Dual Neural Networks Coupling Data Regression with Explicit Priors for Monocular 3D Face Reconstruction

Xin Fan *Senior Member, IEEE,* Shichao Cheng, Kang Huyan, Minjun Hou, Risheng Liu, *Member, IEEE,* and Zhongxuan Luo

***Abstract*—We address the challenging issue of reconstructing a 3D face from one single image under various expressions and illuminations, which is widely applied in multimedia tasks. Methods built upon classical parametric morphable models (3DMMs) gain success on reconstructing the global geometry of a 3D face, but fail to precisely characterize local facial details. Recently, deep neural networks (DNN) have been applied to the reconstruction that directly predicts depth maps, showing compelling performance on detail recovery. Unfortunately, their reconstruction is prone to structural distortions owing to the lack of explicit prior constraints. In this paper, we propose dual neural networks that optimize one energy coupling data fitting with local explicit geometric prior. Specifically, we build one residual network upon traditional convolution layers in order to directly predict 3D structures by fitting an input image. Meanwhile, we devise a novel architecture stacking shallow networks to refine 3D clouds with geometric priors given by Markov random fields (MRFs). Quantitative evaluations demonstrate the superior performance of the dual networks over either end-to-end DNNs or parametric models. Comparisons with the state-of-the-art also show competitive reconstruction quality on various conditions.**

***Index Terms*—3D face reconstruction, Deep optimization, Residual neural networks, Markov random fields**

I. Introduction

RECONSTRUCTING the 3D geometry structure from a single facial image has gained increasing attention for its benefits to multimedia and computer vision tasks. It is especially helpful for various face analysis applications, such as face recognition [1], face retrieval [2], and animated avatars [3]. Rich textures and features extracted from 3D faces can effectively characterize the essential facial structure, improving the performance of face skin detection [4], recognition and retrieval [5], [6]. Additionally, 3D face reconstruction can be regarded as a special depth estimation task [7], and thus inspire the depth estimation in generic scenarios.

A face image is the projection of a 3D human face onto a 2D imaging plane. Given a point q in a 2D image, the point q can be the projection of any 3D points on the line passing through q and the focal center of the camera lens [8]. This ambiguity of projection produces tremendous possible 3D clouds corresponding an image with millions of 2D points, leading to a severely ill-posed problem when recovering a 3D face from one facial image. Various approaches encoding explicit or implicit facial priors have been introduced in order to resolve this ill-posed problem.

Statistical parametric models [9], [10] have become the most popular approach to this challenging issue in the past decade. The common strategy for these parametric methods is to fit the model to a 2D input image by minimizing the error energy between the input and the 2D projection from the 3D model. Recently, researchers resort to regression techniques to accelerate the time consuming iterative optimization process for the model parameters [11], [12], [13]. The regression only employs sparse 2D geometric landmarks from facial images, but totally ignores the abundant textural details available in 2D images. Consequently, these methods can only generate rough 3D geometry without fine facial structures, *e.g.*, those on the nose and cheek, whose rich information is abandoned by 2D landmarks. Deep networks [14] and unsupervised training techniques [15] are also used for 3DMM parameters regression, considering both geometries and appearances of facial images. The absence of precise geometric characteristics is the common bottleneck of these parametric approaches.

Recently, convolutional neural networks (CNN) have been applied to learn the implicit regression from given image appearances to 3D structures [16], [17], [18], motivated by their remarkable success in depth recovery for natural scenes. In a latest work, Feng *et al.* append a position map regression network to the 3D shape reconstruction in order to implicitly encode facial geometric priors [19]. Similarly, Tewari *et al.* [20] learn a latent corrective space to compensate personalized geometric variations. Unfortunately, subtle local variations may bring notorious geometric distortions because these methods lack explicit constraints on facial geometries.

In this study, we construct a dual neural network for 3D face reconstruction by optimizing a unified energy-based model that couples appearance-to-structure regression with explic-

This work is supported by the National Natural Science Foundation of China (Nos. 61733002, 61922019, 61672125 and 61632019), the Natural Science Foundation of Zhejiang Provincial under Grant No. Q20F020062, LiaoNing Revitalization Talents Program (XLYC1807088), the Fundamental Research Funds for the Central Universities (DUT19TD19), and the Fundamental Research Funds for the Central Universities.

X. Fan, K. Huyan and M. Hou, and R. Liu are with the DUT-RU International School of Information Science & Engineering, Dalian University of Technology, Dalian, 116024, China. (Corresponding author: Xin Fan, e-mail: xin.fan@ieee.org).

S. Cheng is with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, 310018, China.

Z. Luo is with the DUT-RU International School of Information Science & Engineering, Dalian University of Technology, Dalian, 116024, China, and is also with the Institute of Artificial Intelligence, Guilin University of Electronic Technology, Guilin, 541004, China.

The code of this paper is available at https://github.com/dlut-dimt/TMM-3D-Face-Reconstruction.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2994506, IEEE Transactions on Multimedia

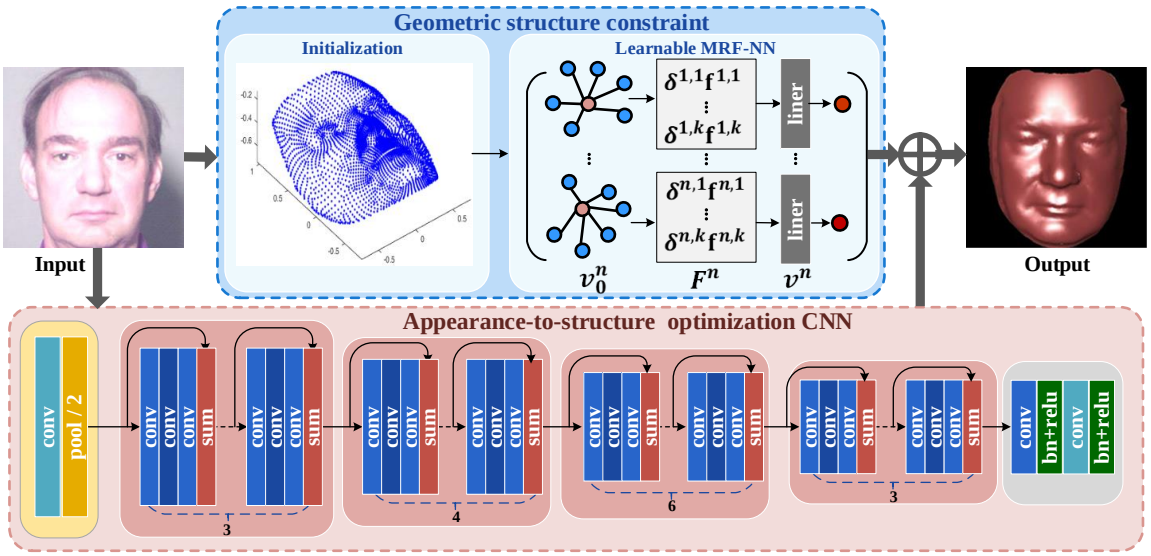


Fig. 1. The overall architecture of our dual neural networks for monocular 3D face reconstruction.

