Pattern Recognition Letters 153 (2022) 51–58

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

High-fidelity 3D face reconstruction with multi-scale details☆

Yiwei Jin, Qingyu Li, Diqiong Jiang, Ruofeng Tong*

Zhejiang University, China

ARTICLE INFO

Article history:
Received 7 April 2021
Revised 2 November 2021
Accepted 18 November 2021
Available online 22 November 2021

Edited by xxx

Keywords:
3D Face reconstruction
Feature-preserving
Multi-scale details
Coarse-to-fine

ABSTRACT

Despite tremendous success has been achieved in faithfully reconstructing face shapes from single images, recovering accurate local details still remains challenging. Previous works propose reprojection-based methods to improve the performance of detail recovering – they render a textured 3D shape into an image and make it approximate to the input during iterations. However, details from textures and shapes are mixed in the rendered image when minimizing the re-projection loss, which leads to limitations in detail recovery. To address this issue, we propose a novel 3D face reconstruction framework that 1) uses a coarse-medium-fine strategy to capture details while preserving the global shape, 2) disentangles details from the texture to enhance local accuracy, and 3) applies a phased optimization to recover details over multiple scales. Experiments demonstrate the capability of our framework to reconstruct high-fidelity face shapes with accurate, fine details.

1. Introduction

Reconstructing 3D face geometry from a single 2D image has been a research focus in the computer vision community and has a wide range of applications in face recognition [41,44,50], face animation [18,38], face manipulation [8,27], etc. With the performance of face reconstruction increasing in both quality and efficiency, there appears a surge of interest in recovering subtle details of faces, such as wrinkles and pores.

However, faithfully recovering local geometry details while preserving global shape fidelity is a challenging task. The main difficulty is a conflict between a strong constraint for preserving accurate global shapes and a weak constraint for capturing fine local details. A common solution is to separately reconstruct the shape and the details [12,39]: firstly rely on prior knowledge to enforce a reasonable shape, and then apply a lower constraint to extract fine details. Especially, differentiable renderers are proposed to address the lack of 3D data and capture details directly from the input.

By using renderers and re-projection losses, previous works [19,23,30,37] are able to project current 3D results back into the 2D space and implement minimization on the difference between this image and the input. However, on such rendered images, texture details and geometry details are mixed. In this way, a rendered image produced by a smooth geometry and detailed textures may be regarded to have minimal difference with the input, which actually is a bad detail-recovering result. As far as we know, little work has been done on disentangling the details between texture and geometry to solve this issue and enhance the detail reconstruction.

In this paper, we propose our solution to the problem: when recovering details, we capture internal face details from the input image and reduce them from the 2D space, so that the geometry is forced to recover more 3D details during the reconstruction. We propose a coarse-medium-fine framework for single-image reconstruction that produces 3D faces with global fidelity as well as local details. In the coarse stage, we reconstruct a smooth face shape by using a prior 3D morphable model (3DMM); in the medium stage, we apply a landmark-conducted Laplace deformation to fine-tune the coarse geometry while preserving the global fidelity; and in the fine stage, we use face segmentation and relative-total-variation [47] algorithms to increase both accuracy and robustness of the albedo, and design a staged shape-from-shading (SFS) optimization to recover multi-scale geometry details from the input image. Furthermore, we design a staged strategy for recovering geometry details of multiple scales, including large-scale wrinkles and small-scale pores (Fig. 1).

Overall, our contributions are summarized as follows:

- We propose a novel 3D face reconstruction framework to capture high-fidelity face shapes from single images. Our framework progressively enhances the reconstruction results in multiple levels and finally produces realistic and distinguishable face geometries.
- We propose to enhance geometry details in face reconstruction by reducing 2D details from the image space, and design a

☆ Editor: Maria De Marsico.

* Corresponding author.
E-mail address: trf@zju.edu.cn (R. Tong).

https://doi.org/10.1016/j.patrec.2021.11.022
0167-8655/

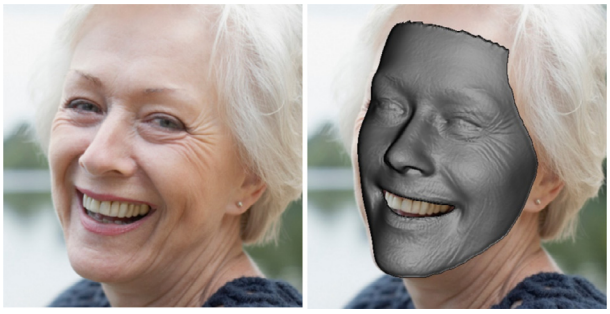

Fig. 1. Demonstration of our reconstruction framework: (left) the input image; (right) the reconstructed geometry.

detail-reducing term as well as an average-skin-color term for the recovering optimization.

- We design a staged strategy to capture subtle image features over multiple scales. We demonstrate it in sufficient experiments that our proposed method has the capability to reconstruct a high-fidelity geometry with visual discrimination.

2. Related work

Model-based face reconstruction. The first 3DMM [5] is a generic 3D morphable face model that uses principal component analysis (PCA) to represent 3D face shapes and textures with linear bases. This seminal work has stimulated a large amount of relevant work afterward. Basel Face Model (BFM) [32] uses better scanning devices and Nonrigid Iterative Closest Point (NICP) registration algorithm of higher performance to improve previous models. Furthermore, by adding bases of different attributes, multilinear models are proposed for higher expressiveness. For example, FaceWarehouse [8] and BFM-2017 [6] build bilinear models by adding expression bases, while FLAME [28] develops an extra pose variation based on facial joints. Huge success has been achieved in model-based face reconstruction [4,13,30,36,43]. However, these methods have an inherent theoretical defect that causes a lack of visual discrimination and geometry details. On the one hand, there is a mismatch between the distribution of sampled faces in datasets and real faces in daily life, making the reconstruction results inaccurate (e.g., a model built from Caucasian faces may fail in modeling Asian faces). On the other hand, high-frequency features are dropped during decomposing, which causes a lack of details in the reconstructed faces.

