91**
CASIA(baseline)	UV-GAN	99.02	-
CASIA+augUV	UV-GAN	**99.22**	-
CASIA(baseline)	ours	98.75	99.88
CASIA+augUV	ours	98.98	99.90

Table 1: Comparison of the face augmentation ability with UV-GAN (Deng et al. 2018) and (Zhou et al. 2020)

Quantitative results

Data augmentation As in many previous works, we use our proposed method to synthesize face images for face data augmentation and evaluate the performance of the model trained on the augmented dataset to demonstrate the merits of our approach. The experiment is based on the CASIA (Yi et al. 2014) dataset. Due to the low resolution of the images in CASIA, we retrained a model with an input/output resolution of 128×128. Moreover, we remove the segmentation head in the UV sampler to increase the diversity of the augmented images. For each image with less than 30° yaw angle, we randomly increase its yaw angle from 15° to 60° and get a synthesized images. Our basic training settings are the same as (Zhou et al. 2020), with ResNet18 (He et al. 2016) for the backbone and ArcFace (Deng et al. 2019a)

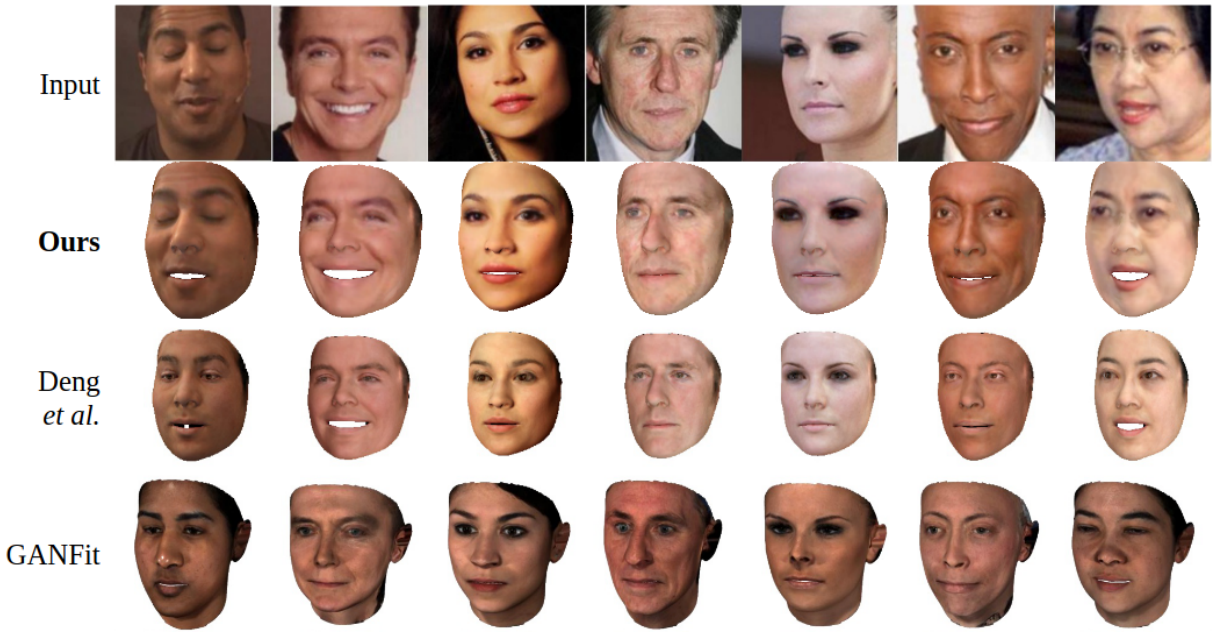

Figure 6: Qualitative comparison with other state-of-the art 3D reconstruction methods.

