# Detailed 3D Face Reconstruction from Single Images Via Self-supervised Attribute Learning

Mingxin Yang
NLPR, CASIA
School of AI, Univ. of CAS
yangmingxin18@ia.ac.cn

Jianwei Guo*
NLPR, CASIA
School of AI, Univ. of CAS
jianwei.guo@nlpr.ia.ac.cn

Juntao Ye
NLPR, CASIA
yejuntao@gmail.com

Xiaopeng Zhang
NLPR, CASIA
School of AI, Univ. of CAS
Xiaopeng.Zhang@ia.ac.cn

## ABSTRACT

We present a novel approach to reconstruct high-fidelity geometric human face model from a single RGB image. The main idea is to add details into a coarse 3D Morphable Model (3DMM) based model in a self-supervised way. Our observation is that most of the facial details like wrinkles are driven by expression and intrinsic facial characteristics which here we refer to as the facial attribute. To this end, we propose an expression related details recovery scheme and a facial attribute representation.

## KEYWORDS

3D Face Reconstruction; 3DMM; Facial Attribute Learning;

**ACM Reference Format:**
Mingxin Yang, Jianwei Guo, Juntao Ye, and Xiaopeng Zhang. 2020. Detailed 3D Face Reconstruction from Single Images Via Self-supervised Attribute Learning. In *SIGGRAPH Asia 2020 (SA '20 Posters), December 04-13, 2020.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3415264.3425455

## 1 INTRODUCTION

3D realistic reconstruction of human face from a single RGB image is a challenging task, because face geometry is not easy to recover due to the gap between the geometric space and texture space. Researchers propose the 3DMM [Blanz and Vetter 2003], which makes it possible to recover facial shape and albedo from a single RGB image. Some variants of 3DMM further add facial expression to the model. However, the intrinsic drawback of 3DMM limits the model's ability to capture facial details. Therefore, realistic and high-fidelity human face reconstruction are still an open problem.

A number of methods are proposed to reconstruct realistic human face by breaking the problem into first recovering a coarse face model and then adding details [Chen et al. 2019]. These works use image-to-image translation paradigm and regress details directly from the input face image. While the results are promising, they suffer two drawbacks. First, these methods rely on the high-quality 3D face data, which is rather laborious to capture. Second, the details generated by these methods are only suitable to current individual in current expression, in other words, not riggable. [Yang et al. 2020] propose to solve this problem by regressing riggable details but they still rely on 3D face dataset to train the network.

Our approach tries to deal with above two problems simultaneously by treating facial details as being driven by two factors: the facial expression and the facial attribute. Then we propose to train a neural network with only 2D image data in a self-supervised way. We combine these two factors and insert this process into a self-supervised paradigm.

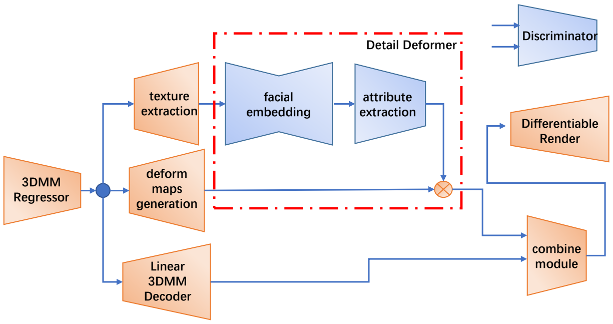

**Figure 1: The overall framework of our approach. Only the blue blocks need to be trained in our pipeline.**

## 2 OUR APPROACH

We propose a coarse-to-fine approach. As shown in Fig. 1, our pipeline includes four modules, the *3DMM Regressor* for dimentional regressing the 3DMM coefficients and environment parameters from the input image, the *Detail Generator* which infers associated details driven by facial expression, the *Differentiable Render* for self-supervised training, and the *Discriminator* for adversarial training.

### 2.1 3DMM Regressor.

Given a 224x224 RGB facial image, the regressor outputs a 257 dimentional parameter vector. We adopt the state-of-the-art pre-trained regressor model [Deng et al. 2019] in our implementation. Then we can derive 3D face model with these parameters by Eq. 1:

$$S = \overline{S} + A\alpha + B\beta, T = \overline{T} + C\gamma, \tag{1}$$

where $S \in R^{3\times N}$ is the 3D face with $N$ vertices, $\overline{S} \in R^{3\times N}$ is a mean face shape, $A$, $B$, $C$ are identity, expression and albedo basis respectively and $\alpha$, $\beta$, $\gamma$ are corresponding parameters. We establish the expression representation-map by computing the mesh vertices differences caused by expression. We then extract the facial texture from image.