With the increasing popularity of deep learning techniques, researchers use one fully connected layer without activation functions to represent 3DMM bases. To improve model expressiveness, some works replace it with multiple fully connected layers or convolution layers to improve the expressiveness of the networks [10,40,41]. For example, Tran and Liu [41] use deep convolution neural networks to encode input images into shape, texture and projection parameters and decode the parameters to apply a reprojection loss between rendered images and real images. More recently, works using graph convolution networks [31,45] have been proposed to further improve reconstruction results [25,26]. The performance of reconstruction is further enhanced through such non-linear face models, yet most of these methods rely heavily on the quantity and quality of 3D data and have relatively high training costs.

Recovering facial details. The aforementioned methods using prior face models are able to produce globally reliable face geometries whereas having fewer details due to an ill-posed problem of balancing the requirements of strong regularization for preserving the global shape and weak regularization for capturing subtle details. In model-free methods, some previous works use facial landmarks as constrains of alignment to fit 3D faces [2,20]. However, sparse landmarks cannot capture sufficient facial details over a dense geometry, thus such methods usually generate generic 3D face shapes without subtle details. Shape-from-shading (SFS) [17] is a classic computer vision technique for recovering geometries from images. Given camera pose, surface reflectance and object shading, an image can be rendered under certain illumination models through SFS theory. Thereby 3D face shapes can be recovered by minimizing the difference between an input image and a rendered image. SFS-based methods [22,42] are able to recover fine geometric details yet are limited in assumption models and rely on initial values to achieve reliable results.

In recent years, a common strategy, known as coarse-to-fine or coarse-medium-fine, has been proposed to separate the reconstruction process for recovering both global shape and local details [9,16,19,23,34]. During the coarse stage, prior knowledge, such as face models and templates, is used to produce a preliminary result of 3D face reconstruction. Next, a fine stage that implements techniques such as SFS [19,29] and DCNNs [33,37] is introduced to capture fine details on the geometry. For example, Richardson et al. [33] first rely on a 3DMM to reconstruct a coarse face shape from a given image, then render the 3D shape into a depth map to integrate it with details through a deep network. Zeng et al. [49] use three cascaded CNN modules to produce a depth map from the input image and gradually add local features. Rather than training costly deep networks, a latest work [23] produces a smooth 3D face and enhances it using local corrective deformation fields, and finally implements the SFS optimization to recover fine geometric details. The similarity between Khan et al. [23] and us lies in the coarse-medium-fine strategy that we use to produce multi-level reconstruction results. While we differ from them in 1) we use a pre-trained network [51] to estimate 3DMM parameters of the face shapes rather than using landmark alignment to fit an example-based face model; 2) we construct a dense correspondence between augmented 2D landmarks and 3D vertices to conduct the Laplace deformation; 3) we reduce details from the texture to enhance 3D details and implement a staged SFS optimization to recover multi-scale details.

3. Methods

We propose a coarse-medium-fine approach for 3D face reconstruction from a single image. In this section, we will show specific theories and methods used in our framework.

3.1. Overview

Our overall framework is composed of three stages, which is shown at the top of Fig. 2 from left to right. In the coarse stage, we input an image into pre-trained CNNs to estimate shape parameters of a 3DMM model and pose parameters for projection. In the medium stage, we detect dense landmarks in the input image to conduct a Laplace deformation over the coarse shape. In the fine stage, where lies our main contribution, we implement a shape-from-shading optimization for integrating subtle details on depth maps. Specifically, as shown in the bottom of Fig. 2, we start with a mean texture and a medium shape to respectively estimate initial values of illumination, albedo and shading components for the SFS optimization. Besides, we design a skin-color loss and a detail-reducing loss for albedo optimization, which are respectively based on face segmentation and relative-total-variation algorithms. After implementing the optimization over multiple scales, realistic fine details are integrated into the depth map.

