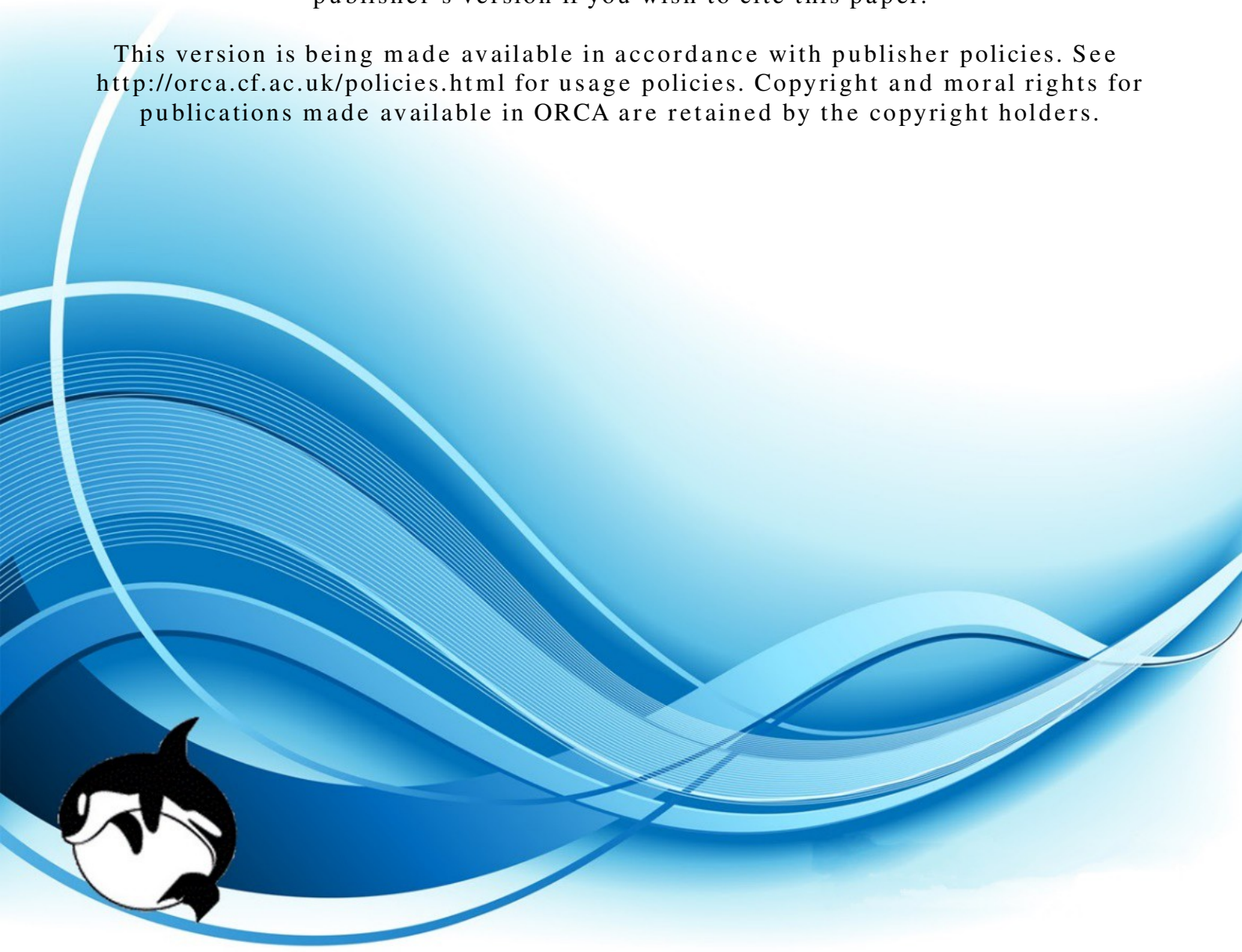


3D Face Reconstruction with Geometry Details from a Single Image

Luo Jiang, Juyong Zhang†, Bailin Deng, *Member, IEEE,* Hao Li, and Ligang Liu, *Member, IEEE,*

***Abstract*—3D face reconstruction from a single image is a classical and challenging problem, with wide applications in many areas. Inspired by recent works in face animation from RGB-D or monocular video inputs, we develop a novel method for reconstructing 3D faces from unconstrained 2D images, using a coarse-to-fine optimization strategy. First, a smooth coarse 3D face is generated from an example-based bilinear face model, by aligning the projection of 3D face landmarks with 2D landmarks detected from the input image. Afterwards, using local corrective deformation fields, the coarse 3D face is refined using photometric consistency constraints, resulting in a medium face shape. Finally, a shape-from-shading method is applied on the medium face to recover fine geometric details. Our method outperforms state-of-the-art approaches in terms of accuracy and detail recovery, which is demonstrated in extensive experiments using real world models and publicly available datasets.**

***Index Terms*—Tensor Model, Shape-from-shading, 3D Face Reconstruction.**

I. Introduction

Reconstruction of 3D face models using 2D images is a fundamental problem in computer vision and graphics [1], with various applications such as face recognition [2], [3] and animation [4], [5]. However, this problem is particularly challenging, due to the loss of information during camera projection.

In the past, a number of methods have been proposed for face construction using a single image. Among them, example-based methods first build a low-dimensional parametric representation of 3D face models from an example set, and then fit the parametric model to the input 2D image. One of the most well-known examples is the 3D Morphable Model (3DMM) proposed by Blanz and Vetter [6], represented as linear combination of the example faces. 3DMM is a popular parametric face model due to its simplicity, and has been the foundation of other more sophisticated face reconstruction methods [3]. Another approach to single image reconstruction is to solve it as *Shape-from-shading* (SFS) [7], a classical computer vision problem of 3D shape recovery from shading variation. For example, Kemelmacher-Shlizerman and Basri [8] reconstruct the depth information from an input face image, by estimating its lighting and reflectance parameters using a reference face shape.

While these existing approaches are able to produce high-quality reconstruction from a single image, they also come with limitations. Although example-based methods are simple and efficient, they rely heavily on the dataset, and may produce unsatisfactory results when the target face is largely different from those in the example set; moreover, due to the limited degrees of freedom of the low-dimensional model, these methods often fail to reproduce fine geometric details (such as wrinkles) that are specific to the target face. SFS-based methods are able to capture the fine-scale facial details from the appearance of the input image; however, they require prior knowledge about the geometry or illumination to resolve the ambiguity of the reconstruction problem, and may become inaccurate when the input image does not satisfy the assumptions.

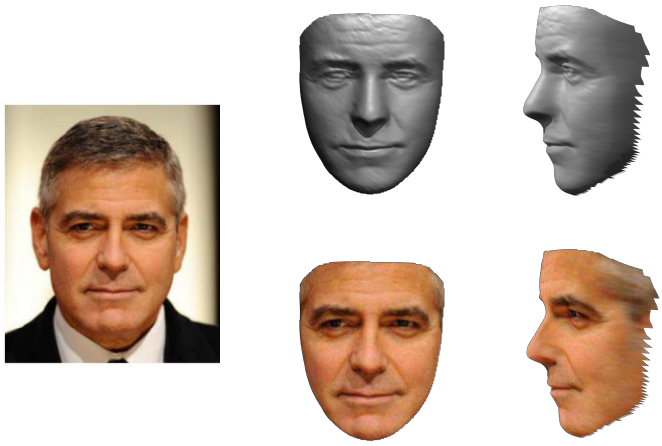


Figure 1: 3D face reconstruction from a single image. Given an input image (left), we reconstruct a 3D face with fine geometric details (right, top row). The input image can be used as texture for rendering the reconstructed face (right, bottom row).

In this paper, we propose a novel coarse-to-fine method to reconstruct a high-quality 3D face model from a single image. Our method consists of three steps:

- First, we compute a coarse estimation of the target 3D face, by fitting an example-based parametric face model to the input image. Our parametric model is derived from FaceWarehouse [9] and the Basel Face Model (BFM2009) [10], two 3D face datasets with large variation in expression and identity respectively. The resulting mesh model captures the overall shape of the target face.
- Afterwards, we enhance the coarse face model by applying smooth deformation that captures medium-scale facial features; we also estimate the lighting and reflectance parameters from the enhanced face model.

L. Jiang, J. Zhang, H. Li, and L. Liu are with School of Mathematical Sciences, University of Science and Technology of China.

B. Deng is with School of Computer Science and Informatics, Cardiff University.

†Corresponding author. Email: juyong@ustc.edu.cn.

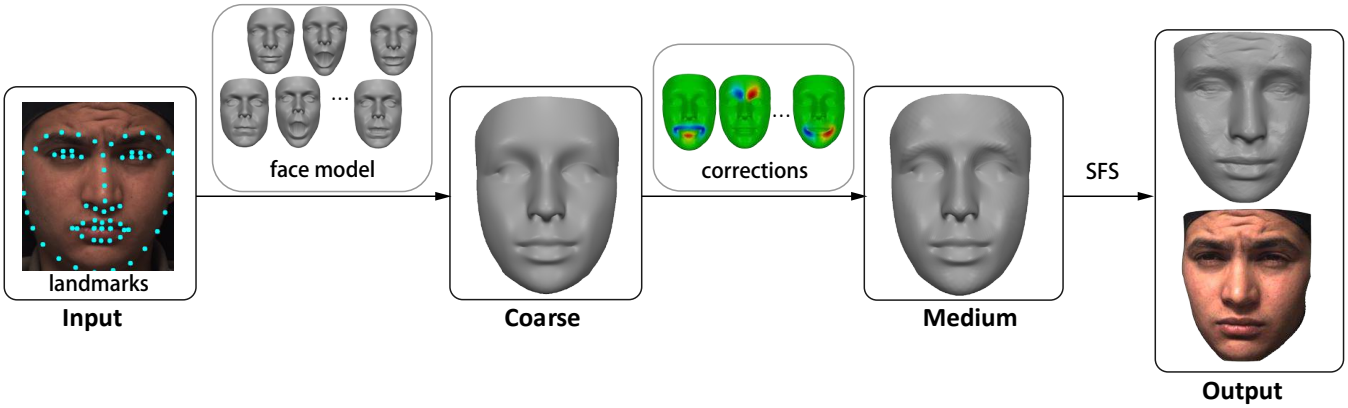


Figure 2: An overview of our coarse-to-fine face reconstruction approach.