	Reconstruction		Recognition	
Method	L_1	SSIM	Recon	Front
Deng *et al.*	0.064	0.698	0.554	0.501
Zhou *et al.*	0.069	0.613	0.780	0.675
Ours	**0.021**	**0.913**	**0.862**	**0.684**

Table 2: Pixel-wise reconstruction and the identity-preserving ability on AFLW2000-3D, non-facial areas of all images are masked out for fair comparison.

for the loss function. Results are shown in Table 1. Since UV-GAN (Deng et al. 2018) uses ResNet27 as its backbone, which is deeper than ours and (Zhou et al. 2020), it is not surprising that it achieves the highest accuracy. Although we take the same settings as Zhou et al., we achieve a slightly lower baseline due to numerous differences in training details (learning rate, batch size, optimizer, etc.). However, by training on the augmented dataset, our model exceeds their accuracy, and our AUC closes the gap with them, demonstrating the efficiency of our data augmentation ability.

Face reconstruction We evaluate the proposed method in two aspects: the pixel-wise reconstruction ability and the identity-preserving ability. As most previous works are not open-sourced, we only compare with Deng *et al.* (Deng et al. 2019b) and Zhou *et al.* (Zhou et al. 2020), SOTA methods based on 3DMM and 2D face image generation, respectively. We conduct the experiments on the AFLW2000-3D (Zhu et al. 2016b), which contains 2000 face images with ground truth shape parameters.

For the reconstruction ability evaluation, we calculate the L1 loss and the structural similarity (Wang, Simoncelli, and Bovik 2003) of the reconstructed face images. As can be seen from Table 2, our method outperforms others in both these metrics.

As for the identity-preserving ability, the evaluation is conducted by features extracted by the pre-trained LightCNN-29 v2 (Wu et al. 2018) model. We calculate the cosine similarity of the features corresponding to the input images and the reconstructed/frontalized images. Results are shown in the two rightmost columns of Table 2. An interesting thing to notice is that, although Zhou *et al.* is inferior to the 3DMM-based model in terms of reconstruction loss, they are more capable of preserving the face identity. However, our proposed method achieves the best performance in both aspects.

Conclusion

This work proposes a novel 2-stage image-to-image translation model that can convert the input face image into its corresponding UV map. In the first stage, with the proposed UV sampler, pixels in the input face images are selectively sampled and adjusted to form an incomplete UV map, which contains all the visible textures of the face. With the help of this module, the inference stage no longer requires the intervention of 3D shapes. In the second stage, the incomplete UV map is further completed by a UV generator. The training is conducted on purely pseudo UV maps, thus weakly-supervised. With the help of two carefully designed partial UV discriminators, we can generate photo-realistic face textures without the supervision of the complete UV map. Qualitative and quantitative experiments validate the reconstruction ability and the identity-preserving ability of the proposed method.

References

Blanz, V.; and Vetter, T. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 187–194.

Booth, J.; Antonakos, E.; Ploumpis, S.; Trigeorgis, G.; Panagakis, Y.; and Zafeiriou, S. 2017. 3D face morphable models" In-The-Wild". In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 48–57.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*.

Bulat, A.; and Tzimiropoulos, G. 2017. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, 1021–1030.

Cao, J.; Hu, Y.; Zhang, H.; He, R.; and Sun, Z. 2018. Learning a high fidelity pose invariant model for high-resolution face frontalization. *arXiv preprint arXiv:1806.08472*.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8789–8797.

Deng, J.; Cheng, S.; Xue, N.; Zhou, Y.; and Zafeiriou, S. 2018. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7093–7102.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019a. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Deng, Y.; Yang, J.; Xu, S.; Chen, D.; Jia, Y.; and Tong, X. 2019b. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Feng, Y.; Wu, F.; Shao, X.; Wang, Y.; and Zhou, X. 2018. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 534–551.

Gecer, B.; Deng, J.; and Zafeiriou, S. 2020. OSTeC: One-Shot Texture Completion. *arXiv preprint arXiv:2012.15370*.

Gecer, B.; Ploumpis, S.; Kotsia, I.; and Zafeiriou, S. 2019. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1155–1164.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Gross, R.; Matthews, I.; Cohn, J.; Kanade, T.; and Baker, S. 2010. Multi-pie. *Image and Vision Computing*, 28(5): 807–813.

