DF^{2}Net: A Dense-Fine-Finer Network for Detailed 3D Face Reconstruction

Xiaoxing Zeng* [1,2] Xiaojiang Peng* [1] Yu Qiao† [1]

[1]ShenZhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab,Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences, China

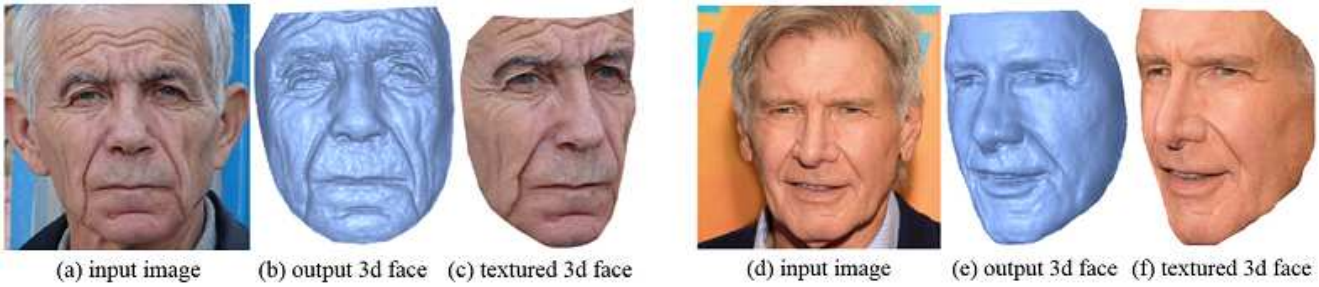


Figure 1: 3D face reconstruction results of the proposed method. Reconstructed geometries are shown next to the corresponding input images.

Abstract

Reconstructing detailed geometric structure from a single face image is a challenging problem due to its ill-posed nature and the fine 3D structures to be recovered. This paper proposes a deep Dense-Fine-Finer Network (DF^{2}Net) to address this challenging problem. DF^{2}Net decomposes the reconstruction process into three stages, each of which is processed by an elaborately-designed network, namely D-Net, F-Net, and Fr-Net. D-Net exploits a U-net architecture to map the input image to a dense depth image. F-Net refines the output of D-Net by integrating features from both depth and RGB domains, whose output is further enhanced by Fr-Net with a novel multi-resolution hypercolumn architecture. In addition, we introduce three types of data to train these networks, including 3D model synthetic data, 2D image reconstructed data, and fine facial images. Qualitative evaluation indicates that our DF^{2}Net can effectively reconstruct subtle facial details such as small crow's feet and wrinkles. Our DF^{2}Net achieves performance superior or comparable to state-of-the-art algorithms in qualitative and quantitative analyses on real-world images and the BU-3DFE dataset. Codes and the collected 70K image-depth dataset are publicly available[1].

1. Introduction

This paper considers the problem of high-fidelity 3D face reconstruction from a single image. Image-based 3D face reconstruction is a fundamental yet important problem in computer vision, with wide applications in face animation[16, 34] ,human-machine interaction[1], medical applications [2, 29], etc. The challenge of this problem comes from its ill-posed nature and the fine facial details to be recovered. The projection from a 3D face to a 2D image depends on its material properties, lighting conditions, viewing directions and other factors. Given an input image, there usually exist multiple choices of 3D structures to generate this image. Moreover, faces always include subtle structures like wrinkles, eye grains, which are difficult to recover accurately.

Early approaches toward this problem can be roughly divided in two categories, namely 3D Morphable Model (3DMM) [4] based methods and Shape-from-Shading (SFS) [41] based ones. 3DMM-based methods represent a textured 3D face as a low-dimensional representation in terms

*Equally-contributed first authors ({xx.zeng,xj.peng}@siat.ac.cn)
†Corresponding author (yu.qiao@siat.ac.cn)

[1]https://github.com/xiaoxingzeng/DF2Net/

of latent variables and corresponding basis vectors. These SFS-based methods utilize rendering principle to recover the underlying shape from shading observations. Though 3DMM-based methods are efficient and simple, they always lead to over-smooth results and can not capture the rich details of input images, partly due to its low-dimensional nature. Compared with 3DMM-based methods, SFS-based methods allow to recover more fine geometric details, but it always needs accurate prior shape information and the iterative optimization process of SFS can be sensitive to noise.

Recently, deep convolutional neural network (CNN) based methods achieve impressive progresses on the recovery of 3D facial geometry [11, 35, 19, 44, 10]. Roth et.al [27] exploit landmark driven 3D warping to produce prior 3D faces and then adjust their geometry with photometric methods. Sela *et al.* [28] predict depth image and correspondence map, by using non-rigid transformation to obtain 3D facial mesh, and further refine it with an iterative process. Compared to previous hand-crafted feature works, deep networks can learn effective features to estimate the mapping from 2D images to their 3D shapes. The recent state-of-the-art methods [25, 22, 11] exploited a coarse-to-fine CNN framework in an end-to-end fashion, where the coarse CNN model mainly regresses the low-dimensional and smooth representation of a 3DMM and the fine CNN model applies a shape-from-shading like refinement to capture the fine facial details. The reconstruction performance mainly depends on both coarse model and the refinement model. However, capturing subtle 3D structure from a single image is still a challenging task.