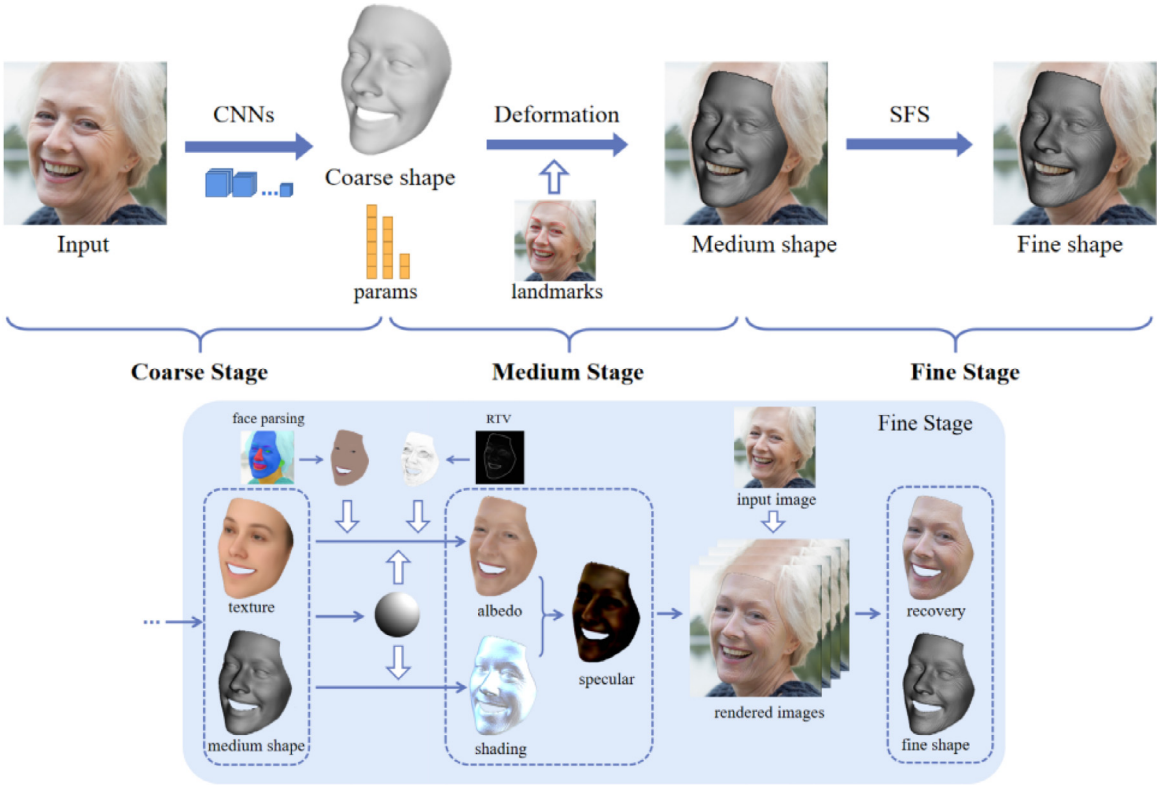

Fig. 2. The overview of our coarse-medium-fine framework. We reconstruct 3D faces with high fidelity and subtle details in three stages: 1) predict 3DMM parameters from an input image, 2) deform the coarse shape for landmark alignment, and 3) integrate subtle details over the geometry.

Fig. 3. Face detection results: 1000 landmarks per face.

Table 1
Face-to-vertex 3D RMSE comparison on MICC dataset (in mm).

Method	Cooperative	Indoor	Outdoor
Tuan Tran et al. [44]	1.97±0.49	2.03±0.45	1.93±0.49
Genova et al. [14]	1.78±0.54	1.78±0.52	1.76±0.54
Deng et al. [11]	1.66±0.52	1.66±0.46	1.69±0.53
Ours	1.59±0.37	1.55±0.43	1.61±0.28

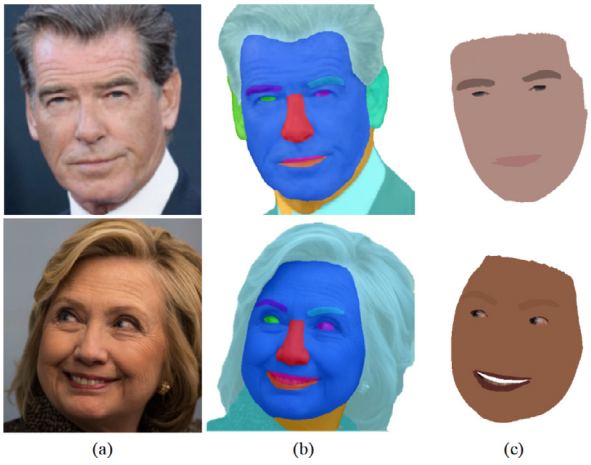

Fig. 4. Face segmentation results. Column (a) are input images, (b) face parsing results, (c) average skin colors.

3.2. Coarse stage: a foundation shape

We use a 3DMM that includes identity and expression bases to represent 3D face shapes. Briefly, we model a face shape S and a face texture T as:

$$S = \overline{S} + A_{id}\alpha_{id} + A_{exp}\alpha_{exp} \quad (1)$$

where $\overline{S}$ indicates the average face shape; $A_{id} \in \mathbb{R}^{80}$, $A_{\text{exp}} \in \mathbb{R}^{64}$ are PCA bases of identity and expression; α_{id}, α_{exp} are corresponding parameter vectors. Specifically, We adopt $\overline{S}$ and A_{id} from BFM [6] and A_{exp} from FaceWarehouse [8].

Given an input image, we use a publicly available method of 3DDFA [51] to estimate face parameters α_{id}, α_{exp} in Eq. (1) and a 6-DOF head pose vector p to generate a foundation face shape. Some results of this stage in Section 4 show that the coarse face is globally approximate to the ground truth but carries local errors. Such coarse geometries are far below our expectation, but enough for an initialization of the subsequent work.

3.3. Medium stage: enhance local shape

In this stage, we enhance the smooth foundation face shape in a global scale by using a landmark conducted Laplace deformation to fine-tune those inaccurate local parts while preserving the global structure.