Guo, J.; Zhu, X.; Yang, Y.; Yang, F.; Lei, Z.; and Li, S. Z. 2020. Towards Fast, Accurate and Stable 3D Dense Face Alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Guo, Y.; Cai, J.; Jiang, B.; Zheng, J.; et al. 2018. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE transactions on pattern analysis and machine intelligence*, 41(6): 1294–1307.

Hassner, T.; Harel, S.; Paz, E.; and Enbar, R. 2015. Effective face frontalization in unconstrained images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4295–4304.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, Y.; Wu, X.; Yu, B.; He, R.; and Sun, Z. 2018. Pose-guided photorealistic face rotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8398–8406.

Huang, R.; Zhang, S.; Li, T.; and He, R. 2017. Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis. In *The IEEE International Conference on Computer Vision (ICCV)*.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 172–189.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; and Kavukcuoglu, K. 2015. Spatial transformer networks. *arXiv preprint arXiv:1506.02025*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, G.-H.; and Lee, S.-W. 2020. Uncertainty-aware mesh decoder for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6100–6109.

Lee, M.; Cho, W.; Kim, M.; Inouye, D.; and Kwak, N. 2020. StyleUV: Diverse and High-fidelity UV Map Generative Model. *arXiv preprint arXiv:2011.12893*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Ma, F.; Ayaz, U.; and Karaman, S. 2018a. Invertibility of Convolutional Generative Networks from Partial Measurements. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Ma, F.; Ayaz, U.; and Karaman, S. 2018b. Invertibility of convolutional generative networks from partial measurements. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 9651–9660.

Mahendran, A.; and Vedaldi, A. 2015. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5188–5196.

Masi, I.; Tran, A. T.; Hassner, T.; Sahin, G.; and Medioni, G. 2019. Face-specific data augmentation for unconstrained face recognition. *International Journal of Computer Vision*, 127(6): 642–667.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.

Paysan, P.; Knothe, R.; Amberg, B.; Romdhani, S.; and Vetter, T. 2009. A 3D Face Model for Pose and Illumination Invariant Face Recognition. In *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments*. Genova, Italy: IEEE.

Pérez, P.; Gangnet, M.; and Blake, A. 2003. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, 313–318.

Pumarola, A.; Agudo, A.; Martinez, A. M.; Sanfeliu, A.; and Moreno-Noguer, F. 2018. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 818–833.

Qian, Y.; Deng, W.; and Hu, J. 2019. Unsupervised face normalization with extreme pose and expression in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9851–9858.

Ravi, N.; Reizenstein, J.; Novotny, D.; Gordon, T.; Lo, W.-Y.; Johnson, J.; and Gkioxari, G. 2020. Accelerating 3D Deep Learning with PyTorch3D. *arXiv:2007.08501*.

Richardson, E.; Sela, M.; and Kimmel, R. 2016. 3D face reconstruction by learning from synthetic data. In *2016 fourth international conference on 3D vision (3DV)*, 460–469. IEEE.

Richardson, E.; Sela, M.; Or-El, R.; and Kimmel, R. 2017. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1259–1268.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Shen, Y.; Gu, J.; Tang, X.; and Zhou, B. 2020. Interpreting the Latent Space of GANs for Semantic Face Editing. In *CVPR*.

Tran, L.; and Liu, X. 2018. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7346–7355.

Tran, L.; Yin, X.; and Liu, X. 2017. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1415–1424.

Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, 1398–1402. Ieee.

Wu, X.; He, R.; Sun, Z.; and Tan, T. 2018. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11): 2884–2896.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*.

Yin, X.; Huang, D.; Yang, H.; Fu, Z.; Wang, Y.; and Chen, L. 2020. Pixel Sampling for Style Preserving Face Pose Editing. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, 1–10. IEEE.

Zhou, H.; Liu, J.; Liu, Z.; Liu, Y.; and Wang, X. 2020. Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5911–5920.

Zhu, J.-Y.; Krähenbühl, P.; Shechtman, E.; and Efros, A. A. 2016a. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, 597–613. Springer.

Zhu, X.; Lei, Z.; Liu, X.; Shi, H.; and Li, S. Z. 2016b. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 146–155.