it geometric priors. Differing from an end-to-end network, we cascade residual architectures to iteratively optimize an appearance-to-structure energy defined upon the Mahalanobis distance. Additionally, we explicitly formulate the geometric constraints on 3D face clouds as an MRF energy, and solve the energy by a stack of learnable networks. Accordingly, this novel dual neural network, one derived from the appearance-to-structure energy, and the other from the geometrical constraint optimization, is able to generate realistic 3D face structures from one single image with expression and illumination variations. Fig. 1 illustrates the overall architecture that includes a residual network for appearance-to-structure optimization and a network characterizing local geometries of a vertex with respect to its neighbors as MRF-NN[1]. We summarize our contributions as:

- For the first time, end-to-end data regression embraces explicit geometric priors into one energy for monocular face reconstruction. The dual neural networks generate the 3D face structure as the solution to the energy.
- We derive a residual neural network structure from the iterative process of appearance-to-structure fitting. This network, unlike the coeffcient regression of parametric models, is able to incorporate the rich textural information given by its convolutional operators.
- We devise a stack of shallow neural networks to predict the solution to the MRF geometric constraints. To our best knowledge, it is the first neural network based optimizer (regressor) for MRF constraints on 3D structures.

We perform quantitative evaluations of the dual networks with geometric priors by comparing with end-to-end regression and popular parametric models. Comparisons with the state-of-the-art also show competitive reconstruction quality on various conditions.

[1]This constraint exhibits a form of Markov random fields.

II. Related Work

Three-dimensional (3D) face reconstruction from a single 2D image has witnessed significant advances in recent years. We focus our discussions on statistical parametric models and regression-based methods in this section.

A. Statistical Parametric Models

The most popular 3D face reconstruction method is the 3D Morpfable Model (3DMM) [9], which represents a 3D face by weighting parameters of latent candidate reference faces based on principal component analysis (PCA). Hereafter, numerous variants of parametric models have been emerging [10], [21]. The work in [22] performed face reconstruction by recovering the shape and appearance parameters from linear error estimation. Romdhani and Vetter [23] integrated pixel intensity, edges, spectacular highlights, and appearance constraints as multi-features to improve reconstruction precision. Some works attempted to expand the number of training examples to learn more accurate models. For instance, Booth *et al.* extended the space of candidate faces and provided a richer shape distribution from around 10,000 facial scans [10]. Later, the same group proposed "in-the-wild" 3DMM by combining a powerful statistical model of facial shapes describing both identity and expression with an in-the-wild appearance model [24]. More techniques including symmetry, multi-scale and spatially-varying details were considered into the modeling framework [25]. However, common issues arise that the fitting optimization requires high time complexity in these approaches.

B. Regression-based Methods

Recently, researchers consider the monocular 3D face reconstruction as a regression task. Liu *et al.* developed direct regression on facial landmarks to improve the efficiency of

single image reconstruction [11]. Jackson *et al.* employed Convolutional Neural Networks (CNN) to perform direct regression of a volumetric representation of the facial structure [26]. Deep networks [27], [28] are also introduced for 3DMM fitting regression to directly optimize the model parameters from an input image. Very deep networks are also used for 3DMM parameters regression [14], considering both geometries and textures of facial images. Following this model, Richardson *et al.* adopted synthetic face images generated from 3DMM as training examples, and then built deep network architectures to learn the regression [29]. Unfortunately, these methods for parameter regression typically lack fine geometric details. Most of them focus on the reconstruction of poses or expressions but ignore textural variations.

End-to-end networks [30], [31] are also introduced to build the regression from 2D image to 3D face clouds. Richardson *et al.* proposed an end-to-end approach for detailed face reconstruction from a single image [31]. Guo *et al.* cascaded a coarse-scale single-image network, a coarse-scale tracking network, and a fine-scale network [32]. Inspired by the depth map regression on natural scenes, Castelan *et al.* developed one network to regress 3DMM parameters [33]. Recent deep approach [34] directly predicted a nonlinear depth map learning network for fine detail recovery and translated it from synthetic images into realistic ones. The popular multi-view fusion technique [35], which combines multiple sources of information with networks, is also adopted in 3D face reconstruction task. Multi-level face models [20] combining 3DMM and deep networks were trained in an end-to-end manner in order to deal with in-the-wild images. However, these networks neglect the *geometric* priors underlying facial surfaces so that their reconstructed results exhibit unnatural local artifacts or distortions. Moreover, these methods heavily rely on training examples as no explicit priors are applied.

III. Dual Neural Networks Paradigm

Targeting at the reconstruction of 3D facial clouds $\mathbf{S}$ from a 2D observed image $\mathbf{I}$, we denote the inference process from $\mathbf{I}$ to $\mathbf{S}$ as $\mathbf{S} = \mathcal{T}(\mathbf{I})$, and formulate the reconstruction as an energy-based model (EBM) [36], [37]:

$$\begin{aligned} &\min_{\mathbf{S}} E(\mathbf{S},\mathbf{I}) + R(\mathbf{S}), \\ &s.t. \quad \mathbf{S} \in \mathbb{S}. \end{aligned} \tag{1}$$

where E denotes the energy loss which measures the compatibility of the reconstructed cloud $\mathbf{S}$ and input $\mathbf{I}$, and R is the regularizer to encode the constraints on the desired 3D face $\mathbf{S}$. $\mathbb{S}$ is a set of constraints for 3D reconstruction $\mathbf{S}$. We decompose (1) into two subproblems, *i.e.*, a deep appearance-to-structure optimization by $E(\mathbf{S},\mathbf{I}) + R(\mathbf{S})$ and a projection from the geometric structure constraint $\mathbf{S} \in \mathbb{S}$.

Different from Structure from Shading (SFS) considering the textural information and 3DMM focusing on 3D facial structures, our model (1) merges both appearance and structure into one framework. Thus, we first directly regress 3D clouds from the image $\mathbf{I}$ using the unconstrained energy function based on abundant textural information. Meanwhile, our model also embraces explicit priors on the 3D geometry. Unfortunately, it is an extremely difficult task to solve (1) by traditional numerical optimization because the iterative calculations of the inference $\mathcal{T}$ from the loss E upon priors R are intractable in an efficient way. In this work, we establish dual neural networks to replace these iterations for appearance-to-structure regression and geometric priors, respectively.