- Finally, the illumination parameters and the enhanced face model are utilized to compute a height-field face surface according to the shading variation of the input image. This final model faithfully captures the fine geometric details of the target face (see Fig. 1).

Our method builds upon the strength of the existing approaches mentioned above: the example-based coarse face enables more reliable estimation of illumination parameters, and improves the robustness of the final SFS step; the SFS-based final face model provides detailed geometric features, which are often not available from example-based approaches. Our method outperforms existing example-based and SFS methods in terms of reconstruction accuracy as well as geometric detail recovery, as shown by extensive experimental results using publicly available datasets.

II. Related Work

Low-dimensional models. Human faces have similar global characteristics, for example the location of main facial features such as eyes, nose and mouth. From a perception perspective, it has been shown that a face can be characterized using a limited number of parameters [11], [12]. The low dimensionality of the face space allows for effective parametric face representations that are derived from a collection of sample faces, reducing the reconstruction problem into searching within the parameter space. A well-known example of such representations is the 3DMM proposed in [6], which has been used for various face processing tasks such as reconstruction [6], [13], [14], [15], [16], recognition [2], [3], face exchange in images [17], and makeup suggestion [18]. Low-dimensional representations have also been used for dynamic face processing. To transfer facial performance between individuals in different videos, Vlasic et al. [19] develop a multilinear face model representation that separately parameterizes different face attributes such as identity, expression, and viseme. In the computer graphics industry, facial animation is often achieved using linear models called blendshapes, where individual facial expressions are combined to create realistic facial movements [20]. The simplicity and efficiency of blendshapes models enable real-time facial animation driven by facial performance captured from RGBD cameras [21], [22], [23], [24], [25] and monocular videos [26], [4], [27], [5]. When using low-dimensional face representations derived from example face shapes, the example dataset has strong influence on the resulting face models. For instance, it would be difficult to reconstruct a facial expression that deviates significantly from the sample facial expressions. In the past, during the development of face recognition algorithms, various face databases have been collected and made publicly available [28]. Among them, BFM2009 provides 3DMM representation for a large variety of facial identities. Recently, Cao et al. [9] introduced FaceWarehouse, a 3D facial expression database that provides the facial geometry of 150 subjects, covering a wide range of ages and ethnic backgrounds. Our coarse face modeling method adopts a bilinear face model that encodes identity and expression attributes in a way similar to [19]. We use FaceWarehouse and BFM2009 as the example dataset, due to the variety of facial expressions and identities that they provide respectively.

Shape-from-shading. Shape-from-shading (SFS) [7], [29] is a computer vision technique that recovers 3D shapes from their shading variation in 2D images. Given the information about illumination, camera projection, and surface reflectance, SFS methods are able to recover fine geometric details that may not be available using low-dimensional models. On the other hand, SFS is an ill-posed problem with potentially ambiguous solutions [30]. Thus for face reconstruction, prior knowledge about facial geometry must be incorporated to achieve reliable results. For example, symmetry of human faces has been used by various authors to reduce the ambiguity of SFS results [31], [32], [33]. Another approach is to solve the SFS problem within a human face space, using a low-dimensional face representation [34], [35]. Other approaches improve the robustness of SFS by introducing an extra data source, such as a separate reference face [8], as well as coarse reconstructions using multiview stereo [36], [37] or unconstrained photo collections [38], [39], [40]. We adopt a similar approach which builds an initial estimation of the face shape and augment it with fine geometric details using SFS. Our initial face estimation combines coarse reconstruction in a low-dimensional face space with refinement of medium-scale geometric features, providing a more accurate initial shape for subsequent SFS processing.

III. Overview

This section provides an overview of our coarse-to-fine approach to reconstructing a high-quality 3D face model from a single photograph. Fig. 2 illustrates the pipeline of our method.

To create a coarse face model (Sec. IV), we first build a bilinear model from FaceWarehouse and BFM2009 to describe a plausible space of 3D faces; the coarse face shape is generated from the bilinear model by aligning the projection of its 3D landmarks with the 2D landmarks detected on the input image, using a fitting energy that jointly optimizes the shape parameters (e.g., identity, expression) and camera parameters. To further capture person-specific features that are not available from the bilinear model, we enhance the coarse face using an additional deformation field that corresponds to medium-scale geometric features (Sec. V); the deformation field is jointly optimized with the lighting and albedo parameters, such that the shading of the enhanced model is close to the input image. Afterwards, the resulting medium face model is augmented with fine geometric details (Sec. VI): the normal field from the medium face model is modified according to the input image gradients as well as the illumination parameters derived previously, and the modified normal field is integrated to achieve the final face shape.

IV. Coarse Face Modeling

Preprocessing. The FaceWarehouse dataset contains head meshes of 150 individuals, each with 47 expressions. All expressions are represented as meshes with the same connectivity, each consisting of 11510 vertices. The BFM2009 dataset contains 200 face meshes, and each mesh consists of 53490 vertices. In order to combine the two datasets, we first mask the face region on the head mesh from FaceWarehouse to extract a face mesh, and fill the holes in the regions of eyes and mouth, to obtain a simply connected face mesh consisting of 5334 vertices. Afterwards, we randomly sample the parameter space for BFM2009 to generate 150 neutral face models, and deform the average face model from FaceWarehouse to fit these models via nonrigid registration [41]. Then we transfer the other 46 expressions of the FaceWarehouse average face model to each of the 150 deformed face models based on the method in [41]. In this way, we construct a new dataset containing 300 individuals (150 from BFM2009 and 150 from FaceWarehouse), each with 47 expressions. We perform Procrustes alignment for all the face meshes in the dataset. Moreover, BFM2009 provides 199 principal components to span the surface albedo space, but these principal albedo components cannot be used for our new dataset directly due to different mesh connectivity. Thus we transfer their albedo information to the new mesh representation using the correspondence identified in the nonrigid registration, to construct 199 principal albedo components for our dataset. These principal components will be used in Sec V.

Bilinear face model. Following [19], we collect the vertex coordinates of all face meshes into a third-order data tensor, and perform 2-mode SVD reduction along the identity mode and the expression mode, to derive a bilinear face model that approximates the original data set. In detail, the bilinear face model is represented as a mesh with the same connectivity as those from the data set, and its vertex coordinates $\mathbf{F} \in \mathbb{R}^{3\times N_v}$ are computed as

$$\mathbf{F} = \mathbf{C}_{\mathrm{r}} \times_2 \mathbf{w}_{\mathrm{id}}^T \times_3 \mathbf{w}_{\mathrm{exp}}^T, \tag{1}$$

where N_v is the number of vertices, $\mathbf{C}_{\mathrm{r}}$ is the reduced core tensor computed from the SVD reduction, and $\mathbf{w}_{\mathrm{id}} \in \mathbb{R}^{100}, \mathbf{w}_{\mathrm{exp}} \in \mathbb{R}^{47}$ are column vectors for the identity weights and expression weights which control the face shape. Note that here we only reduce the dimension along the identity mode, in order to maintain the variety of facial expressions in the bilinear model. For more details on multilinear algebra, the reader is referred to [42].

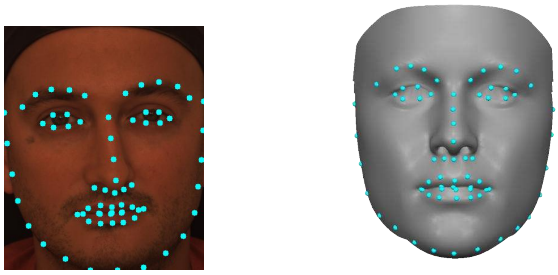


Figure 3: Our coarse face reconstruction is based on aligning the projection of labeled 3D face landmarks (right) with 2D landmarks detected on the input image (left).

To construct a coarse face, we align 3D landmarks on the bilinear face model with corresponding 2D landmarks from the input image. First, we preprocess the bilinear face mesh to manually label 68 landmark vertices. Given an input image, we detect the face as well as its corresponding 68 landmarks using the method in [43] (see Fig. 3 for an example). Assuming that the camera model is a weak perspective projection along the Z direction, we can write the projection matrix as $\mathbf{\Pi} = \begin{bmatrix} \alpha & 0 & 0 \\ 0 & \alpha & 0 \end{bmatrix}$. Then we can formulate the following fitting energy to align the projection of landmark vertices with the detected 2D landmarks

$$\begin{aligned} E_{\mathrm{fit}} &= \sum_{k=1}^{68} \|\mathbf{\Pi}\mathbf{R}\mathbf{F}_{v_k} + \mathbf{t} - \mathbf{U}_k\|_2^2 \\ &\quad + \gamma_1 \sum_{i=1}^{100} \left(\frac{w_{\mathrm{id}}^{(i)}}{\delta_{\mathrm{id}}^{(i)}}\right)^2 + \gamma_2 \sum_{j=1}^{47} \left(\frac{w_{\mathrm{exp}}^{(j)}}{\delta_{\mathrm{exp}}^{(j)}}\right)^2. \end{aligned} \tag{2}$$