Laplace deformation [35] utilizes differential properties of the surface to operate meshes over an intrinsic representation which is invariant to locally linearized rigid transformations. The Laplace

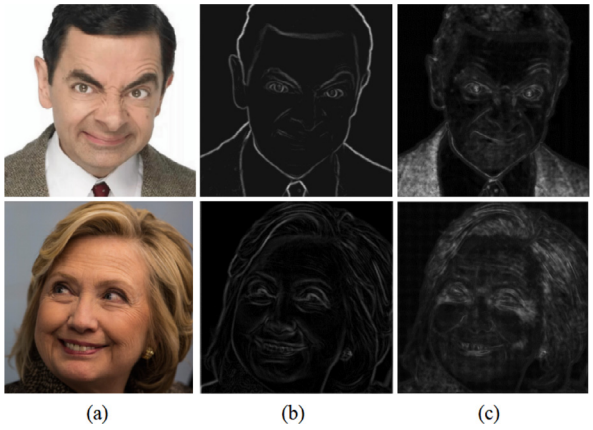

Fig. 5. Detail detection results. Column (a) are input images, (b) are results of Laplace operator, (c) are results of RTV.

coordinate $L(\cdot)$ is given by encoding each vertex relative to its neighborhood as the following:

$$L(v_i) = v_i - \sum_{j=1}^{N_i} \omega_{ij} v_j \tag{2}$$

where N_i is the number of adjacent vertices of v_i, ω_{ij} is the weight between two adjacent vertices while $\sum \omega_{ij} = 1$. Here we take the reciprocal of N_i as the value of ω_{ij}.

Traditional methods that use 2D landmarks to guide the 3D manipulation, usually require a face detection algorithm to detect landmarks on 2D images, and a 3D face model that provides vertex indices of corresponding landmarks. However, 2D landmark detection sometimes falls into inaccuracy owing to large pose variation (especially yaw pose) in the image, which leads to extra correcting steps before further processing. Besides, most existing face models provide 68-point [7,21] or 49-point [46] landmark annotation, which is too sparse compared with tens of thousands of vertices in a face geometry. For better performance and higher efficiency, we use the commercial detector Face++ to automatically detect 1000 face landmarks (shown in Fig. 3) and establish a dense correspondence to skip the calibration step by using index maps. Similar to texture maps, the index map is projected from a 3D face while taking the index of a vertex as the value of a pixel. Given a point of a texture map, by searching the corresponding index map for a pixel with the same coordinate, we are able to quickly find the corresponding vertex in the geometry.

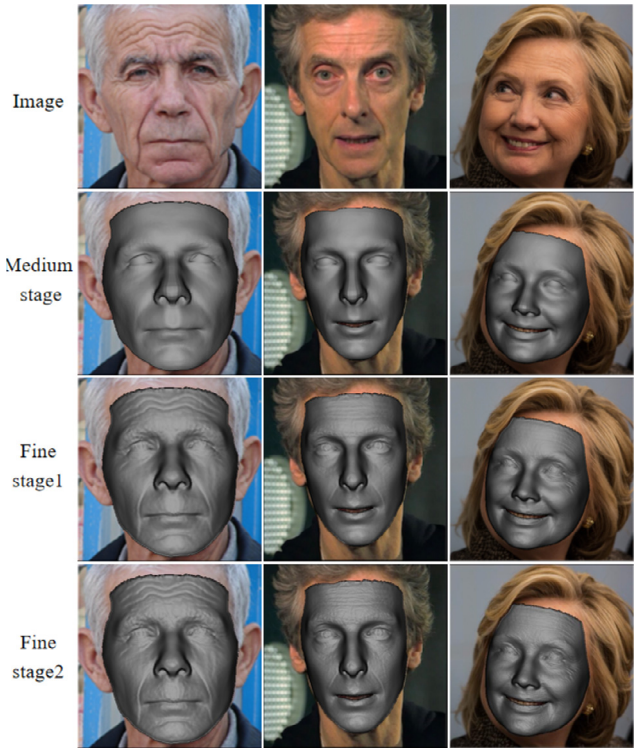

Fig. 7. Reconstruction results of the medium stage and the two phases of the fine stage.

For an initial geometry $V \in \mathbb{R}^N$ and a deformed geometry V', we define our error function of Laplace deformation as:

$$E(v') = \sum_{i=1}^{N} \left\| L(v_i') - L(v_i) \right\|_2 + \sum_{i=1}^{M} w_i \left\| v_i' - \mu_i \right\|_2 \tag{3}$$

where $L(\cdot)$ indicates the Laplace coordinate, μ_i is one of M 3D landmarks and w is a weight vector. Assuming that $s(x, y)$ is a landmark point in an input image and $v(x', y', z')$ is the corresponding vertex in a face geometry, we approximately take μ as

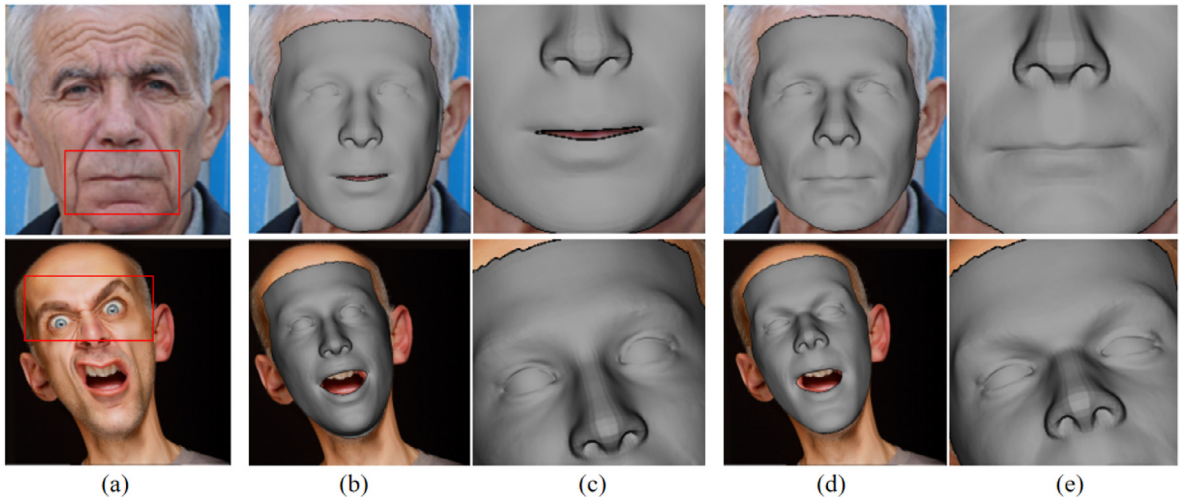

Fig. 6. Efficacy validation of the medium stage: a) input images with a red box highlighting the focusing region, (b) results of the coarse stage, (d) results of the medium stage. Column (c) and (d) are close-ups of the focusing region in coarse and medium shapes respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