A. Deep Appearance-to-structure Optimization

Face images exhibit abundant textures that facilitate detail extraction in 3D reconstruction. To utilize these textures, we first consider the un-constraint subproblem, i.e., $L(\mathbf{S}) = E(\mathbf{S},\mathbf{I}) + R(\mathbf{S})$, and directly build an energy function reflecting the similarity of the desired 3D facial cloud $\mathbf{S}$ and the input image $\mathbf{I}$ under a specific measure for this subproblem. Different from the usual $\|\cdot\|_2$-norm, we consider the Mahalanobis distance $\|\cdot\|_{\mathbf{A}}$ and generalize it as $\langle x, x - \mathcal{A}(x)\rangle$ [2]. Accordingly, we define our energy loss as:

$$E(\mathbf{S};\mathbf{I}) = \frac{1}{2}\left\langle \mathbf{S}, \mathbf{S} - \mathcal{T}_E(\mathbf{S};\mathbf{I})\right\rangle, \tag{2}$$

where $\mathcal{T}_E$ is the regression operator. Herein, we assume that $\mathcal{T}_E$ satisfies the local homogeneity, *i.e.*, $\mathcal{T}_E(c\mathbf{S};\mathbf{I}) = c\mathcal{T}_E(\mathbf{S};\mathbf{I})$ for any $c > 0$, deducing to the approximation about derivatives as $\mathbf{S}\nabla_{\mathbf{S}}\mathcal{T}_E(\mathbf{S};\mathbf{I}) = \mathcal{T}_E(\mathbf{S};\mathbf{I})$. Therefore, the gradient of (2) can be written as:

$$\begin{aligned} \nabla_{\mathbf{S}} E(\mathbf{S};\mathbf{I}) &= \mathbf{S} - \tfrac{1}{2}\mathbf{S}\nabla_{\mathbf{S}}\mathcal{T}_E(\mathbf{S};\mathbf{I}) - \tfrac{1}{2}\mathcal{T}_E(\mathbf{S};\mathbf{I}) \\ &= (\mathbf{S} - \mathcal{T}_E(\mathbf{S};\mathbf{I})). \end{aligned}$$

Considering the gradient descent to solve the un-constraint $L(\mathbf{S})$, we have

$$\begin{aligned} \mathbf{S}_{t+1} &= \mathbf{S}_t - \nabla_{\mathbf{S}} L(\mathbf{S}_t;\mathbf{I}) \\ &= \mathcal{T}_E(\mathbf{S}_t;\mathbf{I}) + \nabla R(\mathbf{S}_t) \doteq \mathcal{T}_L(\mathbf{S}_t;\mathbf{I}), \end{aligned} \tag{3}$$

for the propagation from the input $\mathbf{I}$ to the target $\mathbf{S}$.

It is worthy noting that we are able to design various forms for $\mathcal{T}_L$ in different iterations. This variability inspires us to integrate recent effective network architectures into the propagation. Specifically, we design a residual network including three types of blocks to find the gradient descent in iterations. The light pink box with dash borders in Fig. 1 shows the detailed profiles of each block. The main part of this architecture consists of several "residual" blocks, taking a feature map as the input and outputting a residual feature map to approximate the gradient operator. Each residual block contains a cascade of three convolutions whose kernel sizes and channel numbers are listed in the last column of Tab. I. Besides, one "Conv + max pooling" precedes these residual blocks while two "Conv + BN +ReLU" blocks follow them. The last two "Conv+BN+ReLU" blocks include a 1×1 and a 7×7 convolution layer, respectively. The 1×1 convolutional layer reduces the dimension of the feature map while the the last layer with the output size 1×1 layer recovers the output feature map to a vector with the residual for each vertex of a 3D face cloud. Tab. I gives the detailed network configurations for appearance-to-structure optimization. Thus, the iterations for solving $L(\mathbf{S})$ can be expressed as:

$$\mathbf{S}_{T-1} = \mathcal{T}_L^{T-1}(\mathbf{S}_{T-2};\mathbf{I}) \circ \cdots \circ \mathcal{T}_L^{2}(\mathbf{S}_1;\mathbf{I}),$$

Layer Name	Output Size	52-layer
Conv + max pooling	56×56	7×7, 64, stride 2 3×3 max pooling, stride 2
Residual block 1-3	56×56	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
Residual block 4-7	28×28	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
Residual block 8-13	14×14	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
Residual block 14-16	7×7	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
Conv + BN + ReLU	7×7	1×1, 128
Conv + BN + ReLU	1×1	7×7, 10344

TABLE I
NETWORK CONFIGURATIONS FOR APPEARANCE-TO-STRUCTURE OPTIMIZATION.

where $\mathbf{S}_1 = \mathbf{I}$ extracting the appearance from a 2D face.

To facilitate the training process, we regard the residual between the output of regularizers $\bar{\mathbf{S}}$ and target $\hat{\mathbf{S}}$ [3] as the objective, and formulate the loss function for training as follows:

$$\min \sum_{m=1}^{M} ||\mathcal{T}_L(\mathbf{I}_m; \mathbf{I}_m) - \Delta \mathbf{S}_m||_2^2,$$

where $\mathcal{T}_L = \mathcal{T}_L^{T-1} \circ \cdots \circ \mathcal{T}_L^0$ is the whole residual network, $\Delta \mathbf{S}_m = \hat{\mathbf{S}}_m - \bar{\mathbf{S}}_m$, and M is the number of input images. Thus, we have the reconstruction as $\mathbf{S}_T = \mathcal{T}_L^T(\mathbf{S}_{T-1}; \mathbf{I}) = \bar{\mathbf{S}} + \mathcal{T}_L(\mathbf{I}, \mathbf{I})$. The intermediate 3D reconstruction is denoted as $\bar{\mathbf{S}}$ to which we apply the following geometric constraint.

B. Geometric Structure Constraint

To establish the constraint term in (1), we intend to find an explicit projection $\mathcal{P}$ from the output of un-constraint optimization $\mathbf{S}_T$ to the desired 3D reconstructed structure $\mathbf{S}_\mathcal{P}$:

$$\mathbf{S}_\mathcal{P} = \mathcal{P}(\mathbf{S}_T) \in \mathbb{S}.$$

Considering that the appearance-to-structure optimization focuses on the textural regression, we explore the cloud topology which implies the plenty geometric information critical to 3D face reconstruction by learnable MRFs $\mathcal{T}_\mathcal{P}$.

1) Initializing personalized template: Statistical face models are powerful tools to reconstruct the basic geometry structure from a single facial image. To alleviate the difficult on solving (1), we follow the advantage of 3DMM on pose recovery and prepare a global regularizer by a statistical model performing PCA on a training set with densely registered 3D meshes. Specifically, we generate a 3D face by tuning parameters on the linear bases $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_m]$ of facial surfaces given by PCA, formulated by the following expression:

$$\mathbf{S}_{\boldsymbol{\alpha}} = \bar{\mathbf{u}} + \sum_i^m \alpha_i \sigma_i \mathbf{u}_i,$$

where $\bar{\mathbf{u}}$ denotes the averaged facial mesh of training data, α_i is the sparse coefficient in the shape PCA space and, σ_i is the standard deviation.