Here $\mathbf{F}_{v_k} \in \mathbb{R}^3$ and $\mathbf{U}_k \in \mathbb{R}^2$ are the coordinates of the k-th 3D landmark vertex and the corresponding image landmark, respectively; translation vector $\mathbf{t} \in \mathbb{R}^2$ and rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times 3}$ determine the position and pose of the face mesh with respect to the camera; $w_{\mathrm{id}}^{(i)}$ and $w_{\mathrm{exp}}^{(j)}$ are components of weight vectors $\mathbf{w}_{\mathrm{id}}$ and $\mathbf{w}_{\mathrm{exp}}$, while $\delta_{\mathrm{id}}^{(i)}$ and $\delta_{\mathrm{exp}}^{(j)}$ are the corresponding singular values obtained from the 2-mode SVD reduction; γ_1 and γ_2 are positive weights. As in [6], the last two terms ensure parameters $w_{\mathrm{id}}^{(i)}$ and $w_{\mathrm{exp}}^{(j)}$ have a reasonable range of variation. This fitting energy is minimized with respect to the shape parameters $\mathbf{w}_{\mathrm{id}}, \mathbf{w}_{\mathrm{exp}}$ and the camera parameters $\mathbf{\Pi}, \mathbf{R}, \mathbf{t}$

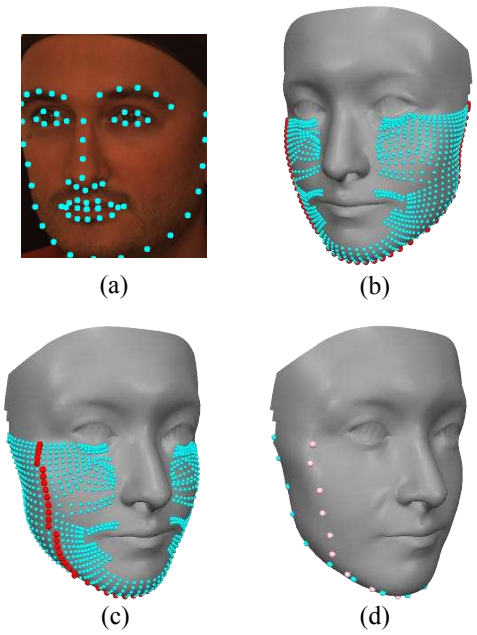


Figure 4: For a non-frontal face images (a), the labeled 3D face silhouette landmarks (shown in cyan in (d)) need to be updated for better correspondence with the detected 2D silhouette landmarks. We construct a set of horizontal lines connecting the mesh vertices (shown in cyan in (b) and (c)), and select among them a set of vertices representing the updated silhouette according to the current view direction (shown in red in (b) and (c)). The new 3D silhouette landmarks (shown in pink in (d)) are selected within the updated silhouette.

via coordinate descent. First we fix the shape parameters and reduce the optimization problem to

$$\min_{\mathbf{\Pi},\mathbf{R},\mathbf{t}} \sum_{k=1}^{68} \|\mathbf{\Pi R F}_{v_k} + \mathbf{t} - \mathbf{U}_k\|_2^2, \tag{3}$$

which is solved using the pose normalization method from [38]. Next we fix the camera and expression parameters, which turns the optimization into

$$\min_{\mathbf{w}_{\text{id}}} \sum_{k=1}^{68} \|\mathbf{\Pi R F}_{v_k} + \mathbf{t} - \mathbf{U}_k\|_2^2 + \gamma_1 \sum_{i=1}^{100} \left(\frac{w_{\text{id}}^{(i)}}{\delta_{\text{id}}^{(i)}} \right)^2. \tag{4}$$

This is a linear least-squares problem and can be easily solved by solving a linear system. Finally, we fix the camera and identity parameters, and optimize the expression parameters in the same way as Eq. (4). These steps are iteratively executed until convergence. In our experiments, four iterations are sufficient for convergence to a good result.

Landmark vertex update. The landmark vertices on the face mesh are labeled based on the frontal pose. For non-frontal face images, the detected 2D landmarks along the face silhouette may not correspond well with the landmark vertices (see Fig. 4(a) for an example). Thus after each camera parameter optimization step, we update the silhouette landmark vertices according to the rotation matrix $\mathbf{R}$, while keeping the internal landmark vertices (e.g., those around the eyes, the nose, and the mouth) unchanged. Similar to [4], we preprocess the original face mesh to derive a dense set of horizontal lines that connect mesh vertices and cover the potential silhouette region from a rotated view (see Fig. 4(b) and 4(c)). Given a rotation matrix $\mathbf{R}$, we select from each horizontal line a vertex that lies on the silhouette, and project it onto the image plane according to the camera parameters $\mathbf{\Pi}, \mathbf{R}, \mathbf{t}$. These projected vertices provide an estimate of the silhouette for the projected face mesh. Then for each 2D silhouette landmark, its corresponding landmark vertex is updated to the silhouette vertex whose projection is closest to it (see Fig. 4(d)).

To determine the silhouette vertex on a horizontal line, we select the vertex whose normal encloses the largest angle with the view direction. Since the face mesh is approximately spherical with its center close to the origin, we approximate the unit normal of a vertex on the rotated face mesh as $\frac{\mathbf{Rv}}{\|\mathbf{Rv}\|_2}$, where $\mathbf{v}$ is the original vertex coordinates. Then the silhouette vertex is the one with the smallest value of $\left|\mathbf{Z} \cdot \frac{\mathbf{Rv}}{\|\mathbf{Rv}\|_2}\right|$ within the horizontal line, where $\mathbf{Z} = [0, 0, 1]^T$ is the view direction.

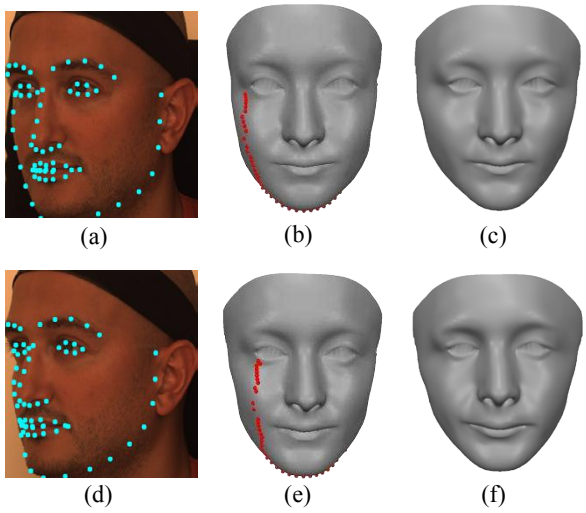


Figure 5: Silhouette update improves accuracy of the coarse face model. Each row shows an input image ((a) and (d)), the corresponding coarse face model with silhouette update ((b) and (e)), and the one without silhouette update ((c) and (f)). The updated silhouette is shown in red. The top row shows an example with $+30°$ yaw, and the bottom row with $+45°$ yaw.

The silhouette update improves the accuracy of the coarse face model for non-frontal images, as shown in Fig. 5 for two examples with $+30°$ and $+45°$ yaws: without the silhouette update, the resulting model will become wider due to erroneous correspondence with between the detected landmarks and the silhouette landmarks. When the yaw becomes larger, the detected 2D landmarks become less reliable, and the coarse face model becomes less accurate even with silhouette update. Our approach does not work well for images with very large poses (beyond $60°$ yaw) unless the invisible landmarks can be accurately detected. On the other hand, our pipeline can be combined with large-pose landmark detection algorithms to produce good results for such images. Some examples are shown in Fig. 13.

V. Medium Face Modeling

Although the coarse face model provides a good estimate of the overall shape, it may not capture some person-specific geometric details due to limited variation of the constructed data set (see Fig. 7). Thus we enhance the coarse face using smooth deformation that correspond to medium-scale geometric features, to improve the consistency between its shading and the input image. During this process we also estimate the lighting and the albedo. The enhanced face model and the lighting/albedo information will provide the prior knowledge required by the SFS reconstruction in the next section. In this paper, we convert color input images into grayscale ones for simplicity and efficiency. However, it is not difficult to extend the formulation to directly process color images.

Lighting and albedo estimation. To compute shading for our face mesh, we need the information about lighting and surface reflectance. Assuming Lambertian reflectance, we can approximate the grayscale level $s_{i,j}$ at a pixel (i,j) using second-order spherical harmonics [44]:

$$s_{i,j} = r_{i,j} \cdot \max(\boldsymbol{\xi}^T \mathbf{H}(\mathbf{n}_{i,j}), 0). \tag{5}$$

Here $r_{i,j}$ is the albedo at the pixel; $\mathbf{n}_{i,j}$ is the corresponding mesh normal, computed via

$$\mathbf{n}_{i,j} = \frac{(\mathbf{v}_2^{i,j} - \mathbf{v}_1^{i,j}) \times (\mathbf{v}_3^{i,j} - \mathbf{v}_1^{i,j})}{\|(\mathbf{v}_2^{i,j} - \mathbf{v}_1^{i,j}) \times (\mathbf{v}_3^{i,j} - \mathbf{v}_1^{i,j})\|_2}, \tag{6}$$

where $\mathbf{v}_1^{i,j}, \mathbf{v}_2^{i,j}, \mathbf{v}_3^{i,j}$ are the vertex coordinates for the mesh triangle that corresponds to pixel (i,j); $\mathbf{H}$ is a vector of second-order spherical harmonics

$$\mathbf{H}(\mathbf{n}) = [1, n_x, n_y, n_z, n_x n_y, n_x n_z, n_y n_z, n_x^2 - n_y^2, 3n_z^2 - 1]^T, \tag{7}$$

and $\boldsymbol{\xi}$ is a vector of harmonics coefficients. For more robust estimation, we follow [6] and parametrize the surface reflectance using a *Principal Component Analysis* (PCA) model:

$$r_{i,j} = \left(\boldsymbol{\Phi}_0 + \sum_{l=1}^{N_r} w_r^l \boldsymbol{\Phi}_l \right) \cdot \mathbf{c}_{i,j}, \tag{8}$$