In this paper, we focus on the coarse-to-fine framework and aim to advance the state-of-the-art with respect to coarse 3D face model and refinement model. Specifically, we propose a novel coarse-to-fine framework, termed as *Dense-Fine-Finer Network* (DF^{2}Net), to reconstruct a high-fidelity 3D face surface from a single image. DF^{2}Net consists of three modules, namely a Dense depth Network (D-Net), a Fine network (F-Net), and a Finer network (Fr-Net), which perform 3D reconstruction in a cascaded way. For the coarse 3D face model, instead of regressing the low-dimensional 3D representation as in previous works, our D-Net estimates a coarse but dense depth map from input image with a U-net [26] like architecture. We train D-Net with both artificial images generated by 3DMM with different poses/illumination conditions, and real images with 3D surfaces obtained by existing reconstruction algorithms [8, 37, 35]. Our coarse model is superior to previous works in that our coarse estimation leverages both 3DMM and other state-of-the-art models, which allows D-Net to capture richer and denser details than previous coarse modules. Surprisingly, with this simple change on training data, the D-Net already obtains performance on par with most of the state-of-the-art methods.

For the refinement model, we propose a fine-to-finer architecture which consists of F-Net and Fr-Net. F-Net takes as input the dense map obtained by D-Net and the original face image. F-Net efficiently fuses the features from depth maps and images for reconstruction. Although being more accurate than D-Net, the output of F-Net still lack the ability to capture the subtle structures and fine details like forehead wrinkles, crows feet. To further enhance the details, we introduce the Fr-Net to estimate a depth displacement (or residual) map with features from different layers and different domains (i.e. depths and RGB images). Fr-Net integrates the original color image into its input layer and multi-resolution features from middle layers. We design multi-resolution hypercolumn blocks to gradually correct some local depth errors in different resolutions inspired by the recent depth denoising work [38].

The main contributions of this paper are summarized as follows,

- We propose a novel Dense-Fine-Finer architecture for high-fidelity 3D face reconstruction. DF^{2}Net decomposes the reconstruction into three cascaded stages, and we introduce different training strategies together with different datasets for each stage.
- We elaborately design a novel Fr-Net which effectively integrates features from different levels and domains with multi-resolution hypercolumn blocks. With the Fr-Net, our DF^{2}Net obtains state-of-the-art qualitative performance in real-world images.
- We collect about 70k high quality image-depth pairs for training 3D face reconstruction networks. We will make this dataset with codes and models public after the publication of this paper.

2. Related Works

Face reconstruction from a single image has been studied extensively in computer vision and computer graphics communities. There are many methods try to handle the intrinsic ambiguities in this problem. This paper mainly reviews the relate methods in this section.

3DMM methods. The 3D morphable model (3DMM) [4] is one of the most popular methods for the task of face reconstruction from a single image [33, 36]. Though this method achieves very impressive 3D face reconstruction performance, 3DMM suffers from high computational costs and requires manual manipulation to align the mean 3D facial shape to the 2D facial image during initialization. To solve this problem, Blanz et al [3] use a sparse set of facial feature landmarks to initialize the 3D face shape. Similarly, the Basel Face Model (BFM) [23] reconstructs 3D face by regressing 2D facial landmarks via a principal component regression (PCR) model. Though 3DMM can provide

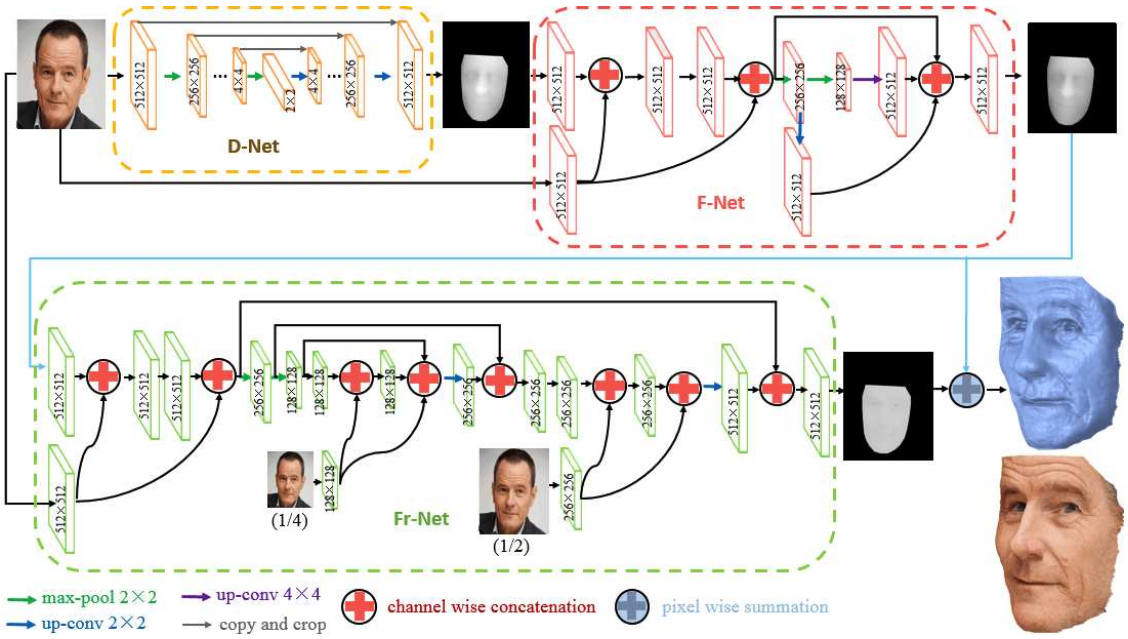


Figure 2: The architecture of our proposed network

whole 3D face from a single image, the facial details like folds and wrinkles may not be well captured since they are not spanned by the principal components. Our proposed method uses image-depth pairs synthesized by 3DMM and several state-of-the-art reconstruction methods to train a coarse dense reconstruction network. Perhaps the most related work is the Image-to-Image Translation method [28] which estimates the depth map with image-depth pairs synthesized by 3DMM. Our method differs from [28] in that i) our synthetic image-depth pairs are richer which lead to better coarse 3D face model, ii) we elaborately design a fine-to-finer refinement architecture for detailed reconstruction.