To recover the primary geometry transformation, we build a cost function inspired by the simple shape-to-landmarks fitting in [38] as follows:

$$\mathbf{S}_0 = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2\sigma_L^2} \sum_{i=1}^{n_L} ||\pi_i(\texttt{Proj}\mathbf{S}_{\boldsymbol{\alpha}}) - \mathbf{L}_i||^2 + \lambda ||\boldsymbol{\alpha}||^2, \quad (4)$$

where λ is the regularization weight, n_L is the number of facial landmarks, and $\mathbf{L}_i = [x_i, y_i]$ denoting the 2D coordinate of one facial landmark. σ_L is an optional variance for these landmarks. `Proj` and π_i are the projection from a 3D face $\mathbf{S}_\alpha$ to a 2D plane and the operator to extract facial landmarks [39], respectively. We can solve (4) using a standard least square algorithm, generating a personalized template $\mathbf{S}_0$ with its facial contour.

The personalized 3D template $\mathbf{S}_0$ also establishes the dense correspondence between 3D vertices and 2D image pixels. We employ this correspondence to produce a color map for each vertex of the 3D template [38].The r, g, b value at the corresponding pixel is taken as the color of the 3D vertex. For those vertices invisible in the image, we set their r, g, b values as 0. Hence, the personalized template including the 3D cloud and color map is ready for learning the structural constraint.

2) Learnable Markov random fields constraint: The personalized template restores the basic global geometry and appearance for a given face image using its facial landmarks. These 2D facial landmarks are not able to convey the geometric information of non-flat regions on a human face such as the surfaces of nose and cheek. Meanwhile, only hundreds of 3D faces, which build the statistical parametric model, can hardly cover a wide range of changes in age, gender, and race. Therefore, we resort to a non-parametric local constraint on 3D structures.

From this respect, we learn the transformation $\mathcal{P}$ from the personalized template $\mathbf{S}_0$ to the desired target $\mathbf{S}$. Alternating to an explicit form of $\mathcal{P}(\mathbf{S})$, we separate the complicated geometric mapping as follows by transforming each vertex $\mathbf{v}$ in turns,

$$\mathcal{P}(\mathbf{S}) = \mathcal{P}(\mathbf{S}; \mathbf{S}_0) = \prod_{\mathbf{v}^i \in \mathbf{S}} \mathcal{P}^i(\mathbf{v}^i; \mathbf{S}_0), \quad (5)$$

where $\mathcal{P}^i$ denotes the prior for a vertex $\mathbf{v}^i$. We apply the MRF constraint [40] on the 3D cloud $\mathbf{S}$, i.e., each vertex $\mathbf{v}^i \in \mathbf{S}$ is only relates to its neighbors, considering the local Markov assumption widely adopted in computer vision tasks [41]. Following the local conditional probability density, (5) turns to:

$$p(\mathbf{S}|\mathbf{S}_0) = \prod_{\mathbf{v}^i \in \mathbf{S}} p(\mathbf{v}^i|\mathbf{S}_0) = \prod_{\mathbf{v}^i \in \mathbf{S}} p(\mathbf{v}^i|B(\mathbf{v}_0^i)), \quad (6)$$

where $B(\mathbf{v}_0^i)$ denotes the K nearest neighbors of $\mathbf{v}_0^i \in \mathbf{S}_0$. The second equality above holds upon the observation that the

[2]The detailed derivation will be appeared in appendix.

[3]We will detail the generation of $\hat{\mathbf{S}}$ in Sec. IV-A.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2994506, IEEE Transactions on Multimedia

vertices in a facial cloud are only effected by the K nearest neighbors in practice.

3) Local shallow network: To learn the probability density p of each vertex in (6), we propose a local shallow neural network $\mathcal{T}_{\mathcal{P}}$ including feature extraction and regression. Firstly, we regard that each vertex is associated with an undirected and connected local graph as $\mathcal{G}^i = (\mathcal{V}^i, \mathcal{E}^i, \mathcal{W}^i)$, where the set $\mathcal{E}^i$ consists of the edges between $\mathbf{v}_0^i$ and its neighbors, and $\mathcal{W}^i$ denotes the weighted adjacency matrix encoding the connection weight in each edge. We concatenate the coordinate $I(\cdot)$ and color feature $C(\cdot)$ of each mesh vertex $\mathbf{v}_0^i$ as the node of the local graph, i.e., $\mathcal{V}^i = [I(\mathbf{v}_0^i)^\top; C(\mathbf{v}_0^i)^T]$. To obtain the K nearest neighbors of $\mathbf{v}_0^i$, we define the distance of $\mathbf{v}_0^i$ and $\mathbf{v}_0^j \in \mathbf{S}_0$ as

$$w^{i,j} = \left\| [\mathbf{v}_0^i, C(\mathbf{v}_0^i)] - [\mathbf{v}_0^j, C(\mathbf{v}_0^j)] \right\|^2,$$

and then index the K neighbors of $\mathbf{v}_0^i$ as $\{j_k\}_{k=1,\cdots,K}$ such that $w^{i,j_1} \leq w^{i,j_2} \leq \cdots \leq w^{i,j_K}$.

For each mesh vertex, we only introduce the local geometry and appearance information as the descriptor. The feature F of a vertex is defined as:

$$\mathbf{F}^i(\mathbf{v}_0^i) = [\delta^{i,j_1}\mathbf{f}^{i,j_1}, \cdots, \delta^{i,j_k}\mathbf{f}^{i,j_k}, \cdots, \delta^{i,j_K}\mathbf{f}^{i,j_K}]^T,$$

where $\mathbf{f}^{i,j_k} = [\mathbf{v}_0^{j_k}, C(\mathbf{v}_0^{j_k})]$, and the weight δ^{i,j_k} can be calculated by $\delta^{i,j_k} = 1 - k/K$ when $k = 1, \cdots, K$.

To tackle the difficulties in the training stage, we build a shallow neural network for each vertex. The "Learnable MRF-NN" part in Fig. 1 demonstrates the architecture of our network, which only feeds the designed feature $\mathbf{F}^i$ as the input and outputs the coordinates of the target 3D point.

4) Training strategy: Xie *et al.* develop a ResNet to improve the accuracy of image classification by training multiple simple-structure CNNs for the same target [42]. They impose an additional "cardinality" (the size of the transformation set) as an essential factor in addition to the dimensions of depth and width in order to improve the ability of CNNs. We take advantage of their practice to train multiple simple-structured neural networks, but differently, we use one shallow neural network on vertex to cope with the local geometric structure for residual regression. Our method divides the original problem into a number of sub-problems that is easier to solve. The shallow neural network has only one hidden layer, fully connected with the input and output nodes. The training objective for the shallow neural network in node i is:

$$\min \sum_{m=1}^{M} \|\mathcal{T}_{\mathcal{P}}^i(\mathbf{F}^i(\mathbf{v}_{0,m}^i)) - \hat{\mathbf{v}}_m^i\|^2,$$

where m is the index of input models and $\hat{\mathbf{v}}_m^i \in \hat{\mathbf{S}}_m$ is the target 3D facial clouds.