where $[c_{i,j}^1, c_{i,j}^2, c_{i,j}^3] \in \mathbb{R}^3$ is the barycentric coordinate of the triangle corresponding to $r_{i,j}$, $[\boldsymbol{\Phi}_0, \boldsymbol{\Phi}_1, ..., \boldsymbol{\Phi}_{N_r}] \in \mathbb{R}^{N_v \times (N_r+1)}$ is a basis of vertex albedos with N_v being the number of vertices of the face mesh, $\mathbf{w}_r = (w_r^1, ..., w_r^{N_r}) \in \mathbb{R}^{N_r}$ is a vector for the albedo weights; $\mathbf{c}_{i,j} \in \mathbb{R}^{N_v}$ is a vector whose components for the three vertices of the triangle that contains pixel (i,j) are equal to the barycentric coordinates of the pixel within the triangle, and the components for other vertices are zero. Among the 199 principal albedo components derived from BFM2009, we choose N_r principal components with the largest variance as $\boldsymbol{\Phi}_1, ..., \boldsymbol{\Phi}_{N_r}$. We set $N_r = 100$ in our experiments. The lighting and albedo are then estimated by solving an optimization problem

$$\min_{\mathbf{r},\boldsymbol{\xi},\mathbf{d}} \quad \sum_{i,j} \left(r_{i,j} \boldsymbol{\xi}^T \mathbf{H}(\mathbf{n}_{i,j}) - I_{i,j} \right)^2 + \mu_1 \sum_{l=1}^{N_r} \left\| \frac{w_r^l}{\delta_{\mathrm{r}}^{(l)}} \right\|_2^2, \tag{9}$$

where vectors $\mathbf{r}, \mathbf{d}$ collect the values $\{r_{i,j}\}, \{d_{i,j}\}$, respectively; $I_{i,j}$ denotes the grayscale value at pixel (i,j) of the input image; $\{\delta_{\mathrm{r}}^{(l)}\}$ are the standard deviations corresponding to the principal directions; μ_1 is a user-specified positive weight. To optimize this problem, we first set $\mathbf{w}_r$ to zero and optimize the harmonics coefficients $\boldsymbol{\xi}$. Then we optimize the reflectance weights $\mathbf{w}_r$ while fixing $\boldsymbol{\xi}$. Both sub-problems reduce to solving a linear system. This process is iterated three times in our experiment.

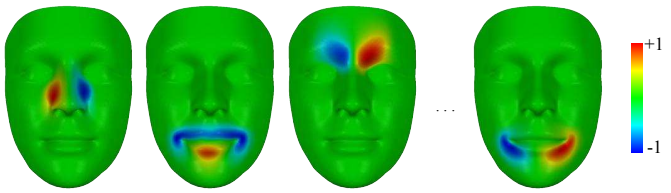


Figure 6: Some Laplacian eigenfunctions of local regions on the face mesh (displayed via color coding).

Facial detail enhancement. With an estimate of lighting and albedo, we can now enhance the coarse face mesh to reduce the discrepancy between the mesh shading and the input image. We apply a smooth 3D deformation field to the N_v vertices of the frontal face mesh to minimize the following discrepancy measure with respect to the vertex displacements $\mathbf{D} \in \mathbb{R}^{3 \times N_v}$:

$$E_{\text{shading}}(\mathbf{D}) = \sum_{i,j} \left(r_{i,j} \max(\boldsymbol{\xi}^T \mathbf{H}(\widetilde{\mathbf{n}}_{i,j}), 0) - I_{i,j} \right)^2, \tag{10}$$

where $\{\widetilde{\mathbf{n}}_{i,j}\}$ are the new mesh face normals. Specifically, since our final goal is to recover a depth field defined on the facial pixels in the given image, we sum over the pixels in Eq. (10). The correspondence between pixels and triangles are computed by the Z-buffer method [45]. However, this nonlinear least-squares problem can be very time-consuming to solve, due to the high resolution of the mesh. Therefore, we construct a low-dimensional subspace of smooth mesh deformations and solve the optimization problem within this subspace, which significantly reduces the number of variables. Specifically, if we measure the smoothness of a deformation field using the norm of its graph Laplacian with respect to the mesh, then the Laplacian eigenfunctions associated with small eigenvalues span a subspace of smooth deformations. Indeed, it is well known in 3D geometry processing that the Laplacian eigenvalues can be seen as the frequencies for the eigenfunctions, which indicate how rapidly each eigenfunction oscillates across the surface [46]. Thus by restricting the deformation to the subspace with small eigenvalues, we inhibit the enhancement of fine-scale geometric features, leaving them to the SFS reconstruction step in Sec VI. Since most facial variations are local, we select some local regions on the mesh, and perform Laplacian eigenanalysis on each region separately (see Fig. 6). The selected eigenfunctions are then combined to span a space of facial variations. Specifically, for the i-th selected region, we preprocess the frontal face mesh to construct its graph Laplacian matrix $\mathbf{K}^i \in \mathbb{R}^{N_v \times N_v}$ based on mesh connectivity, and add a large positive value to the j-th diagonal element if vertex j is outside the selected region. Then we perform eigendecomposition to obtain $k+1$ eigenvectors $\mathbf{e}_0^i, \mathbf{e}_1^i, \ldots, \mathbf{e}_k^i$ corresponding to the smallest eigenvalues $\lambda_0^i \leq \lambda_1^i \leq \ldots \leq \lambda_k^i$. Among them, $\mathbf{e}_0^i$ has a constant value inside the selected region,

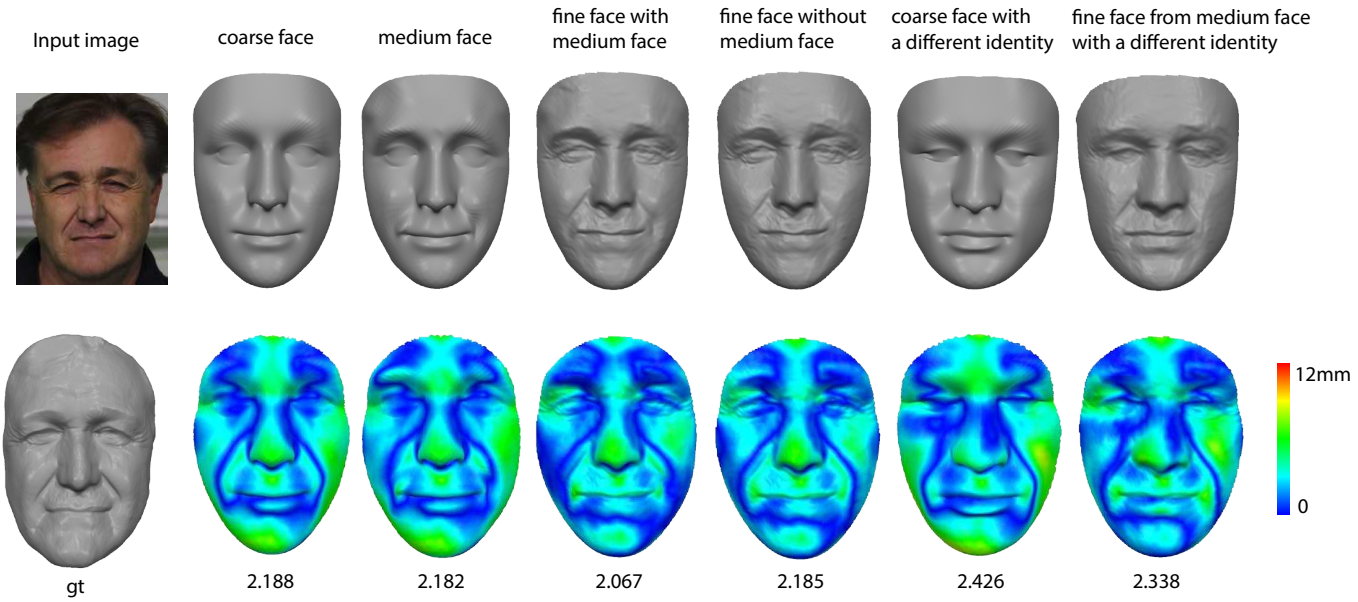


Figure 7: Quantitative results on the dataset [47]. The input image and its ground truth shape are shown in the first column. In the other columns, we show different face reconstructions and their corresponding error maps (according to Eq. (24)): the coarse face model, the medium face model, the fine reconstruction with and without medium face modeling, the coarse model with modified identity parameters, and the fine reconstruction with medium face modeling from the modified coarse face. In the bottom, we show the reconstruction error values.

representing a translation of the whole region [46]. Since it does not represent variation within the region, we discard $\mathbf{e}_0^i$ to get k eigenvectors $\mathbf{E}^i = [\mathbf{e}_1^i, \ldots, \mathbf{e}_k^i]$. Combing all the eigenvectors to span the x-, y-, and z-coordinates of the vertex displacement vectors, we represent the deformation field as

$$\mathbf{D} = (\mathbf{E}\boldsymbol{\eta})^T, \tag{11}$$

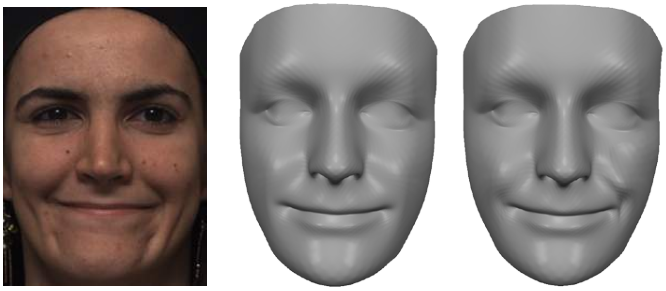
where $\mathbf{E} = [\mathbf{E}^1, \ldots, \mathbf{E}^{N_e}] \in \mathbb{R}^{N_v \times (k \cdot N_e)}$ stacks the basis vectors, and $\boldsymbol{\eta} = [\lambda_1^1, \ldots, \lambda_k^1, \ldots, \lambda_1^{N_e}, \ldots, \lambda_k^{N_e}]^T \in \mathbb{R}^{(k \cdot N_e) \times 3}$ collects their linear combination coefficients. Then the deformation is determined by solving the following optimization problem about $\boldsymbol{\eta}$:

$$\min_{\boldsymbol{\eta}} \quad E_{\text{shading}}(\mathbf{D}) + \mu_2 \sum_{i=1}^{N_e} \sum_{j=1}^{k} \left\| \frac{\boldsymbol{\eta}_j^i}{\lambda_j^i} \right\|_2^2. \tag{12}$$

Here the second term prevents large deformations, with more penalty on basis vectors of lower frequencies; μ_2 is a user-specified weight. Our formulation is designed to induce more enhancement for finer geometric features, since the coarse face already provides a good estimate of the overall shape. In our experiments, we set $k = 5$ and $N_e = 9$, which means we select nine local regions and the first five eigenfunctions of the corresponding Laplacian matrix for each region. These local regions are manually selected in a heuristic way. More specifically, given the mean face shape, we first compute the vertex displacements from its neutral expression to each of the other 46 expressions, and manually select nine regions with the largest variation as the local regions.