Shape-from-Shading (SFS) based methods. SFS [41] is a computer vision technique that recovers the underlying shape from shading variation in 2D images by image rendering principle. Given the lighting coefficient and reflectance parameters, SFS method can revover subtle geometry details with a optimization process. However, SFS is a typical ill-posed problem with well-known ambiguities such as the convex/concave ambiguity [24]. Solving such ambiguities for face reconstruction requires handling priors on the facial surface, i.e. SFS needs a reliable initial 3D shape. For example, Kemelmacher-Shlizerman and Basri [21] use a single 3D reference to align input facial image manually. With this prior information, a shape from shading method was then exploited to recover the geometry. The characteristic of facial symmetry has been used by various researchers to constrain the problem [43, 42, 31]. As an efficient and simple low-dimensional representation method, 3DMM is also widely used as an initial facial shape [27, 10, 20, 22, 32]. In this paper, we use the dense reconstruction of D-Net as an initial shape rather than the low-dimensional parametric 3DMM-based shape.

Deep learning based methods. With the help of convolution neural networks, many deep learning based 3D face reconstruction methods get admirable results. Deep learning based methods can be classified as coarse face reconstruction and dense face reconstruction. For coarse face reconstruction, face landmarks or 3DMM parameters are treated as supervision signal for traning. Patrik *et al.* [14] apply a L_2 cost function between 2D facial landmarks and plane projection of estimated 3D landmarks for shape fitting. Yang *et al* [39] proposed weighted landmark regression method for 3DMM shape fitting. Yao *et al.* [8] encode 3D coordinates of vertexs into a position map, and process 2D position map with convlutional nueral network learning. Dense reconstruction methods predict dense shape variation rather than low-dimensional parameters. Tran *et al.* [35] estimate a coarse 3D face shape which acts as a foundatioin, and then separately layer this foundation shape with details represented by a bump map. Huynh et al [15] produce plausible facial detail at medium and fine scales from texture maps with learning-based approach. Feng et al [8] train their proposed FaceLFnet with light field images to reconstruct 3D faces. Richardson et al [25] use a coarse-to-fine network to learn detailed face reconstruction, where a CoarseNet is applied to estimate coarse shape with 3DMM model and a FineNet is trained to refine de-

tails of a 3D face with shape-from-shading based unsupervised constraints. There are many other algorithms which use coarse-to-fine architecture to learn detailed face reconstruction [28, 22, 20, 11, 27]. Our method differs from these coarse-to-fine works by the elaborately-designed fine-to-finer refinement model which allows high-fidelity 3D face reconstruction.

3. Methods

In this section, we first overview our Dense-Fine-Finer (DF^{2}Net) framework for high-fidelity 3D face reconstruction, and then introduce our training data, and finally we respectively present its architecture, components and training details.

3.1. Overview of Our DF^{2}Net

As illustrated in Figure 2, our proposed DF^{2}Net framework consists of three sub-networks, namely a Dense depth Network (D-Net), a Fine network (F-Net) and a Finer network (Fr-Net). A face image is first input to the D-Net for dense but rough 3D reconstruction. Then the output depth map of D-Net along with the orignal image are fed into the F-Net (a hypercolumn network) for refinement. The refined depth map of F-Net and the multi-scale face images are jointly processed with into our elaborately-designed Fr-Net to estimate a residual depth map for finer detail reconstruction. We detail our training data and each components of DF^{2}Net in the following sections.

3.2. Training Data

Training data plays an essential role in deep learning based methods. For 3D facial reconstruction, it is very difficult and time-costly to collect real face images together their ground truth 3D shapes (usually capture with depth devices). Current public 3D face datasets only include a few hundreds of subjects, which are insufficient for training deep networks with millions of parameters.

To address this problem, we construct three types of data which are elaborately designed for training in different stages of DF^{2}Net. 1) *3D model synthetic dataset*. We generate different 3D facial shapes using 3DMM and render 2D image with these shapes, similar to existing methods [28, 25]. Specifically, we generate identity, expression [6] and texture basis elements randomly for constructing 3D textured shapes. To make our method robust to illumination, we render 2D images under different light conditions. We generate 20K 3D models with morphable model, and project each 3D model to image plane with 3 different poses which leads to 60K image-depth pairs. One problem of using 3DMM is that it always lacks fine structures like wrinkles, in real images. To alleviate this problem, we introduce another synthetic data: 2) *2D image synthetic dataset*. We collect 10K 2D face images from the Internet where 5K of them include smooth faces (e.g faces of youth or makeup woman) and the rest of them contain rich facial structures (e.g. man or the old) . We obtain the 3D shapes of these real face images by using previous reconstructing methods [8] and [35] (each method processes 5k images). Then these reconstructed shapes are projected into frontal view. Totally, we obtain 70K image-depth training pairs, which are mainly used to train D-Net and F-Net. Several examples of image-depth pairs are shown in Figure 3. We experimentally find the deep networks trained with the above two datasets still meet difficulties in reconstructing facial shapes with fine structures (e.g faces of old man). To address this problem, we further construct the third dataset, 3) *2D image dataset with rich details*. Specifically, we select 5K face images with rich facial details from the CACD dataset [5]. The details of these facial images are hard to reconstruct by existing methods. We use them for the unsupervised training of Fr-Net. A summarization of training data for each component of our DF^{2}Net is presented in Table 1.