Our shallow network imposes learnable local geometric constraints to the global restoration $\mathbf{S}_{R_G}$ vertex by vertex, and thus the explicit updating of $\mathbf{S}$ can be generalized by

$$\mathbf{S}_{\mathcal{P}} = [\mathcal{T}_{\mathcal{P}}^1(\mathbf{F}^1(\mathbf{v}_0^1)), \cdots, \mathcal{T}_{\mathcal{P}}^N(\mathbf{F}^N(\mathbf{v}_0^N))],$$

where N is the number of 3D points. Recalling the training loss of the appearance-to-structure residual network, we can regard $\mathbf{S}_{\mathcal{P}}$ as $\bar{\mathbf{S}}$ in the appearance-to-structure regression.

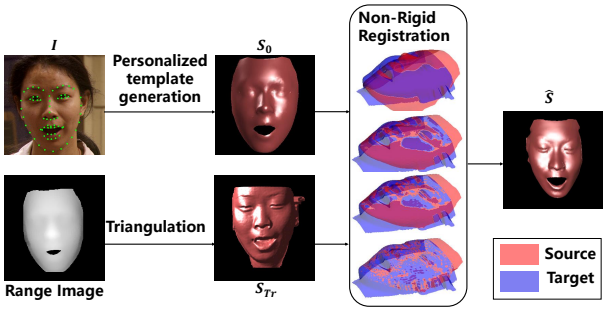


Fig. 2. The algorithmic pipeline of the data preparation.

IV. Experimental Results

In the experiments, we first provide the preparation process of training examples and then discuss several parameter settings for the dual networks including the residual CNN and learnable shallow networks for MRF priors. The quantitative comparisons on FRGC v2.0 [43] validate the effectiveness of our mechanism. Finally, the qualitative inspections on the reconstruction from arbitrary facial images are compared with the state-of-the-art methods. All experiments are performed on a PC with Intel Core i7 CPU at 3.4 GHz, 32 GB RAM and a NVIDIA GeForce GTX 1050 Ti GPU.

A. Data Preparation

Our training examples mainly come from the FRGC v2.0 data set, containing 4,007 range images and 2D textural images of 466 individuals. Most of them are in a frontal pose with various facial expressions. These 2D textural and range images are well registered with 68 facial landmarks labeled.

A large number of training examples are always critical to train deep networks. Unfortunately, the original FRGC v2.0 data set provides no enough 3D faces along with the dense correspondence with 2D facial images. We leverage a non-rigid ICP algorithm [44] for surface registration to augment the data set. Fig. 2 illustrates the data augmentation process. Firstly, we triangulate the range images into 3D meshes as the target 3D faces $\mathbf{S}_{Tr}$. Secondly, we fit a statistical face model to the corresponding 2D facial images using labeled facial landmarks as the initialization for the non-rigid ICP algorithm. This initial template $\mathbf{S}_0$ works not only to start ICP but also initialize our learnable MRF model as $\mathbf{S}_{\mathcal{P}}$. Finally, we align these source face models $\mathbf{S}_0$ to their targeting face models $\mathbf{S}_{Tr}$, producing the models $\hat{\mathbf{S}}$ with dense correspondences. Consequently, all these ground truth 3D face models have the same number of vertices and topological structures. This process is also applicable to the data augment for training more complex deep networks other than ours with the MRF prior.

B. Parameter Setting

One characteristic of our MRF prior on 3D facial structures distinguished from traditional ones lies in the learnable shallow network associate with each vertex for a fast energy inference. In this part, we provide practical trials on the parameter settings for the shallow network, including the value

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2994506, IEEE Transactions on Multimedia

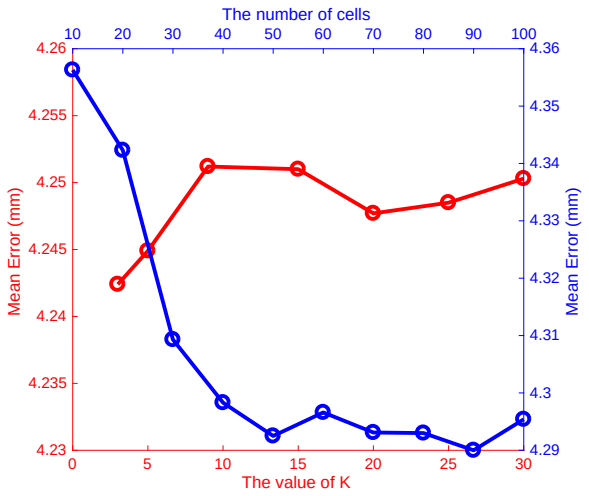


Fig. 3. Mean error of our method with different k value and number of hidden layer cell in learnable MRF-NN.

Architecture	Mean Error (mm)	Running Time (s)
ResNet-18	4.1377	**0.13**
ResNet-34	4.1199	0.18
ResNet-50	4.1200	0.14
ResNet-101	**4.1198**	0.19
ResNet-152	4.1121	0.24

TABLE II
THE COMPARISON OF DIFFERENT RESNET ARCHITECTURES.

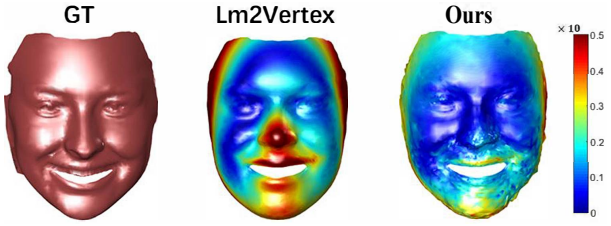


Fig. 4. Heat maps of Lm2Vertex and our method.

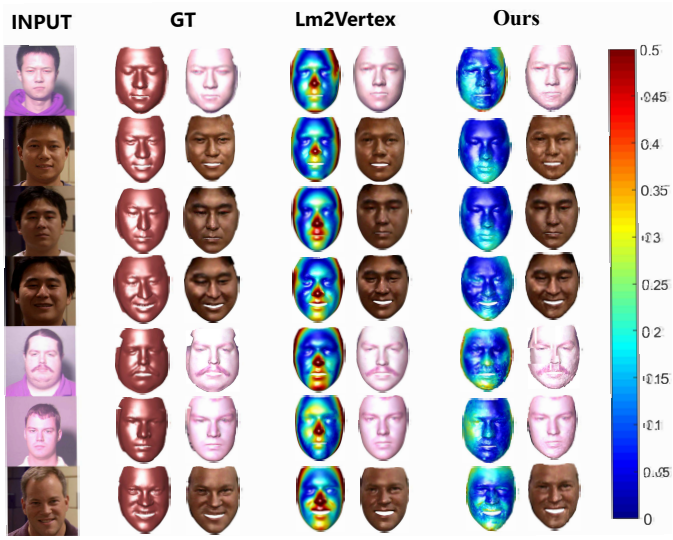


Fig. 5. Heat maps and visible results of Lm2Vertex and ours on FRCG v2.0.