As the number of variables are significantly reduced in (12), this nonlinear least-squares problem can be solved efficiently using the Levenberg-Marquardt algorithm [48]. We then apply the optimized deformation field to the frontal face mesh, and update the correspondence between image pixels and mesh triangles. With the new correspondences, we solve the optimization problems (9) and (12) again to further improve the lighting/albedo estimate and the face model. This process is iterated twice in our experiments.

Medium face modeling can improve the accuracy of medium-scale facial features such as those around the laugh lines, as shown in Figs. 7 and Figs. 8. Fig. 7 compares the fine face reconstruction results with and without medium face modeling. We can see that the use of medium face leads to more accurate results numerically and visually. Indeed, eigendeomposition of the Laplacian matrix corresponds to Fourier analysis of geometric signals defined on the mesh surface [46], thus our use of basisvectors is similar to approximating the displacement from the coarse face to the ground truth shape in each local region using its Fourier components of lowest frequencies, which is a classical signal processing technique. On the other hand, our approach cannot reconstruct facial features whose frequency bands have limited overlap with those corresponding to the chosen basisvectors. One example is shown in Fig. 8, where the dimples cannot be reconstructed. Finally, as the medium face modeling is applied on local regions, it cannot reduce reconstruction errors of global scales. As an example, in Fig. 7 we alter the identity parameters to generate a different coarse face model, and apply medium and fine face modeling. We can see that although medium and fine face modeling help to introduce more details, they cannot change the overall face shape.



(a) (b) (c)

Figure 8: An input image with smile expression (a), and its coarse (b) and medium (c) face models. The use of Laplacian eigenvectors improves the accuracy of features around the laugh lines, but cannot reconstruct the dimples as the eigenvectors provide limited cover of their frequency band.

VI. Fine Face Modeling

As the final step in our pipeline, we reconstruct a face model with fine geometric details, represented as a height field surface over the face region Ω of the input image. Using the medium face model and the lighting/albedo information computed in Sec. V, we first compute a refined normal map over Ω, to capture the details from the input image. This normal map is then integrated to recover a height field surface for the final face shape.

Overall approach. Specifically, the normal map is defined using a unit vector $\mathbf{n}'_{i,j} \in \mathbb{R}^3$ for each pixel $(i,j) \in \Omega$. Noting that each face pixel corresponds to a normal vector facing towards the camera [8], we represent $\mathbf{n}'_{i,j}$ using two variables $p_{i,j}, q_{i,j}$ as

$$\mathbf{n}'_{i,j} = \frac{(p_{i,j}, q_{i,j}, -1)}{\sqrt{p_{i,j}^2 + q_{i,j}^2 + 1}}. \tag{13}$$

The values $\{p_{i,j}\}$, $\{q_{i,j}\}$ are computed by solving an optimization problem that will be explained later. The final height-field face model, represented using a depth value $z_{i,j}$ per pixel, is then determined so that the height field normals are as close as possible to the normal map. We note that the height field normal $\widehat{\mathbf{n}}_{i,j}$ at pixel (i,j) can be computed using three points $\mathbf{h}_{i,j} = (i, j, z_{i,j})$, $\mathbf{h}_{i,j+1} = (i, j+1, z_{i,j+1})$, $\mathbf{h}_{i+1,j} = (i+1, j, z_{i+1,j})$ on the height field surface via

$$\begin{aligned} \widehat{\mathbf{n}}_{i,j} &= \frac{(\mathbf{h}_{i,j+1} - \mathbf{h}_{i,j}) \times (\mathbf{h}_{i+1,j} - \mathbf{h}_{i,j})}{\|(\mathbf{h}_{i,j+1} - \mathbf{h}_{i,j}) \times (\mathbf{h}_{i+1,j} - \mathbf{h}_{i,j})\|_2} \\ &= \frac{(z_{i+1,j} - z_{i,j}, z_{i,j+1} - z_{i,j}, -1)}{\sqrt{(z_{i+1,j} - z_{i,j})^2 + (z_{i,j+1} - z_{i,j})^2 + 1}}. \end{aligned} \tag{14}$$

Comparing this with Eq. (13) shows that for the height field normal to be consistent with the normal map, we should have

$$z_{i+1,j} - z_{i,j} = p_{i,j}, \quad z_{i,j+1} - z_{i,j} = q_{i,j} \tag{15}$$

for every pixel. As these conditions only determine $\{z_{i,j}\}$ up to an additional constant, we compute $\{z_{i,j}\}$ as the minimum-norm solution to a linear least-squares problem

$$\min_{\{z_{i,j}\}} \sum_{(i,j)} (z_{i+1,j} - z_{i,j} - p_{i,j})^2 + (z_{i,j+1} - z_{i,j} - q_{i,j})^2. \tag{16}$$

Normal map optimization. For high-quality results, we enforce certain desirable properties of the computed normal map $\mathbf{n}'_{i,j}$ by minimizing an energy that corresponds to these properties. First of all, the normal map should capture fine-scale details from the input image. Using the lighting and albedo parameters obtained during the computation of the medium face, we can evaluate the pixel intensity values from the normal map according to Eq. (5), and require them to be close to the input image. However, such direct approach can suffer from the inaccuracy of spherical harmonics in complex lighting conditions such as cast shadows, which can lead to unsatisfactory results. Instead, we aim at minimizing the difference in intensity gradients, between the input image and the shading from the normal map. This difference can be measured using the following energy

$$E_{\text{grad}} = \sum_{(i,j)} \left\| \begin{bmatrix} s'_{i+1,j} - s'_{i,j} \\ s'_{i,j+1} - s'_{i,j} \end{bmatrix} - \begin{bmatrix} I_{i+1,j} - I_{i,j} \\ I_{i,j+1} - I_{i,j} \end{bmatrix} \right\|_2^2, \tag{17}$$

where $\{I_{i,j}\}$ are intensity values from the input image, and

$$s'_{i,j} = r_{i,j} \cdot \max(\boldsymbol{\xi}^T \mathbf{H}(\mathbf{n}'_{i,j}), 0) \tag{18}$$

are shading intensities for the normal map according to Eq. (5), using the optimized albedo $\{r_{i,j}\}$ and spherical harmonic coefficients $\boldsymbol{\xi}$ from Sec. V. Minimizing the difference in gradients instead of intensities helps to attenuate the influence from illumination noises such as cast shadows, while preserving the features from the input image. Another benefit is that its optimality condition is a higher-order PDE that results in smoother solution and reduces unnatural sharp features [49]. One example is shown in Fig. 9, where the formulation with gradient difference reduces the sharp creases around the nose and the mouth. (see Fig. 9).

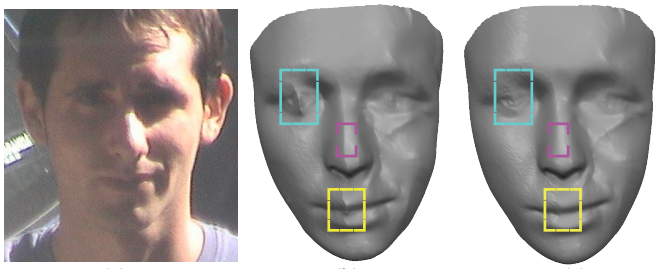
Optimizing E_{grad} alone is not sufficient for good results, since the problem is under-constrained. Thus we introduce two additional regularization terms for the normal map. First we note that the medium face model from Sec. V provides good approximation of the final shape. Thus we introduce the following energy to penalize the deviation between normal map and the normals from the medium face

$$E_{\text{close}} = \sum_{(i,j)} \|\mathbf{n}'_{i,j} - \mathbf{n}_{i,j}\|_2^2, \tag{19}$$

where $\mathbf{n}_{i,j}$ is computed from the medium face mesh according to Eq. (6). In addition, we enforce smoothness of the normal map using an energy that penalizes its gradient

$$E_{\text{smooth}} = \sum_{(i,j)} \|\mathbf{n}'_{i+1,j} - \mathbf{n}'_{i,j}\|_2^2 + \|\mathbf{n}'_{i,j+1} - \mathbf{n}'_{i,j}\|_2^2. \tag{20}$$

Finally, we need to ensure the normal map is *integrable*, i.e., given the normal map there exists a height field surface such that conditions (15) are satisfied. Note that if (15) are satisfied,



(a) (b) (c)

Figure 9: An input image with cast shadow and noise (a), and its reconstruction results by minimizing the intensity difference (b) and the gradient difference (c), respectively. Compared with intensity difference minimization, the formulation with gradient difference produces a smoother result and reduces unnatural sharp creases at the eye, the nose, and the mouth (highlighted with rectangles).

then $p_{i,j}$ and $q_{i,j}$ are the increments of function z along the grid directions. Moreover, the total increment of z along the close path that connects pixels $(i,j),(i+1,j),(i+1,j+1),(i,j+1)$ should be zero, which results in the condition

$$p_{i,j} + q_{i+1,j} - p_{i,j+1} - q_{i,j} = 0. \tag{21}$$

For the normal map to be integrable, this condition should be satisfied at each pixel. Indeed, with condition (15) we can interpret p and q as partial derivatives $\frac{\partial z}{\partial u}, \frac{\partial z}{\partial v}$ where u, v are the grid directions; then condition (21) corresponds to $\frac{\partial p}{\partial v} = \frac{\partial q}{\partial u}$, which is the condition for (p,q) to be a gradient field. We can then enforce the integrability condition using an energy

$$E_{\text{int}} = \sum_{(i,j)} (p_{i,j} + q_{i+1,j} - p_{i,j+1} - q_{i,j})^2. \tag{22}$$

Combining the above energies, we derive an optimization problem for computing the desirable normal map

$$\min_{\mathbf{p},\mathbf{q}} \quad E_{\text{grad}} + \omega_1 E_{\text{close}} + \omega_2 E_{\text{smooth}} + \omega_3 E_{\text{int}}, \tag{23}$$

where the optimization variables $\mathbf{p}, \mathbf{q}$ are the values $\{p_{i,j}\}, \{q_{i,j}\}$, and $\omega_1, \omega_2, \omega_3$ are user-specified weights. This nonlinear least-squares problem is again solved using the Levenberg-Marquardt algorithm.