3.3. Dense Network

D-Net is a pixel-level mapping network, which generates target depth maps from input images. We choose U-Net [26] as its architecture due to its impressive performance in many related pixel-wise tasks [18, 17, 28, 11]. Compared to previous coarse models [25, 22, 11] which regress the low-dimensional representation of 3DMM, our D-Net learns to recover dense depth information directly.

Training. We use two loss functions to train D-Net with the image-depth pairs from both the 3D model synthetic dataset and the 2D image synthetic dataset constructed in Section 3.1. The first loss is a pixel-wise L_1 loss between ground-truth and predicted depth maps,

$$L_{rec} = \|D_{gt} - D_{pre}\|_1 , \tag{1}$$

where D_{gt} is the ground-truth depth, and D_{pre} is the predicted depth. As shown in [28], L1 loss often lead to oversharpen predictions in local areas. Hence, we also use another L_1 constraint on the normals of depth maps with the formulation as follows,

$$L_{nor} = \|N_{gt} - N_{pre}\|_1 , \tag{2}$$

where N_{gt} and N_{pre} are normals of gound-truth depth and prediction depth, respectively. For the computation of normal vectors, we take surrounding 4 pixels into account. During training, the loss weight of L_{rec} is set to 10 by default, and to 0.1 for L_{nor}.

3.4. Fine Network

With the large amout of training data and powerful CNN, our D-Net already provides 3D surface with coarse shapes such as large wrinkles and facial parts. However,

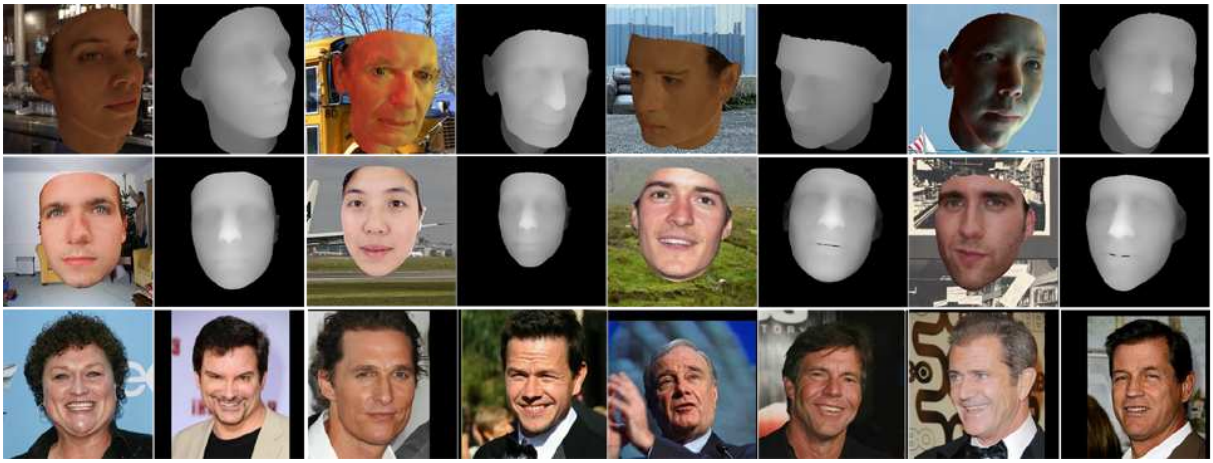


Figure 3: Examples of training images and their target depth maps.The first row is generated by 3DMM, the second row by state-of-the-art (SOA) methods [8, 35], and the third row selected from CACD

Table 1: The summarization of training data for different stages in DF^{2}Net.

	3D model synthetic data:60K from 3DMM	**2D image reconstructed data:**10K from the internet reconstructed by [8] and [35]	**Fine facial images:**5K from CACD [5]
D-Net	✓	✓	
F-Net		✓	
Fr-Net			✓

some fine details of realistic faces are not able to reconstructed due to the lack of fine and accurate details in the synthesized training data. To reconstruct the fine details, we introduce fine reconstruction network (F-net) and resort to a SFS-based refinement approach which is demonstrated as an effective way to recover high-frequency information of geometry [25, 11].

The fundamental of SFS-based refinement method is an image formation term, which illustrates the connection between the reconstructed depth map and the input intensity image. The image formation term is usually defined as follows,

$$\hat{I}(l, N, R) = R \sum_{i=1}^{9} l_i H_i(N), \tag{3}$$

where $\hat{I}$ is reflected irradiance, R is the albedo map estimated by the SfSNet [30] in our work, $H_i(N)$ is the basis function of spherical harmonics (SH) obtained by unit depth normal N. l is the second-order SH coefficients, which can be obtained by solving the following least squared problem,

$$l^* = \arg\min_{l} \left\| R \sum_{i=1}^{9} l_i H_i(N_{gt}) - I \right\|_2^2, \tag{4}$$

where I is the input intensity image. N_{gt} is the normal of ground-truth depth.