of k and the number of hidden layer cells. We randomly pick 1,000 from all 2,800 training examples generated from the data preparation procedure, and the other 500 from the rest for testing. We fix the number of hidden layer cells to 50, and train the shallow neural network in our local MRF varying K from 3 to 30 as $K = [3, 5, 9, 15, 20, 25, 30]$. The red curve in Fig. 3 demonstrates the mean error of the shallow network with different K values. The mean errors exhibit such subtle differences for different K values that the maximal gap is less than 0.01. Meanwhile, the mean errors become slightly larger when increasing K. These results verify that only a few nearest neighbors on the graph suffice to well characterize the local geometric structures of human faces. To ensure accuracy and efficiency of local MRFs, we set the value of $K = 10$ in the subsequent experiments. The blue curve in Fig. 3 illustrates the mean errors versus the cell numbers in the hidden layer. When the number of hidden layer cells increases, the mean error gradually decreases, and becomes stable when the cell number is greater than 50. We set the number of hidden layer cells to 50 in the subsequent experiments in order to prevent overfitting and simplify the network structure.

We also conduct the quantitative comparison on different configurations of ResNet 18-152 for appearance-to-structure optimization [45]. Tab. II demonstrates that ResNet-50 well balances accuracy and efficiency, giving almost identical mean error (0.0002mm more) as ResNet-101 while running close to ($0.01s$ more) the fastest one, ResNet-18. Therefore, we adopt ResNet-50 as the backbone network for appearance-to-structure optimization in all our experiments.

C. Quantitative Comparisons on FRGC v2.0

Quantitative results play an important role in measuring the quality of the visual results in a more detailed way. Thus, in this section, we normalize the mean error to generate the heat map for better visualization:

$$e_{ij} = \frac{\hat{e}_{ij}}{\max(\hat{e}_{i1}, \hat{e}_{i2}, ..., \hat{e}_{iM})} \times 10, \tag{7}$$

where $\hat{e}_{ij}$ is the origin mean error, M is the number of compared methods, and i and j are the indices of the testing image and method, respectively.

First, to verify the performance of our method, we compare with a classical technique Lm2Vertex [11][4], a representative variant of 3DMM models. Lm2Vertex trains a linear regressor from 2D facial landmarks to model parameters of 3DMM and generates 3D faces by applying the parameters to 3DMM. Fig. 4 shows the corresponding heat maps of testing images from FRGC v2.0. It can be seen that Lm2Vertex has a poor ability to accurately recover the face outline and some facial components, especially in the regions of nose and chin, though it tries to predict the face surface upon explicit facial geometry. In contrast, our method provides a much better reconstruction

[4]We implemented this method having the same vertices in all 3D faces for a fair quantitative comparison upon (7). The other approaches, especially end-to-end deep networks, generate different meshes using various numbers of learned parameters.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2994506, IEEE Transactions on Multimedia

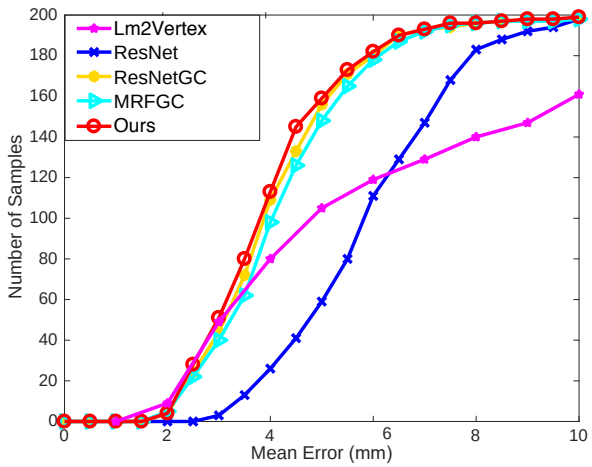


Fig. 6. The comparisons between ResNet, MRFGC, ResNetGC, and our method on FRGC v2.0. The x-axis is the mean error, and the y-axis indicates the number of samples on which mean errors are lower than the x value.

	ResNet	ResNetGC	MRFGC	Ours
ResNet	√	√		√
GC		√	√	√
L-MRFs			√	√
Mean Error (mm)	5.9661	4.1363	4.2680	**4.0121**

TABLE III
THE COMPONENT AND MEAN ERROR OF DIFFERENT STRATEGIES.

in terms of face outline and local structures neighboring to prominent facial components like nose, mouth and eyes, showing great effectiveness on local geometry reconstruction. Fig. 5 provides more comparisons with Lm2Vertex on the FRGC v2.0 benchmark. It indicates that our approach has smaller mean errors and richer textures than Lm2Vertex.

As ablation analysis, we secondly compare among different strategies including the end-to-end (image-to-3D) data-driven network (ResNet), explicit 3DMM geometric constraints (GC and Lm2Vertex), naive cascade of ResNet and 3DMM (ResNetGC), naive cascade of 3DMM and learnable MRF (MRFGC), and our dual networks. Tab. III lists detailed configuration for each strategy and the mean errors on FRGC v2.0. Both data-driven residual networks and learnable MRF-NN improve the reconstruction accuracy over the personalized template. Our dual network has the least mean error among these strategies beacause both textural information available in 2D images and geometrical structures in 3D meshes are well characterized and incorporated. We demonstrate the cumulative error distribution curves for all compared strategies on 200 testing images from FRGC v2.0 in Fig. 6 in order to peer into the respective impacts. The errors are much higher when the geometrical constraint is missing (ResNet). Our method obtains the best by coupling both data-driven and learnable geometric priors.

For more detailed analysis, we provide comparisons on different conditions including facial images with the neutral expression, large expression variations and illumination changing [46]. Figs. 7, 8, and 9 show the reconstruction

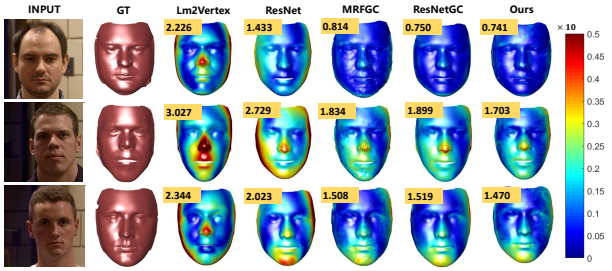


Fig. 7. The comparison of Lm2Vertex, ResNet, MRFGC, ResNetGC, and our method on facial images with neural expression. The highlighted numbers give the normalized mean errors.

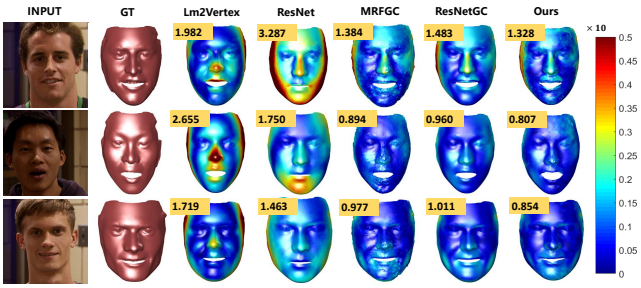


Fig. 8. The comparison of Lm2Vertex, ResNet, MRFGC, ResNetGC, and our method on facial images with large expression. The highlighted numbers give the normalized mean errors.