Fig. 7 shows a fine face model reconstructed using our method. Compared with the medium face model, it captures more geometric details and reduces the reconstruction error. Besides, it can be observed from the reconstruction results in last two columns that the initial coarse face model has a large influence on reconstruction accuracy.

VII. Experiments

This section presents experimental results, and compares our method with some existing approaches.

Experimental setup. To verify the effectiveness of our method, we tested it using the data set from the Bosphorus database [50]. This database provides structured-light scanned 3D face point clouds for 105 subjects, as well as their corresponding single-view 2D face photographs. For each subject, the database provides point clouds and images for different facial expressions and head poses. We ran our algorithm on the 2D images, and used the corresponding point clouds as ground truth to evaluate the reconstruction error. 55 subjects with low noises in their point clouds were chosen for testing. The reconstructed face is aligned with its corresponding ground truth face using iterative closest point (ICP) method [51]. After alignment, we crop the face model at a radius of 85mm around the tip of the nose, and then compute the 3D Root Mean Square Error (3DRMSE):

$$\sqrt{\sum_i (\mathbf{X} - \mathbf{X}^*)^2 / N}, \tag{24}$$

where $\mathbf{X}$ is the reconstructed face, $\mathbf{X}^*$ is the grund truth, N is the number of vertices of the cropped frontal reconstructed face. We also computed the mean and standard deviation of all these errors.

Table I: The mean and standard variation of our reconstructions for each pose and expression.

Pose	Yaw $+10°$	Yaw $+20°$	Yaw $+30°$
3DRMSE	1.73 ± 0.33	1.51 ± 0.24	1.44 ± 0.32
Expression	happy	surprise	disgust
3DRMSE	1.71 ± 0.34	2.05 ± 0.49	1.98 ± 0.42

Our algorithm is implemented in C++ and is tested on a PC with an Intel Core i7-4710MQ 2.50 GHz CPU and 7.5 GB RAM. The weights in optimization problems (2), (9), (12), (23) are set as follows: $\gamma_1 = \gamma_2 = 1.5 \times 10^3$;$\mu_1 = 5$;$\mu_2 = 20$; $\omega_1 = 10, \omega_2 = 10, \omega_3 = 1$. The nonlinear least-squares problems are solved using the Ceres solver [52], with all derivatives evaluated using automatic differentiation. To speed up the algorithm, we downsample the high-resolution 2D images from the database to 30% of their original dimensions before running our algorithm. The down-sampled images have about 400×500 pixels, for which the coarse, medium, and fine face construction steps take about 1 second, 2 minutes, and 1 minute respectively using our non-optimized implementation.

Frontal and neutral faces. We first tested our method on facial images of frontal pose and neutral expression, from 55 subjects in the Bosphorus database. For comparison we also ran the face reconstruction method from [3], which is based on a 3DMM built from BFM2009 and FaceWarehouse. Fig. 10 presents the reconstruction results of six subjects using our method and [3], and compares them with the ground truth faces. Thanks to the enhancement in the medium face step and the SFS recovery in the fine face step, our approach can not only obtain a more realistic global facial shape, but also accurately capture the person-specific geometric details such as wrinkles. Fig. 10 also shows the 3DRMSE for our results and the results using [3]. The mean and standard variation of 3DRMSE is 1.97 ± 0.35 for the results by method [3], and 1.56 ± 0.24 for the results by our method. It can be seen that the mean error from our results are consistently lower than those from the method of [3].

Near-frontal poses and expressions. We also tested our method on face images with near-frontal poses and expressions. First, for each of the 55 subjects, we applied our method on

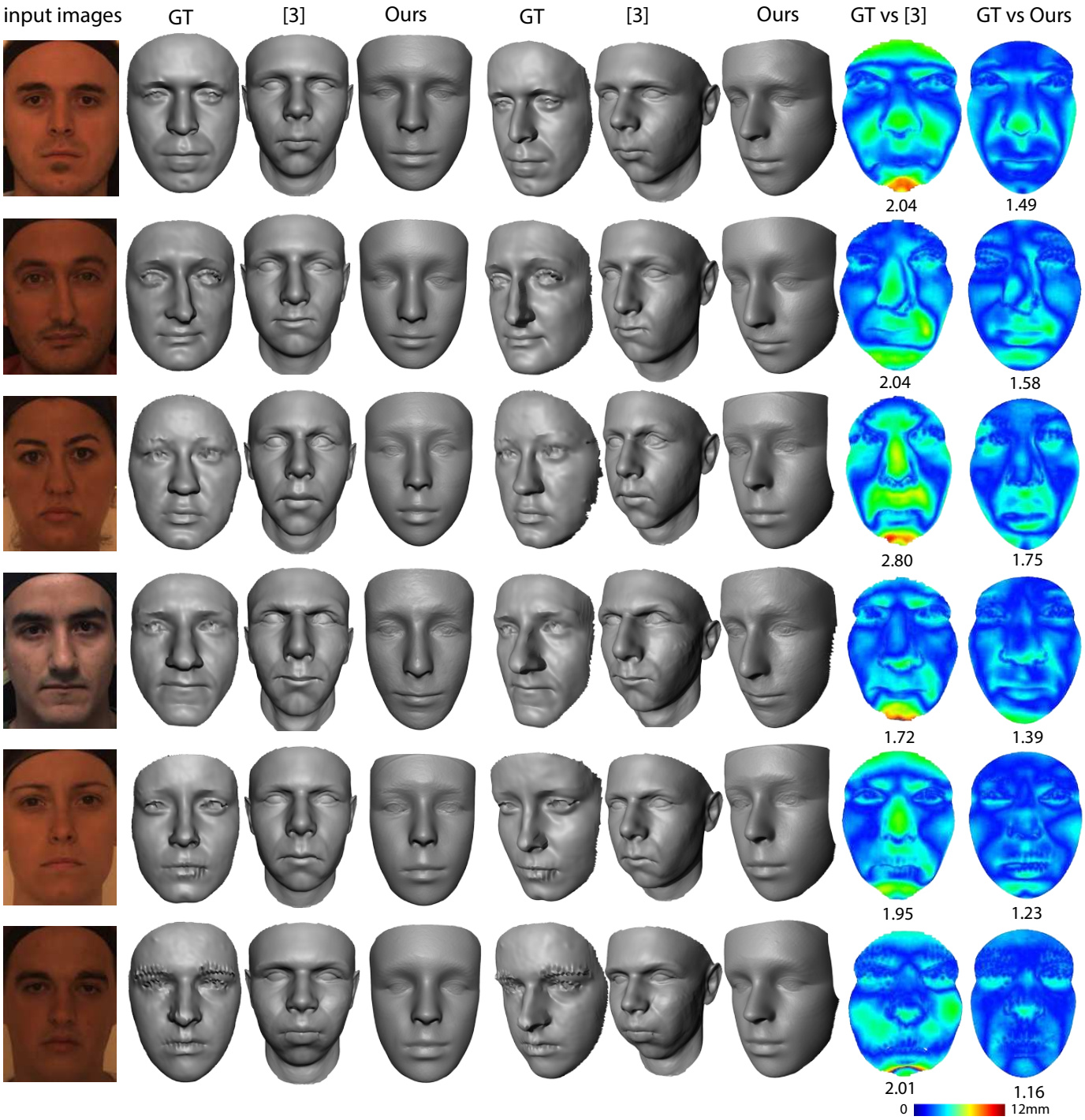


Figure 10: Facial reconstruction from images of frontal pose and neutral expression. For each input image, we show the ground truth (GT) as well as the results using out method and the method from [3], each in two viewpoints. We also show the error maps (according to Eq. (24)) for the two methods, together with their 3DRMSE.

their images of neutral expression with three types of poses: Yaw $+10°$, $+20°$, and $+30°$. Then, we tested our approach on frontal faces with three non-neutral expressions: happy, surprise, and disgust. Among the 55 subjects, there are 25 of them with all three expressions present. We apply our method on these 25 subjects, and Table I shows the mean and standard deviation of 3DRMSE for each pose and expression. We can observe that the reconstruction results by our method are consistent for different poses and expressions, and the reconstruction errors are small. This is verified in Fig. 11, where we show the reconstruction results of four subjects under different poses and expressions.

Furthermore, using landmark detection methods designed for facial images with large pose (e.g., $90°$), our approach can also reconstruct the 3D model well for such images. Two examples are shown in Fig. 13, where the landmarks are detected using

Figure 11: Face reconstructions of four subjects from images of frontal pose with different expressions (happy, surprise, disgust), and of different poses (Yaw $+10°$, $+20°$, $+30°$) with neutral expression. For each input image, we show the reconstructed face mesh as well as its textured rendering.