Training. F-Net takes both RGB and depth as input and has a hypercolumn architecture, which allows to efficiently fuse the output responses from different convolution layers along the forward path. Similar architecture is also adopted in [12, 25]. The detailed architecture is illustrated in Figure 2. Different from [12, 25], the input layer and the middle layer of F-Net integrate both depth and RGB features, which is inspired by the recent depth denoising work [38]. In this way, F-Net can capture fine structures from both color image domain and depth domain.

We use SFS refinement criterion to train F-Net with the 2D image synthetic dataset (Table 1). Given R and l^*, we train the fine reconstruction network to predict a fine depth map conditioned by for normal vectors. Denote the normal of a predicted depth map as N_{pre}, we use the following pixel-wise shading loss,

$$L_{sh} = \left\| \hat{I}(l^*, N_{pre}, R) - I \right\|_2. \tag{5}$$

L_{sh} penalizes the intensity difference between the rendered image and the original input image, and drives the

network to recover fine details. For training F-Net, we use both L_{rec} and L_{sh} with the weight 10 and 0.1, respectively. Experimental results demonstrate that our F-net can learn fine facial details beyond the 3DMM and other reconstruction method used for data generation.

3.5. Finer Network

Since the synthesized depths are used for training both D-Net and F-Net, the reconstruction is more or less impacted by the 3DMM and other data generation methods. To get rid of the impact of these synthesized training image-depth pairs, we further introduce a novel finer reconstruction network, i.e. Fr-Net, and train it with another 5K 2D face images from CACD dataset [5] which contain complicated geometries and are hard to be reconstructed by D-Net and F-Net. We aim to reconstruct subtler details of 3D face surface by Fr-Net.

As depicted in Figure 2, Fr-Net has two novel aspects compared to previous reconstruction networks. As the first aspect, it integrates the original color image of different sizes at the input layer and several middle layers, which is called *multi-resolution hypercolumn blocks*. In addition to RBG images, F-Net aslo takes as input the depth map estimated by F-Net. As the CNN feature map resolution decreases to 1/4 and 1/2, we integrate similar blocks with the input layer where the RBG images are resized according the sizes of feature maps. By integrating the multi-resolution images step-by-step, we can gradually correct local depth errors with contexts of different resolutions. Similar block proves efficient in depth denoising [38]. As the second aspect, we add a residual connection from the input depth map to the output depth map, which ensures our Fr-Net retains the similar geometry as the input depth. In this way, Fr-Net learns to regress a displacement/residual depth map which makes training easier and improves generalization [13].

Training. Using the collected 5K images, we train our Fr-Net with both reconstruction and SFS losses, similar as F-Net. Specifically, we use both L_{rec} and L_{sh} with the loss weight 1 and 100, respectively. Here, the "ground-truth" depth used in L_{rec} and L_{sh} is approximated by the output depth map of F-Net.

3.6. Implementation Details

DF^{2}Net, which is naturally implemented in a cascaded way, can be trained end-to-end stochastic gradient descent (SGD). Since the SFS process is time-consuming, we implement and run it on GPU devices. As D-Net does not include the SFS process, we train our DF^{2}Net in a 3-step scheme.

In the first step, we train the D-Net with all the above-mentioned 70K image-depth pairs with a batchsize of 64. In the second step, we fix D-Net, and train the F-Net with a batchsize of 8. Similarly, we fix D-Net and F-Net, and train the Fr-Net with a batchsize of 8 in the last step. It is worth noting that the small batchsize in the last two steps is owing to the SFS implementation on GPUs. The sizes of input image and output depth map are fixed to 512×512. For all the steps, we initialize the learning rate to 0.001, and divide it by 10 at epoch 8 and 9, and we stop training after the 10 epoch. We implement the deep networks with Pytorch toolbox, and run training in 4 NVIDIA Tesla K40c GPUs.

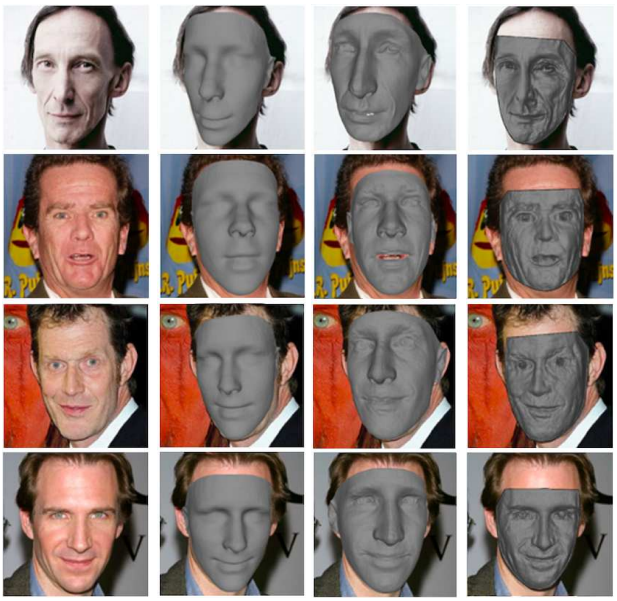


Image PRNet [8] ExtremeNet[35] Ours

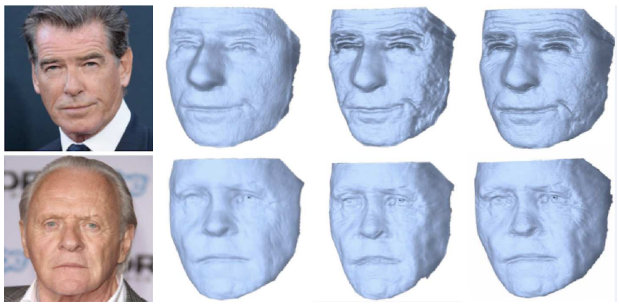
Figure 4: Qualitative comparison for our DF^{2}Net, PRNet [8], and ExtremeNet [35]. Note that PRNet [8] and ExtremeNet [35] are used to generate a part of our training data. Best viewed in PDF