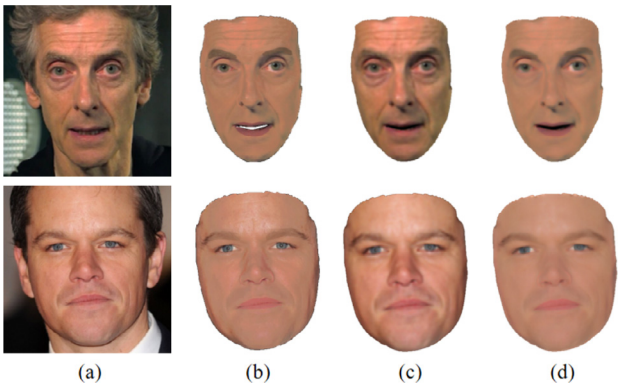

Fig. 8. Albedos with different loss terms. In each row, an input image is followed by three albedos optimized under (b) the skin-color term only, (c) the skin-color term with a detail-reducing term using Laplace operator, and (d) the skin-color term with a RTV-based detail-reducing term.

(x, y, z'). Such approximation is not accurate especially for those boundary points, thus we manually take higher weights for central landmarks and lower weights for the boundary.

By applying a landmark-conducted Laplace deformation over the coarse shape, we make the face geometry better aligned to the input image in local shape.

3.4. Fine stage: integrate subtle details

During the fine stage, we render the face shape into a depth map and apply SFS optimization over it to recover high-fidelity details from the input image. Overall, we first estimate coefficients of the spherical harmonics lighting model, then successively estimate the albedo and the specular lighting, and finally use these values to optimize the depth map.

Compared with parameterized face models, SFS has the advantage in recovering more subtle and more realistic local details (such as wrinkles and teeth). Lambertian model [3,15] is a common illumination model used in SFS problems. As shown in Eq. (4), we base on the Lambertian model and decompose an image I into three components: albedo γ, shading $S(\cdot)$ and specular β:

$$I = \gamma \cdot S(n) + \varepsilon\beta \tag{4}$$

$$S(n) = \sum_{k=1}^{m} r_k Y_k(n) \tag{5}$$

where n indicates a normal vector and ε is the weight of specular term. The shading term $S(\cdot)$ is defined as a function of 3D shapes based on spherical harmonics lighting, where $Y(n_i)$ are the given spherical harmonics bases and r are the spherical harmonics coefficients to be estimated. While predicting r, we use the mean texture in BFM and normal of the medium face as initiations of albedo and shape, and set $\varepsilon = 0$ in Eq. (4) to ignore the specular term. In this way, the objective function becomes an overdetermined equation and can be solved with least square method.

After getting an approximate illumination, we are to optimize with designed skin-color and detail-reducing loss terms for albedos with higher fidelity and less local details. The new objective function is defined as the following:

$$E(\gamma) = \|I - I\|_2^2 + \lambda_{\gamma 1}\|\gamma - \gamma_{avg}\|_2^2 + \lambda_{\gamma 2}\|RTV(\gamma))\|_2^2 \tag{6}$$

where the first term in Eq. (6) is the pixel constraint between a rendered image I' and the input image I, the second term is a skin-color loss between the albedo γ and the average skin color γ_{avg} with a weight $\lambda_{\gamma 1}$, and the third term is a detail-reducing loss with a weight $\lambda_{\gamma 2}$. Here we use a normal map calculated from the depth map to fill in Eq. (4) for a rendered image.

For getting the above γ_{avg}, we apply a face segmentation algorithm [48] to divide a face into several semantic regions and compute a mean skin color for each region (including eyebrow, mouth and skin part). As shown in Fig. 4, this term allows albedos to retain original skin colors as much as possible and reduces the influence of illumination. In addition, it helps remove the illumination from a texture in practice.

As mentioned before, we design a detail-reducing term for the albedo optimization. It is based on the relative total variation technique to extract internal details of a face. The $RTV(\cdot)$ is the sum of results in two directions, where the result along the X-axis is defined in Eq. (7) and the result along the Y-axis shares a symmetric formulation with it. It's worth mention that edge detection operators, such as Canny and Prewitt, have a similar capability to capture sharp changes of image brightness. However, these operators are much sensitive to face contours rather than internal details, which probably leads to indistinct facial contours with remained local details (shown in Fig. 5).

$$RTV_x(\gamma) = \frac{G_{R(\gamma)} * |\delta_x(\gamma)|}{\left|G_{R(\gamma)} * \delta_x(\gamma)\right|} \tag{7}$$

where $G_{R(\gamma)}$ is the Gaussian kernel with a window size of $R(\gamma)$, $\delta_x(\gamma)$ indicates the gradient of γ along the X axis, $*$ indicates a convolution operation and $|\cdot|$ means taking the absolute value.