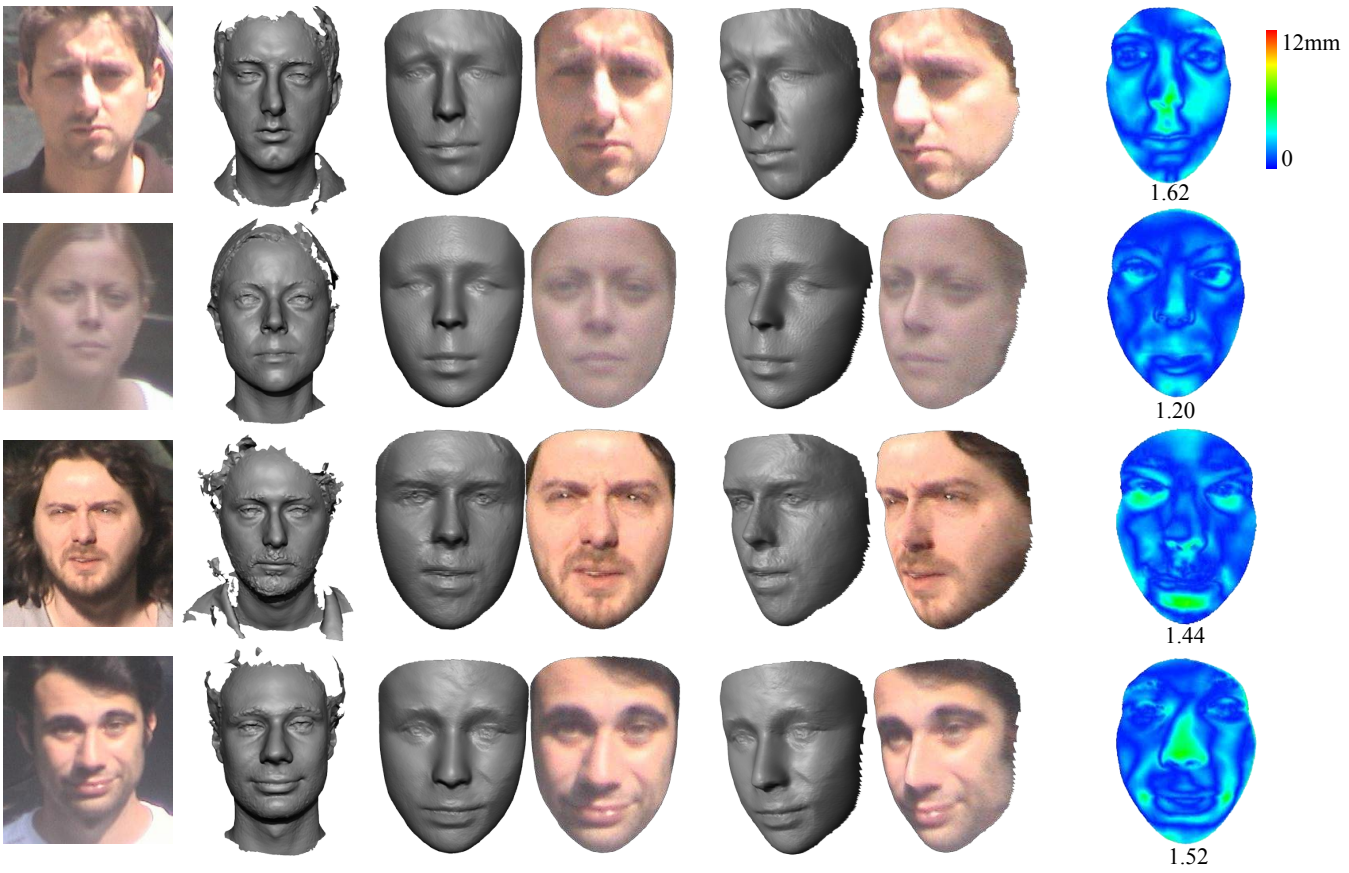


Figure 12: Face reconstructions of four subjects from the MICC dataset [53] using our method. We show from left to right the input image, the ground truth, our reconstruction result (with texture) in two view points, and error map (according to Eq. (24)).

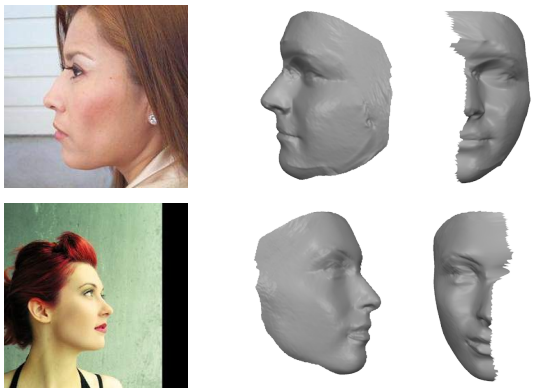


Figure 13: Face reconstructions of face images with very large pose using our method. We show from left to right the input image, and the reconstruction result from two viewpoints.

the method from [54].

Unconstrained facial images. To demonstrate the robustness of our approach on general unconstrained facial images, we compare our method with the *structure from motion* (SFM) method [55] and the learning-based method [56] using the MICC dataset [53]. The MICC dataset contains 53 video sequences of varying resolution, conditions and zoom levels for each subject, which is recorded in controlled, less controlled or uncontrolled environment. There is a structured-light scanning

Table II: Quantitative results on the MICC dataset [53]. The mean and standard variation of 3DRMSE, the runtimes.

Approach	3DRMSE	run time
SFM [55]	1.92 ± 0.39	CPU 1min 13s
CNN-based methods [56]	1.53 ± 0.29	GPU 0.088s
Ours	1.75 ± 0.29	CPU 3min

for each subject as the ground truth, and the reconstruction errors of the reconstruction results are computed following the way described in the above. For each subject, we select the most frontal face image from the corresponding outdoor video and reconstruct the 3D face model by setting it as input. Table II shows that our reconstruction error is close to [56] and lower than [55]. With the prior of reliable medium face and SFS recovery, our approach can also have good estimations on unconstrained images. Fig. 12 presents the reconstruction results of four subjects using our method.

We also compared our method with the SFS approach of [8] on more general unconstrained facial images. Since there are no ground truth shapes for these images, we only compared them visually. For reliable comparison, we directly ran our algorithm on the example images provided in [8]. Fig. 14 presents the comparison results, showing both the reconstructed face geometry and its textured display. We can see that our approach produced more accurate reconstruction of the overall shape, and recovered more geometrical details such as winkles

Figure 14: Face reconstructions from unconstrained images, using the method from [8] and our method.

and teeth. Although both methods perform SFS reconstruction, there is major difference on how the shape and illumination priors are derived. In [8] a reference face model is utilized as the shape prior to estimate illumination and initialize photometric normals; as the reference face model is not adapted to the target face shape, this can lead to unsatisfactory results. In comparison, with our method the medium face model is optimized to provide reliable estimates of the target shape and illumination, which enables more accurate reconstruction.

VIII. Discussion and Conclusion

The main limitation of our method is that its performance for a given image depends on how well the overall face shape is covered by our constructed face model. This is because medium and fine face modeling have little effect on the coarse face shape; thus in order to achieve good results, the coarse face model needs to be close enough to the ground-truth overall shape, which can be achieved if the ground-truth face is close to the space spanned by our linear face model. By combining FaceWarehouse and BFM2009 to construct the face model, our approach achieves good results on a large number of images. But for faces with large deviation from both FaceWarehouse and BFM2009, our method may not work well. One potential future work is to improve the face model by incorporating a larger variety of face datasets.

Since we compute pixel values by multiplying albedo with lighting, there is an inherent ambiguity in determining albedo and lighting from given pixel values. Our approach alleviates the problem by using PCA albedo and second-order spherical harmonics lighting, but it does not fully resolve the ambiguity. Nevertheless, as we only intend to recover face geometry, such approach is sufficient for achieving good results.

In this paper, we present a coarse-to-fine method to reconstruct a high-quality 3D face model from a single image. Our approach uses a bilinear face model and local corrective deformation fields to obtain a reliable initial face shape with large- and medium-scale features, which enables robust shape-from-shading reconstruction of fine facial details. The experiments demonstrate that our method can accurately reconstruct 3D face models from images with different poses and expressions, and recover the fine-scale geometrical details such as wrinkles and teeth. Our approach combines the benefits of low-dimensional face models and shape-from-shading, enabling more accurate and robust reconstruction.

Acknowledgments

We would like to thank the reviewers for their time spent on reviewing our manuscript and their insightful comments helping us improving the article. This work was supported by the National Key R&D Program of China (No. 2016YFC0800501), the National Natural Science Foundation of China (No. 61672481, No. 61672482 and No. 11626253), the Youth Innovation Promotion Association of CAS, and the One Hundred Talent Project of the Chinese Academy of Sciences.

References

[1] G. Stylianou and A. Lanitis, "Image based 3D face reconstruction: A survey," *International Journal of Image and Graphics*, vol. 9, no. 2, pp. 217–250, 2009.

[2] V. Blanz and T. Vetter, "Face recognition based on fitting a 3d morphable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1063–1074, 2003.

[3] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 787–796.

[4] C. Cao, Q. Hou, and K. Zhou, "Displaced dynamic expression regression for real-time facial tracking and animation," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 43:1–43:10, 2014.

[5] J. Thies, M. Zollhfer, M. Stamminger, C. Theobalt, and M. Niener, "Face2face: Real-time face capture and reenactment of rgb videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2387–2395.

[6] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH*, 1999, pp. 187–194.

[7] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, "Shape-from-shading: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 8, pp. 690–706, 1999.