4. Experiments

In this section, we first conduct an experimental exploration for each module of DF^{2}Net, and then compare to several state-of-the-art singe image based 3D face reconstruction methods in both qualitative and quantaitative aspects.

4.1. Exploration of DF^{2}Net

Since PRNet [8] and ExtremeNet [35] are used to generate the ground-truth depth maps of real-world images for our training, we conduct a comparison between our DF^{2}Net and these two methods to check our improvement upon them. Figure 4 shows a qualitative comparison. PRNet uses a simple encoder-decoder network to directly regress the 3D facial structure and dense alignment with training data generated by 3DMM. As can be seen in Figure 4, it is only good at reconstructing the smooth 3D facial surfaces. As a method which can capture some local structure with relative



(a) (b) (c) (d)

Figure 5: Qualitative evaluation of alternative dense-fine-finer architectures. From left to right: (a) original images, *(b) results of stacking two F-Net*, *(c) results of Fr-Net without multi-resolution hypercolumn blocks*, and (d) results of our default DF^{2}Net. Best viewed in PDF.

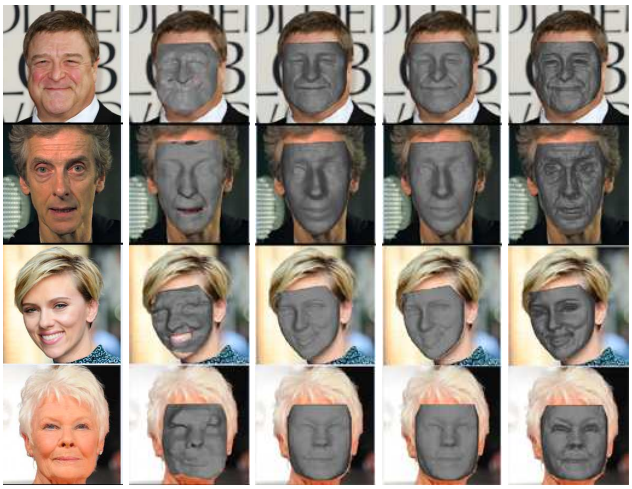


Image Coarse [28] D-Net F-Net Fr-Net

Figure 6: Our qualitative reconstruction results in different stages. Best viewed in PDF.

high speed, ExtremeNet is chosen as another data generation method since it can provide facial details and is robust to extreme conditions such as occlustions and large poses. Still, our DF^{2}Net is obviously superior to ExtremeNet [35] in reconstructing facial details such as cheek wrinkles.

Alternative dense-fine-finer architectures. We also consider two another dense-fine-finer architectures in our work, namely i) stacking two F-Net and ii) Fr-Net without multi-resolution hypercolumn blocks. The first one means that the default Fr-Net of our DF^{2}Net is replaced by another F-Net without skip connection. The second one uses the same Fr-Net architecture without multi-resolution module. The training data of these alternative architectures is the same as the one of default DF^{2}Net. Figure 5 shows a qualitative comparison between our default DF^{2}Net and the alternative architectures. Several observations can be concluded as follows. First, although these alternative ones provide sufficient facial details, our default DF^{2}Net reconstructs more accurate subtle details such as forehead wrinkles and crow's feet. Second, the second alternative one performs better than the naive stacking one which demonstrates the effectiveness of the skip connection. Last but not the least, the superiority of default DF^{2}Net upon the second alternative one indicates that the multi-resolution hypercolumn blocks are helpful for reconstructing subtle facial details.

Qualitative results in each stage of DF^{2}Net. To further investigate the reconstruction process of DF^{2}Net, we present the qualitative results in each stage in Figure 6. As shown in Figure 6, the reconstructed 3D faces include increasing details from D-Net to Fr-Net. Surprisingly, D-Net already obtains detailed 3D face reconstruction on par with several state-of-the-art methods [7, 11, 22, 25, 28], see the results in Figure 7. Though F-Net provides more subtle details in cheeks than D-Net, both D-Net and F-Net are limited by the generated ground truth of training data. In the finer reconstruction stage, richer details like forehead wrinkles and crow's feet are reconstructed thanks to elaborately-designed Fr-Net and its training strategy.

Since [28] uses the same U-Net to estimate a dense depth as its coarse model, we also illustrate the coarse results of [28] in Figure 6. Our coarse results (outputs of D-Net) are superior to [28], thanks to the large-scale mixed training data from both 3DMM and state-of-the-art algorithms.

4.2. Comparison to the state of the art

Qualitative Comparison. We compare our DF^{2}Net to several state-of-the-art algorithms[25, 11, 7, 22, 28] in Figure 7. The results of [7] are mostly limited by 3DMM since it only reconstructs the smooth surfaces. Compared to these coarse-to-fine methods [25, 11, 22, 28], our DF^{2}Net captures complete shapes including teeth, and reconstructs finer details such as beard and crow's feet.