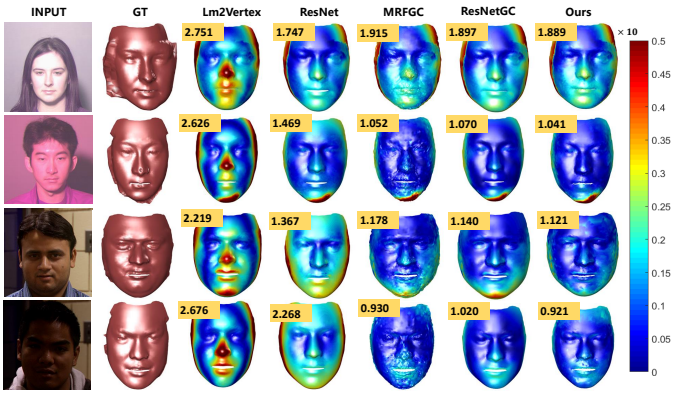


Fig. 9. The comparison of Lm2Vertex, ResNet, MRFGC, ResNetGC, and our method on facial images with illumination variation. The highlighted numbers give the normalized mean errors.

results of these methods in both quantitative and qualitative way simultaneously. As we can see, Lm2Vertex inherits the drawbacks of statistical models as it is trained on the data from a facial statistical model, which makes it work fine on recovering rough facial geometry but fail in generating subtle facial details. Additionally, Lm2Vertex only considers the geometric prior of input facial images, while neglects the most important textural information. As for ResNet, this approach predicts the 3D cloud from an input facial image without any initialization. Although the mean error of ResNet is a bit lower than Lm2Vertex, the reconstructed surfaces are rather close to the mean shape of the training dataset as shown in Fig. 7, where all the results of ResNet present a similar shape. Even worse, it cannot characterize expression variations as showed in Fig. 8. Initialized by the personalized templates, ResNetGC

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2994506, IEEE Transactions on Multimedia

Fig. 10. The comparison between Lm2Vertex, Pix2Vertex, Pix2Parameters, Pix2Voxel, PR-Net and our method on FRGC v2.0.

is able to correctly converge and all the reconstructed faces are globally smoothing with their fundamental geometric structures well preserved. Also, ResNetGC is ideally robust to illumination variations and generates a satisfactory face even in the darkness as shown in the fourth row of Fig. 9. However, there still exists a little gap between ResNetGC and MRFGC as shown in Fig. 6, indicating a weaker ability in handling details recovery in various illuminations. Absorbing all the aforementioned benefits, our method couples texture with structure to jointly restore 3D mesh from the input facial images, ideally producing the lowest mean error among all the methods as illustrated in Fig. 6 and Tab. III.

D. Comparisons with the State-of-the-art

To further evaluate the performance of our method, we compare with the state-of-the-art approaches on both synthetic and in-the-wild benchmarks.

1) Comparison Results on FRGC v2.0: Fig. 10 shows comparisons with Lm2Vertex [47], Pix2Vertex [34], Pix2Voxel [26], Pix2Parameters [14], PG-Net [19] and our method on FRGC 2.0. Since 3D clouds from Pix2Vertex, Pix2Voxel, Pix2Parameters, and PR-Net have different numbers of vertices, it is hard to align them to a standard face shape in our dataset and to generate their heat maps. Thus, we only show a qualitative comparison on reconstructed faces. Pix2Vertex is a CNN based method that recovers a depth map for the input facial image. It is sensitive to local variations and ignores the inherent global geometry structure of human faces. Its results present rich details but have very large distortions. Although Lm2Vertex reconstructs face shapes with geometry prior, it can only recover their basic geometry structures with wrong shapes and details. The noses of its results are too high showing large errors. Although Pix2Parameters combines local and global texture information for 3DMM parameter regression, its results are very different from their images. It verifies that a 3DMM model can hardly cover large variations of 3D faces on races, expressions, and ages. The results of Pix2Voxel look so good that all of them are very similar to their corresponding 2D images. However, their surface are not smooth enough and some details of face cannot be reconstructed. PR-Net has better restoration on details, but there are some missing structures around eyes. In contrast, our method presents better both on geometric structures and details than the others.

2) Comparison Results on LFPW: We further compare with some state-of-the-art approaches including Lm2Vertex [11], Pix2Vertex [34], Pix2Parameters [14], Pix2Voxel [26], and PR-Net[19] on facial images from LFPW [48] which is a wild benchmark including 1432 faces from images downloaded from the Web. Fig. 11 demonstrates the 3D structure reconstruction with texture maps[5]. Our method does not on-

[5]We cannot provide quantitative results on this set since no corresponding ground truth 3D face model is available. Additionally, if there is no preprocessing for registration in the compared algorithms, we first need to register the 3D reconstructed point cloud, and then calculate the mean error. Unfortunately, the approximate operation from registration brings errors leading to unfair comparisons.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2994506, IEEE Transactions on Multimedia

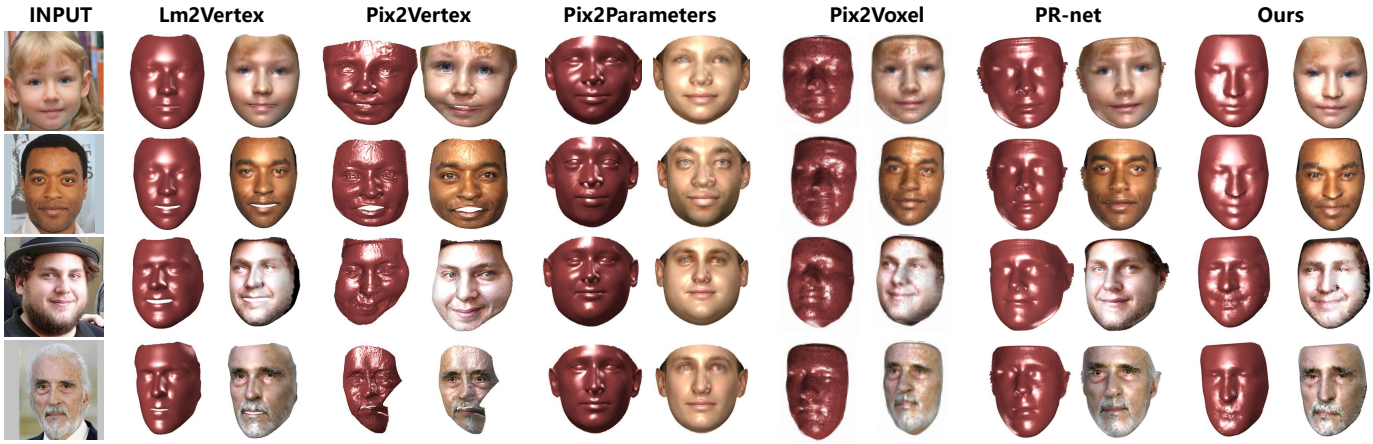


Fig. 11. The comparison between Lm2Vertex, Pix2Vertex, Pix2Parameters, Pix2Voxel, PR-Net and our method on LFPW.