Next, we are to estimate the specular illumination that was set to 0 before. We simply use an approximate value β_0, which is the residual between the input image and the current rendered image, to optimize the specular β:

$$E(\beta) = \left\|I' - I\right\|_2^2 + \lambda_\beta\|\beta - \beta_0\|_2^2 \tag{8}$$

where λ_β is the weight of the regularization term.

Finally, for integrating details of multiple scales on the depth map, we rewrite the objective function as a function of depth map d as Eq. (9), where $\triangledown$ indicates Laplace operator, $\triangle$ indicates an image gradient operator, and λ_d are hyper-parameters of weight. The second term in Eq. (9) is designed to preserve the global shape, and the third is to remove abnormal noise; in the last term, we enforce the depth map to have similar gradients with the input image for capturing subtle local details. Setting ε to 1, we alternately iterate albedo, specular and depth map terms to produce fine details on the depth map.

$$\begin{aligned} E(d) = \quad & \|I' - I\|_2^2 + \lambda_{d1}\|d - d_0\|_2^2 + \lambda_{d2}\|\triangledown d\|_2^2 \\ & + \lambda_{d3}\|\triangle I - \lambda_{d4} \triangle d\|_2^2 \end{aligned} \tag{9}$$

Moreover, by manually adjusting the value of λ_{d4}, we implement a step-wise strategy to integrate multi-scale details on the depth map (the higher the value of λ_{d4}, the finer details we get).

4. Experiments

In this section, we show results of quantitative and qualitative experiments to demonstrate that our framework is superior and comparative to some optimizing and deep-learning methods in performance. However in efficiency, though we use matrix calculation and convolution operations instead of traditional low-efficient loops, our framework is still time-consuming compared with methods using neural networks.

4.1. Implementation details

In preliminary stages, we use the pre-trained network from 3DDFA [51] to predict a pose coefficient and 3DMM coefficients of BFM [6]. Note that since 3DDFA tends to produce bad results in

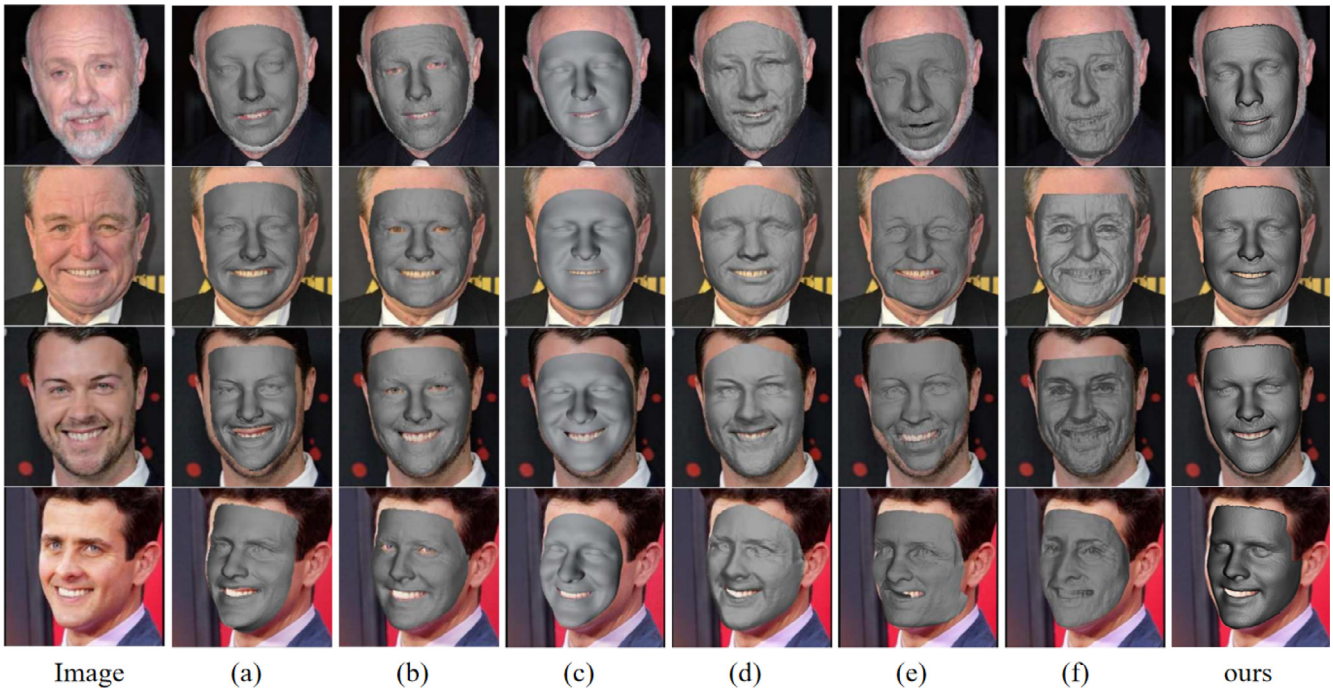

Fig. 9. Qualitative comparison between our framework and (a) [33], (b) [16], (c) [12], (d) [29], (e) [34], (f) [49].

some profile cases, we manually remove those failure cases before the medium stage. The 2D landmarks we utilize in the medium stage are from the MEGVII commercial Face++ detector. And for face parsing, we use the face segmentation algorithm of Yu et al. [48].