Quantitative Comparison We conduct a quantitative analysis on the popular BU-3DFE dataset [40] which contains facial images aligned with ground truth 3D shape. The BU-3DFE dataset includs 100 subjects, and we conduct the evaluation on all subjects.

We use the Dlib toolbox to detect the landmark with 68 points, and create a binary mask according to the outer points to represent the valid pixels in the evaluation. Following [28], we apply Random Sample Concensus (RANSAC) approach [9] to normalize the estimated depth to ground truth. We compute the absolute depth error and evaluate its mean, standard deviation, median, and the average 90% largest error. The quantitative comparison between our method and other state-of-the-art algorithms is presented in Table 2. As shown in Table 2, our method achieves

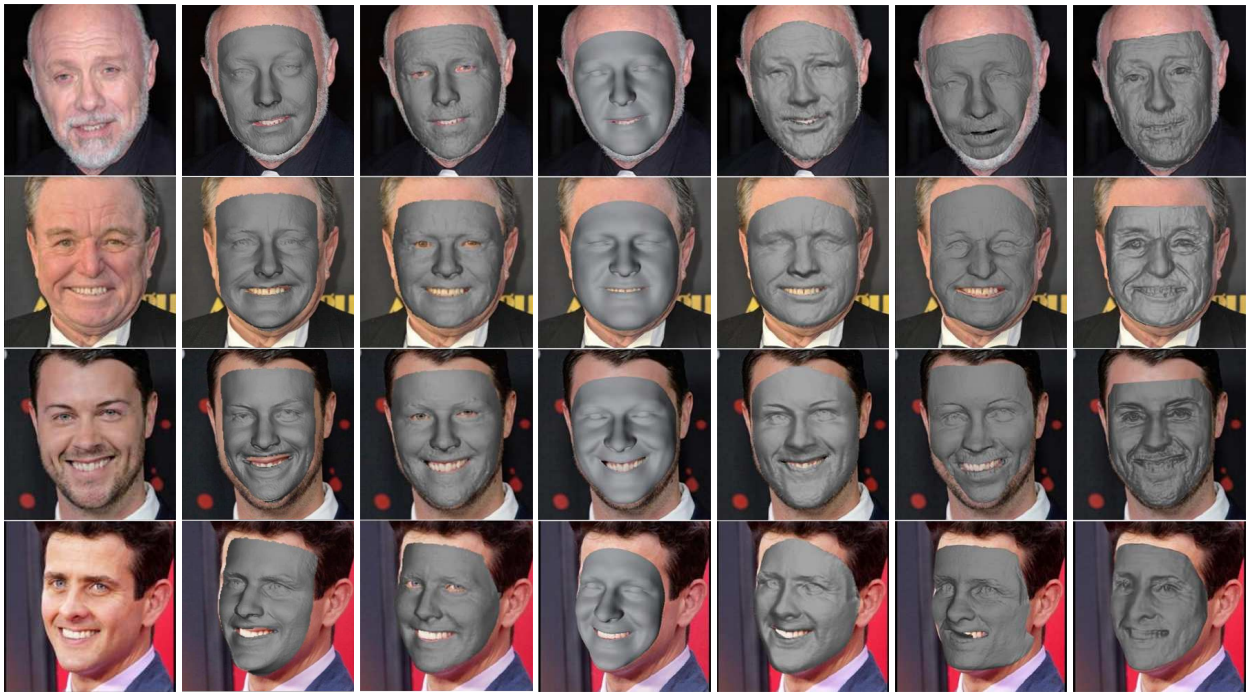


Figure 7: Qualitative comparison between our DF^{2}Net and several state-of-the-art methods.

	Mean Err.	Std Err.	Median Err.	90% Err.
SFS [21]	3.89	4.14	2.94	7.34
HPEN [44]	3.85	3.23	2.93	7.91
Coarse-Fine Net[25]	3.61	2.99	2.72	6.82
Pix2vertex [28]	3.51	2.69	2.65	6.59
Ours	3.37	2.59	2.53	6.41

Table 2: Quantitative comparison results on the BU-3DFE Dataset.

lower depth error than the other algorithms. We also give a analysis about absolute error distribution of whole face, which depicted in Fig.8.

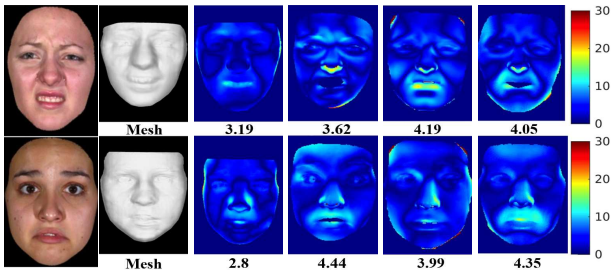


Figure 8: Comparison of error heat maps in percentile of ground truth depth between our DF^{2}Net and several state-of-the-art methods.

5. Conclusion

This paper proposes a deep Dense-Fine-Finer Network (DF^{2}Net) to address the challenging problem of high-fidelity 3D face reconstruction from a single image. DF^{2}Net is composed of three modules, namely D-Net, F-Net, and Fr-Net. It progressively refines the subtle facial details such as small crow's feet and wrinkles. We introduce three types of data to train DF^{2}Net with different training strategies. We also evaluate several alternative choices for the Dense-Fine-Finer framework to demonstrate the efficiency of our DF^{2}Net. DF^{2}Net achieves performance superior or comparable to state-of-the-art methods in qualitative and quantitative analyses on real-world images and BU-3DFE.