[8] I. Kemelmacher-Shlizerman and R. Basri, "3d face reconstruction from a single image using a single reference face shape." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 394–405, 2011.

[9] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 3, pp. 413–425, 2014.

[10] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and 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*, 2009, pp. 296–301.

[11] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *Journal of the Optical Society of America A*, vol. 4, no. 3, pp. 519–524, 1987.

[12] M. Meytlis and L. Sirovich, "On the dimensionality of face space," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 7, pp. 1262–1267, 2007.

[13] S. Romdhani and T. Vetter, "Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 986–993.

[14] M. Keller, R. Knothe, and T. Vetter, "3d reconstruction of human faces from occluding contours," *Computer Vision/Computer Graphics Collaboration Techniques*, pp. 261–273, 2007.

[15] A. Bas, W. A. Smith, T. Bolkart, and 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.

[16] P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W. J. Christmas, M. Ratsch, and J. 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.

[17] V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel, "Exchanging faces in images," *Computer Graphics Forum*, vol. 23, no. 3, pp. 669–676, 2004.

[18] K. Scherbaum, T. Ritschel, M. Hullin, T. Thormhlen, V. Blanz, and H.-P. Seidel, "Computer-suggested facial makeup," *Computer Graphics Forum*, vol. 30, no. 2, pp. 485–492, 2011.

[19] D. Vlasic, M. Brand, H. Pfister, and J. Popovic, "Face transfer with multilinear models." *ACM Trans. Graph.*, vol. 24, no. 3, pp. 426–433, 2005.

[20] J. P. Lewis, K. Anjyo, T. Rhee, M. Zhang, F. Pighin, and Z. Deng, "Practice and theory of blendshape facial models," in *Eurographics 2014 - State of the Art Reports*, S. Lefebvre and M. Spagnuolo, Eds. The Eurographics Association, 2014.

[21] T. Weise, H. Li, L. Van Gool, and M. Pauly, "Face/off: Live facial puppetry," in *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. ACM, 2009, pp. 7–16.

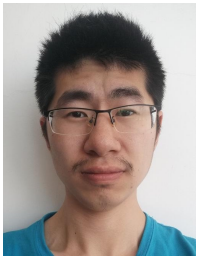
[22] T. Weise, S. Bouaziz, H. Li, and M. Pauly, "Realtime performance-based facial animation," *ACM Trans. Graph.*, vol. 30, no. 4, p. 77, 2011.

[23] S. Bouaziz, Y. Wang, and M. Pauly, "Online modeling for realtime facial animation." *ACM Trans. Graph.*, vol. 32, no. 4, pp. 40:1–40:10, 2013.

[24] H. Li, J. Yu, Y. Ye, and C. Bregler, "Realtime facial animation with on-the-fly correctives," *ACM Transactions on Graphics (Proceedings SIGGRAPH 2013)*, vol. 32, no. 4, July 2013.

[25] P. Hsieh, C. Ma, J. Yu, and H. Li, "Unconstrained realtime facial performance capture," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1675–1683.

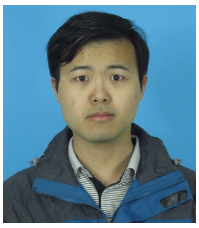
[26] C. Cao, Y. Weng, S. Lin, and K. Zhou, "3d shape regression for real-time facial animation." *ACM Trans. Graph.*, vol. 32, no. 4, 2013.
[27] C. Cao, D. Bradley, K. Zhou, and T. Beeler, "Real-time high-fidelity facial performance capture," *ACM Trans. Graph.*, vol. 34, no. 4, 2015.
[28] R. Gross, "Face databases," in *Handbook of Face Recognition*. Springer New York, 2005, pp. 301–327.
[29] J.-D. Durou, M. Falcone, and M. Sagona, "Numerical methods for shape-from-shading: A new survey with benchmarks," *Computer Vision and Image Understanding*, vol. 109, no. 1, pp. 22 – 43, 2008.
[30] E. Prados and O. Faugeras, "Shape from shading," in *Handbook of Mathematical Models in Computer Vision*, N. Paragios, Y. Chen, and O. Faugeras, Eds. Springer US, 2006, pp. 375–388.
[31] I. Shimshoni, Y. Moses, and M. Lindenbaum, "Shape reconstruction of 3d bilaterally symmetric surfaces," *International Journal of Computer Vision*, vol. 39, no. 2, pp. 97–110, 2000.
[32] W. Y. Zhao and R. Chellappa, "Illumination-insensitive face recognition using symmetric shape-from-shading," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2000, pp. 286–293.
[33] Zhao, Wen Yi and Chellappa, Rama, "Symmetric shape-from-shading using self-ratio image," *International Journal of Computer Vision*, vol. 45, no. 1, pp. 55–75, 2001.
[34] J. J. Atick, P. A. Griffin, and A. N. Redlich, "Statistical approach to shape from shading: Reconstruction of three-dimensional face surfaces from single two-dimensional images," *Neural Computation*, vol. 8, pp. 1321–1340, 1996.
[35] R. Dovgard and R. Basri, "Statistical symmetric shape from shading for 3d structure recovery of faces," in *ECCV*, T. Pajdla and J. Matas, Eds., 2004, pp. 99–113.
[36] C. Wu, B. Wilburn, Y. Matsushita, and C. Theobalt, "High-quality shape from multi-view stereo and shading under general illumination," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 969–976.
[37] A. E. Ichim, S. Bouaziz, and M. Pauly, "Dynamic 3d avatar creation from hand-held video input." *ACM Trans. Graph.*, vol. 34, no. 4, 2015.
[38] I. Kemelmacher-Shlizerman and S. M. Seitz, "Face reconstruction in the wild," in *International Conference on Computer Vision*, 2011.
[39] J. Roth, Y. Tong, and X. Liu, "Unconstrained 3d face reconstruction," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
[40] J. Roth, Y. Tong, and X. Liu, "Adaptive 3d face reconstruction from unconstrained photo collections," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4197–4206.
[41] R. W. Sumner and J. Popović, "Deformation transfer for triangle meshes," *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3, 2004.
[42] L. De Lathauwer, *Signal processing based on multilinear algebra*. Katholieke Universiteit Leuven, 1997.
[43] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *ECCV*, 2014, pp. 109–122.
[44] D. Frolova, D. Simakov, and R. Basri, "Accuracy of spherical harmonic approximations for images of lambertian objects under far and near lighting," in *ECCV*. Springer, 2004, pp. 574–587.
[45] W. Straßer, "Schnelle kurven-und flächendarstellung auf grafischen sichtgeräten," Ph.D. dissertation, 1974.
[46] H. Zhang, O. Van Kaick, and R. Dyer, "Spectral mesh processing," *Computer Graphics Forum*, vol. 29, no. 6, pp. 1865–1894, 2010.
[47] L. Valgaerts, C. Wu, A. Bruhn, H. Seidel, and C. Theobalt, "Lightweight binocular facial performance capture under uncontrolled lighting," *ACM Trans. Graph.*, vol. 31, no. 6, p. 187, 2012.
[48] K. Madsen, H. B. Nielsen, and O. Tingleff, "Methods for non-linear least squares problems," Informatics and Mathematical Modelling, Technical University of Denmark, 2004, 2nd edition.
[49] M. Lysaker, A. Lundervold, and X.-C. Tai, "Noise removal using fourth-order partial differential equation with applications to medical magnetic resonance images in space and time," *IEEE Transactions on image processing*, vol. 12, no. 12, pp. 1579–1590, 2003.
[50] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3d face analysis," in *European Workshop on Biometrics and Identity Management*, 2008.
[51] S. Rusinkiewicz and M. Levoy, "Efficient variants of the ICP algorithm," in *3rd International Conference on 3D Digital Imaging and Modeling*, 2001, pp. 145–152.
[52] S. Agarwal, K. Mierle, and Others, "Ceres solver," http://ceres-solver.org.
[53] A. D. Bagdanov, A. Del Bimbo, and 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.
[54] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 146–155.
[55] M. Hernandez, T. Hassner, J. Choi, and G. Medioni, "Accurate 3d face reconstruction via prior constrained structure from motion," *Computers & Graphics*, 2017.
[56] A. T. Tran, T. Hassner, I. Masi, and G. G. Medioni, "Regressing robust and discriminative 3d morphable models with a very deep neural network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1493–1502.



Luo Jiang is currently working towards the PhD degree at the University of Science and Technology of China. He obtained his bachelor degree in 2013 from the Huazhong University of Science and Technology, China. His research interests include computer graphics, image processing and deep learning.

Juyong Zhang is an associate professor in the School of Mathematical Sciences at University of Science and Technology of China. He received the BS degree from the University of Science and Technology of China in 2006, and the PhD degree from Nanyang Technological University, Singapore. His research interests include computer graphics, computer vision, and numerical optimization. He is an associate editor of The Visual Computer.

Bailin Deng is a lecturer in the School of Computer Science and Informatics at Cardiff University. He received the BEng degree in computer software (2005) and the MSc degree in computer science (2008) from Tsinghua University (China), and the PhD degree in technical mathematics from Vienna University of Technology (Austria). His research interests include geometry processing, numerical optimization, computational design, and digital fabrication. He is a member of the IEEE.



Hao Li received the BSc degree in 2011 from the University of Science and Technology of China. His research interests include computer graphics and image processing.

Ligang Liu is a Professor at the School of Mathematical Sciences, University of Science and Technology of China. His research interests include digital geometric processing, computer graphics, and image processing. He serves as the associated editors for journals of IEEE Transactions on Visualization and Computer Graphics, IEEE Computer Graphics and Applications, Computer Graphics Forum, Computer Aided Geometric Design, and The Visual Computer. He served as the conference co-chair of GMP 2017 and the program co-chairs of GMP 2018, CAD/Graphics 2017, CVM 2016, SGP 2015, and SPM 2014. His research works could be found at his research website: http://staff.ustc.edu.cn/lgliu.