In the medium stage, we set $w = 0.6$ for landmarks of eye and nose region and $w = 0.1$ for eyebrow and mouth to achieve higher accuracy. And during the fine stage, we first set $\varepsilon = 0$, $\lambda_{\gamma 1} = 0.3$, $\lambda_{\gamma 2} = 2.5$ and a learning rate of 1.5 for initial values of albedo and spherical harmonics lighting. When optimizing the depth, we keep the values of λ_γ and lower the learning rate of albedo to 0.05, while setting $\varepsilon = 1$, $\lambda_\beta = 0.5$ and a learning rate of 0.25 for the specular. The depth optimization is phased into two stages: first we set $\lambda_{d1} = 0.001$, $\lambda_{d2} = 0.05$, $\lambda_{d3} = \lambda_{d4} = 0.001$ with a learning rate of 0.015 to capture local details; then we change λ_d successively into 0.001, 0.03, 0.025, 2.5 with a learning rate of 0.05 to capture finer details. We use Adam [24] optimizer for all the experiments.

4.2. Quantitative comparison

The quantitative results are based on the MICC [1] dataset. It contains both 2D and 3D data from 53 subjects, with a scanned 3D ground-truth and several video sequences (outdoor, indoor and cooperative) for each subject. We recover 3D faces for every sampled frame in a video and use the average shape as the reconstruction result of this video. We crop the meshes to 95mm around the nose tip and run the ICP (Iterative Closest Point) algorithm for alignment. As shown in Table 1, we achieve better face-to-vertex 3DRMSE results than some excellent works [11,14,44] under cooperative, indoor and outdoor conditions. Note that data of the first three rows are from Deng et al. [11]; in practice, our retested results of Deng's work are slightly higher, which is probably caused by the image preprocessing.

4.3. Qualitative comparison

We display close-ups of the coarse and the medium shape in Fig. 6 to demonstrate the effectiveness of our medium stage. Compared with coarse shapes represented by 3DMM bases, the medium shapes are improved in local accuracy through landmark alignment but remain smooth. And in Fig. 7, we demonstrate that our staged optimization has the capability to capture multi-scale details. Compared with the results of the medium stage, the results in fine stage 1, which is produced under small-scale coefficients, distinctly integrate local details over the face shape. And after optimizing with a larger scale of coefficients, the results in fine stage 2 are further enhanced in fine details.

To validate the efficacy of our proposed loss terms, we conduct an ablation study on the albedo optimization. As is shown in Fig. 8, the influence of illumination is reduced in column (b) after applying an average-skin-color constraint. The results in column (c) become fuzzy when adding a Laplace-operator-based detail-reducing term, probably because the contours are much more weakened than internal details. In column (d), by conducting a detail-reducing term based on the RTV technique, illumination and details are eliminated while facial contours are preserved. We demonstrate that it is reasonable and effective to use skin-color and detail-reducing terms for better albedos.

We also compare our result with other works in Fig. 9. Note that results of column (a-f) are from Zeng et al. [49], we add our results under same inputs to the final column for intuitive comparison. Our results are visually more natural, with high fidelity of the global shape and fine local details.

5. Conclusion

We have proposed a coarse-medium-fine framework for the challenging task of high-fidelity 3D face reconstruction from a single image. We enhance the detail recovering in reprojection-based methods by using the relative-total-variation technique for disentangling facial details from the texture. Furthermore, we apply a

phased optimization to recover details over multiple scales, which is feasible to extend for multi-view face reconstruction. Extensive experiments demonstrate the capability of our framework to reconstruct high-fidelity face shapes with accurate, fine details.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research is supported in part by National Natural Science Foundation of China (NSFC) (61972342) and the Science and Technology Department of Zhejiang Province (2018C01080).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejps.2020.105216.