Acknowledgement.
This work is partially supported by National Natural Science Foundation of China (U1813218U1613211),the National Key Research and Development Program of China (No. 2016YFC1400704) Shenzhen Research Program (JCYJ20170818164704758CXB201104220032A), the Joint Lab of CAS-HK.

References

[1] Oleg Alexander, Mike Rogers, William Lambeth, Jen-Yuan Chiang, Wan-Chun Ma, Chuan-Chang Wang, and Paul Debevec. The digital emily project: Achieving a photorealistic digital actor. *IEEE Computer Graphics and Applications*, 30(4):20–31, 2010.

[2] Israel Amirav, Anthony S Luder, Asaf Halamish, Dan Raviv, Ron Kimmel, Dan Waisman, and Michael T Newhouse. Design of aerosol face masks for children using computerized 3d face analysis. *Journal of aerosol medicine and pulmonary drug delivery*, 27(4):272–278, 2014.

[3] Volker Blanz, Albert Mehl, Thomas Vetter, and H-P Seidel. A statistical method for robust 3d surface reconstruction from sparse data. In *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, pages 293–300. IEEE, 2004.

[4] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. In *Siggraph*, volume 99, pages 187–194, 1999.

[5] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *European conference on computer vision*, pages 768–783. Springer, 2014.

[6] Baptiste Chu, Sami Romdhani, and Liming Chen. 3d-aided face recognition robust to expression and pose variations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1899–1906, 2014.

[7] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision*, 126(12):1269–1287, 2018.

[8] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551, 2018.

[9] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[10] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8377–8386, 2018.

[11] Yudong Guo, Jianfei Cai, Boyi Jiang, Jianmin Zheng, et al. 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, 2018.

[12] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[14] Patrik Huber, Guosheng Hu, Rafael Tena, Pouria Mortazavian, P Koppen, William J Christmas, Matthias Ratsch, and Josef Kittler. A multiresolution 3d morphable face model and fitting framework. In *Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016.

[15] Loc Huynh, Weikai Chen, Shunsuke Saito, Jun Xing, Koki Nagano, Andrew Jones, Paul Debevec, and Hao Li. Mesoscopic facial geometry inference using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8407–8416, 2018.

[16] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. Dynamic 3d avatar creation from hand-held video input. *ACM Transactions on Graphics (ToG)*, 34(4):45, 2015.

[17] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):107, 2017.

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[19] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1031–1039, 2017.

[20] Luo Jiang, Juyong Zhang, Bailin Deng, Hao Li, and Ligang Liu. 3d face reconstruction with geometry details from a single image. *IEEE Transactions on Image Processing*, 27(10):4756–4770, 2018.

[21] Ira Kemelmacher-Shlizerman and Ronen Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE transactions on pattern analysis and machine intelligence*, 33(2):394–405, 2010.

[22] Yue Li, Liqian Ma, Haoqiang Fan, and Kenny Mitchell. Feature-preserving detailed 3d face reconstruction from a single image. In *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, page 1. ACM, 2018.

[23] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301. Ieee, 2009.

[24] Yvain Quéau, Jean Mélou, Fabien Castan, Daniel Cremers, and Jean-Denis Durou. A variational approach to shape-from-shading under natural illumination. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 342–357. Springer, 2017.

[25] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1259–1268, 2017.

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[27] Joseph Roth, Yiying Tong, and Xiaoming Liu. Unconstrained 3d face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2606–2615, 2015.

[28] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1576–1585, 2017.

[29] Matan Sela, Nadav Toledo, Yaron Honen, and Ron Kimmel. Customized facial constant positive air pressure (cpap) masks. *arXiv preprint arXiv:1609.07049*, 2016.

[30] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *CVPR*, 2018.

[31] Ilan Shimshoni, Yael Moses, and Michael Lindenbaum. Shape reconstruction of 3d bilaterally symmetric surfaces. *International Journal of Computer Vision*, 39(2):97–110, 2000.

[32] William AP Smith and Edwin R Hancock. Recovering facial shape using a statistical model of surface normal direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1914–1930, 2006.

[33] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. 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*, pages 2549–2559, 2018.

[34] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016.

[35] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard G Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *CVPR*, pages 3935–3944, 2018.

[36] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7346–7355, 2018.

[37] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. 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*, pages 5163–5172, 2017.

[38] Shi Yan, Chenglei Wu, Lizhen Wang, Feng Xu, Liang An, Kaiwen Guo, and Yebin Liu. Ddrnet: Depth map denoising and refinement for consumer depth cameras using cascaded cnns. In *ECCV*, 2018.

[39] Yu Yanga, Xiao-Jun Wu, and Josef Kittler. Landmark weighting for 3dmm shape fitting. *arXiv preprint arXiv:1808.05399*, 2018.

[40] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006.

[41] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 21(8):690–706, 1999.

[42] Wen Yi Zhao and Rama Chellappa. Illumination-insensitive face recognition using symmetric shape-from-shading. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 286–293. IEEE, 2000.

[43] Wen Yi Zhao and Rama Chellappa. Symmetric shape-from-shading using self-ratio image. *International Journal of Computer Vision*, 45(1):55–75, 2001.

[44] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796, 2015.