[*] Corresponding author: jianwei.guo@nlpr.ia.ac.cn

*SA '20 Posters, December 04-13, 2020, Virtual Event, Republic of Korea*
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8113-0/20/12.
https://doi.org/10.1145/3415264.3425455

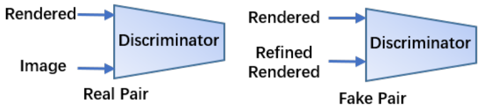

**Figure 2: The input pairs for constructing our adversarial loss.**

## 2.2 Detail Generator.

Our *detail generator* module tries to regress facial skin attribute representation and apply it on the expression-map we derived. Here,we regard facial attribute as local filters which directly convolve on the face expression map to get detail representation. According to this assumption, we cast the problem as regressing local filters from image, which is similar to the idea of dynamic convolution. Therefore, we adopt the method of DRConv [Chen et al. 2020] to design our attribute filter regression network, which involves regressing several local region masks and corresponding local filters shared in each region. For the network structure, we directly adopt a pixel-to-pixel generator to get attribute embedding and insert a layer of DRConv after it. For detail representation, we choose displacement map in the uv-space.

## 2.3 Differentiable Renderer.

To train our network self-supervisedly, we take advantage of the differentiable rendering technique, which aims at rendering face mesh to a 2D image in a differentiable way. With this rendering layer, we could compute the difference between rendered image and input image in order to construct our loss function.

## 2.4 Discriminator.

The discriminator is designed for adversarial training, which aims at improving the mapping from rendered image to refined rendered image.

## 2.5 Loss Function.

Four losses are adopted in our training process of *detail generator* module, including:

*2.5.1 Mask pixel-wise loss.* This is the basic loss of our framework, which tries to minimize the difference between the rendered image and input image. We adopt a pre-trained face segmentation method to generate a facial mask on this loss in order to only compute facial region difference as supervision.

*2.5.2 Perceptual loss.* This loss compares the facial image and rendered image in feature space. We adopt FaceNet as our feature extraction network and using cosine distance in feature space to supervise the network.

*2.5.3 Adversarial loss.* The two losses introduced above has been used in training the 3DMM Regressor network. Therefore, we add adversarial loss by treating detail regression problem as an image-to-image translation problem(from coarsely rendered images to real images). We construct an adv loss used in lsgan [Mao et al. 2017].

*2.5.4 Detail Regularization loss.* In our experiments, self-supervised training provides chaos results at the beginning and is hard to converge. To regularize our training process, we add a detail regularization loss in the early stage of training process. That is, we minimize the point differences between detail reconstruction and coarse reconstruction in order to regularize the detail generator. Then we gradually reduce the detail regularization loss coefficient:

$$L_{DetailReg}(x_{coarse}, x_{refined}) = \frac{1}{N}\left\|x_{coarse} - x_{refined}\right\|_2, \quad (2)$$

where $x_{coarse}$ is the 3DMM face shape, $x_{refined}$ is the face shape after combining our detail generator outputs, $N$ is number of vertices.

## 3 RESULTS AND DISCUSSION

Fig. 3 shows some reconstruction results.

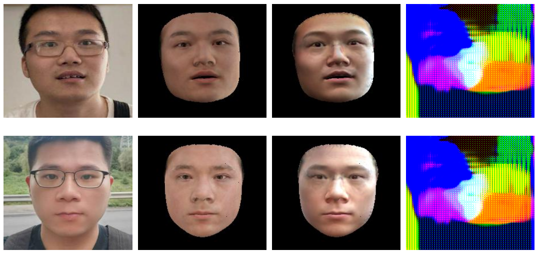

**Figure 3: From left to right: input face image, coarse reconstruction by 3DMM, our detailed reconstruction, and the facial attribute embedding feature map.**

Our current approach is an initial attempt to represent expression related facial attribute in a self-supervised way. There are still some room to improve our work. Our detail reconstruction result is not as good as the state-of-the-art model, we suspect one reason is that the detail geometry has little effect on rendered image in the refined pipeline. We will try to implement our idea in a fully trained detail reconstruction pipeline.

## ACKNOWLEDGMENTS

This work was supported by CCF–Tencent Open Research Fund (RAGR20190105), National Natural Science Foundation of China (61802406, 61761003), and the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (VRLAB2019B02).

## REFERENCES

Volker Blanz and Thomas Vetter. 2003. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on pattern analysis and machine intelligence* 25, 9 (2003), 1063–1074.

Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. 2019. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE International Conference on Computer Vision*. 9429–9439.

Jin Chen, Xijun Wang, Zichao Guo, Xiangyu Zhang, and Jian Sun. 2020. Dynamic Region-Aware Convolution. *arXiv preprint arXiv:2003.12243* (2020).

Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. 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 Workshops*. 0–0.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2794–2802.

Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. 2020. 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*. 601–610.