References

[1] A.D. Bagdanov, A. Del Bimbo, I. Masi, The florence 2D/3D hybrid face dataset, in: Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding, 2011, pp. 79–80.
[2] A. Bas, W.A. Smith, T. Bolkart, S. Wuhrer, Fitting a 3D morphable model to edges: a comparison between hard and soft correspondences, in: Asian Conference on Computer Vision, Springer, 2016, pp. 377–391.
[3] R. Basri, D.W. Jacobs, Lambertian reflectance and linear subspaces, IEEE Trans. Pattern Anal. Mach. Intell. 25 (2) (2003) 218–233.
[4] V. Blanz, A. Mehl, T. Vetter, H.P. Seidel, A statistical method for robust 3D surface reconstruction from sparse data, International Symposium on 3D Data Processing, 2004.
[5] V. Blanz, T. Vetter, A morphable model for the synthesis of 3D faces, in: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, 1999, pp. 187–194.
[6] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, D. Dunaway, A 3D morphable model learnt from 10,000 faces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5543–5552.
[7] A. Bulat, G. Tzimiropoulos, Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3706–3714.
[8] C. Cao, Y. Weng, S. Zhou, Y. Tong, K. Zhou, FaceWarehouse: a 3D facial expression database for visual computing, IEEE Trans. Vis. Comput. Graph. 20 (3) (2013) 413–425.
[9] Y. Chen, F. Wu, Z. Wang, Y. Song, Y. Ling, L. Bao, Self-supervised learning of detailed 3D face reconstruction, IEEE Trans. Image Process. 29 (2020) 8696–8705.
[10] S. Cheng, M. Bronstein, Y. Zhou, I. Kotsia, M. Pantic, S. Zafeiriou, MeshGAN: non-linear 3D morphable models of faces, arXiv preprint arXiv:1903.10384(2019).
[11] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, X. Tong, 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, 2019.
[12] B. Egger, S. Schönborn, A. Schneider, A. Kortylewski, A. Morel-Forster, C. Blumer, T. Vetter, Occlusion-aware 3D morphable models and an illumination prior for face image analysis, Int. J. Comput. Vis. 126 (12) (2018) 1269–1287.
[13] B. Gecer, S. Ploumpis, I. Kotsia, S. Zafeiriou, GANFIT: generative adversarial network fitting for high fidelity 3d face reconstruction, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[14] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, W.T. Freeman, Unsupervised training for 3D morphable model regression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8377–8386.
[15] R. Grosse, M.K. Johnson, E.H. Adelson, W.T. Freeman, Ground truth dataset and baseline evaluations for intrinsic image algorithms, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 2335–2342.
[16] Y. Guo, J. Cai, B. Jiang, J. Zheng, et al., CNN-based real-time dense face reconstruction with inverse-rendered photo-realistic face images, IEEE Trans. Pattern Anal. Mach. Intell. 41 (6) (2018) 1294–1307.
[17] B.K. Horn, Shape from shading: a method for obtaining the shape of a smooth opaque object from one view (1970).
[18] L. Hu, S. Saito, L. Wei, K. Nagano, J. Seo, J. Fursund, I. Sadeghi, C. Sun, Y.-C. Chen, H. Li, Avatar digitization from a single image for real-time rendering, ACM Trans. Graph. (ToG) 36 (6) (2017) 1–14.
[19] L. Jiang, J. Zhang, B. Deng, H. Li, L. Liu, 3D face reconstruction with geometry details from a single image, IEEE Trans. Image Process. 27 (10) (2018) 4756–4770.
[20] A. Jourabloo, X. Liu, Large-pose face alignment via CNN-based dense 3D model fitting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4188–4196.
[21] V. Kazemi, J. Sullivan, One millisecond face alignment with an ensemble of regression trees, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1867–1874.
[22] I. Kemelmacher-Shlizerman, R. Basri, 3D face reconstruction from a single image using a single reference face shape, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2010) 394–405.
[23] A. Khan, S. Hayat, M. Ahmad, J. Cao, M.F. Tahir, A. Ullah, M.S. Javed, Learning-detailed 3D face reconstruction based on convolutional neural networks from a single image, Neural Comput. Appl. (2020) 1–14.
[24] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980(2014).
[25] T.N. Kip F, M. Welling, Semi-supervised classification with graph convolutional networks (2016).
[26] G.-H. Lee, S.-W. Lee, Uncertainty-aware mesh decoder for high fidelity 3D face reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6100–6109.
[27] C. Li, K. Zhou, S. Lin, Simulating makeup through physics-based manipulation of intrinsic image layers, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4621–4629.
[28] T. Li, T. Bolkart, M.J. Black, H. Li, J. Romero, Learning a model of facial shape and expression from 4D scans, ACM Trans. Graph. 36 (6) (2017). 194–1
[29] Y. Li, L. Ma, H. Fan, K. Mitchell, Feature-preserving detailed 3D face reconstruction from a single image, in: Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production, 2018, pp. 1–9.
[30] J. Lin, Y. Yuan, T. Shao, K. Zhou, Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5891–5900.
[31] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, M.M. Bronstein, Geometric deep learning on graphs and manifolds using mixture model CNNs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5115–5124.
[32] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, T. 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, Ieee, 2009, pp. 296–301.
[33] E. Richardson, M. Sela, R. Or-El, R. Kimmel, Learning detailed face reconstruction from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1259–1268.
[34] M. Sela, E. Richardson, R. Kimmel, Unrestricted facial geometry reconstruction using image-to-image translation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1576–1585.
[35] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, H.-P. Seidel, Laplacian surface editing, in: Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing, 2004, pp. 175–184.
[36] A. Tewari, M. Zollhfer, H. Kim, P. Garrido, F. Bernard, P. Pérez, C. Theobalt, MoFA: model-based deep convolutional face autoencoder for unsupervised monocular reconstruction, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017.
[37] A. Tewari, M. Zollhoefer, F. Bernard, P. Garrido, H. Kim, P. Perez, C. Theobalt, High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2) (2018) 357–370.
[38] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Nießner, Face2Face: real–time face capture and reenactment of RGB videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2387–2395.
[39] A.T. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, G.G. Medioni, Extreme 3D face reconstruction: Seeing through occlusions, in: CVPR, 2018, pp. 3935–3944.
[40] L. Tran, F. Liu, X. Liu, Towards high-fidelity nonlinear 3D face morphable model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1126–1135.
[41] L. Tran, X. Liu, Nonlinear 3D face morphable model, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7346–7355.
[42] G. Trigeorgis, P. Snape, I. Kokkinos, S. Zafeiriou, Face normals “in-the-wild” using fully convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 38–47.
[43] X. Tu, J. Zhao, Z. Jiang, Y. Luo, M. Xie, Y. Zhao, L. He, Z. Ma, J. Feng, 3D face reconstruction from a single image assisted by 2D face images in the wild (2019).
[44] A. Tuan Tran, T. Hassner, I. Masi, G. 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, 2017, pp. 5163–5172.
[45] H. Wei, S. Liang, Y. Wei, 3D dense face alignment via graph convolution networks, arXiv preprint arXiv:1904.05562(2019).
[46] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 532–539.
[47] L. Xu, Q. Yan, Y. Xia, J. Jia, Structure extraction from texture via relative total variation, ACM Trans. Graph. 31 (6) (2012) 1.

[48] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, BiSeNet: bilateral segmentation network for real-time semantic segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 325–341.
[49] X. Zeng, X. Peng, Y. Qiao, DF2Net: a dense-fine-finer network for detailed 3D face reconstruction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2315–2324.
[50] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, J. Feng, 3D-aided dual-agent GANs for unconstrained face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 41 (10) (2018) 2380–2394.
[51] X. Zhu, Z. Lei, X. Liu, H. Shi, S.Z. Li, Face alignment across large poses: a 3D solution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 146–155.