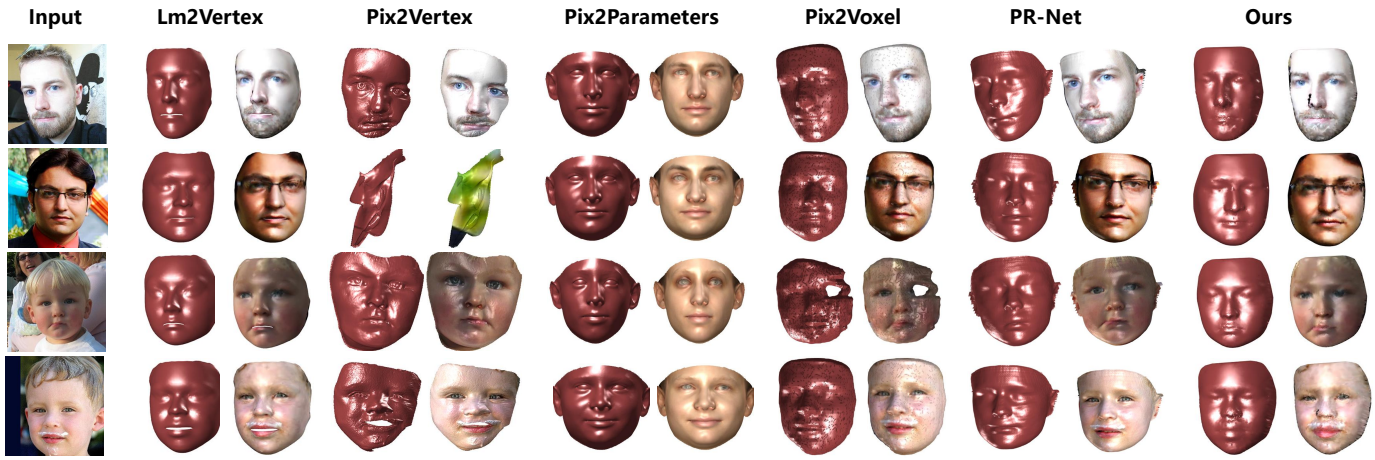


Fig. 12. The comparison between Lm2Vertex, Pix2Vertex, Pix2Parameters, Pix2Voxel, Y-GCRN and GRN on Helen.

ly produce accurate global facial geometries on prominent facial components including chin, cheek, and nose, but also impose facial details, *e.g.*, beards and wrinkles. PR-Net also presents compelling detail recovery, but brings excessive artifacts (wrinkles) on the forehead, especially for the little girl, since PR-net focuses on learning more details from images including plenty of textures. The other end-to-end pixel recovery methods (Pix2Vertex and Pix2Voxel) recover facial details but present evident geometric distortions while those from explicit geometrical models (Lm2Vertex and Pix2Parameters) smear facial details though maintaining basic structures. Specifically, as for Pix2Vertex, it is a CNN-based method that directly regresses a depth map from the input image. Actually, the direct regression only takes the local texture information into account, while ignoring the inherent global geometry structure of human faces. Hence, it is sensitive to local variations. Its results exhibit rich and fine details, but have very large distortions from plausible faces. Lm2Vertex works well on preserving the global geometry of 3D faces. Consequently, the resultant 3D faces well present basic geometry structures, but fail to generate facial details so that their faces look far away from the corresponding facial images. Pix2Parameters predicts 3DMM parameters using a CNN, incorporating both the geometry and texture of facial images. However, the results look fatter than the faces in the corresponding images, showing that the global statistical model can hardly recover subtle facial details. Pix2Voxel regresses the volumetric representation of the 3D facial geometry from a single 2D image. However, the reconstruction appears neither smoothing nor rich details.

3) Comparison Results on Helen: Fig. 12 shows qualitative comparisons among Lm2Vertex [11], Pix2Vertex [34], Pix2Parameters [14], Pix2Voxel [26], PR-Net[19] on arbitrary facial images from the Helen data set [49]. Although Pix2Vertex shows a good result on the second facial image, it is hard for this method to overcome variations on poses, expressions, and illuminations without any global geometry constraints. Most of its results still have very large distortions, and the method fails reconstructing the fourth image. Lm2Vertex shows good results on reconstructing global geometry structures, but it fails to describe prominent facial details, such as eyes, nose, and check. The results of Pix2Parameters are still very different from their images. It cannot recover the expression variation of the third image. The results on the third row of Pix2Voxel show that it cannot well recover the expressions from input images and the results on the fifth row show large distortions. Pix2Voxel provides similar visual results that recover the basic facial geometry and also well characterize the details on a face including the height of the

nose, shapes of cheeks, eyes and facial expressions. PR-Net also has some over fitting on the forehead of children. Again, our method provides better visual results over the others.

V. Conclusion and Future Work

In this paper, we propose a unified energy function incorporating end-to-end data regression and explicit geometric priors, which brings dual neural networks to solve the proposed energy. One data-driven network recovers 3D details from abundant facial textures, and the other prior-derived stack of shallow networks characterizes 3D facial geometries. We compare our methods with the sate-of-the-art on the FRGC v2.0, LFPW and Helen data sets. Our method presents superior performance in both qualitative and quantitative comparisons. The proposed paradigm of jointly learning data-driven and prior-derived deep networks demonstrate great power on handling both implicit textural and explicit geometric constraints.

The convergence of iterative learnable networks to the defined energy on 3D structures remains unresolved. We will investigate the theoretic properties of the proposed learnable optimizer for the energy upon the framework of [50] in the future. Our approach gains satisfactory performance on both synthetic and in-the-wild challenging datasets with 4000+ range images from 400+ individuals showing variations on ages, poses and expressions. Unfortunately, these instances only cover a very limited fraction of human faces. A larger scale of training and testing sets including more significant changes on textures and geometries need to be explored in the future.

Appendix A
Preliminaries derivation in Sec. III-A

Before deriving the iterative formulation appeared in deep appearance-to-structure optimization, we first present how to transfer the Mahalanobis distance $\|\cdot\|_{\mathbf{A}}$ to $\langle x, x-\mathcal{A}(x)\rangle$. Assuming $\mathbf{A}=\mathbf{e}-\mathbf{B}$ and $\mathbf{e}$ is the identity matrix, we have

$$\begin{aligned}\|\mathbf{S}\|_{\mathbf{A}} &= \langle \mathbf{S}, \mathbf{A}\mathbf{S}\rangle = \langle \mathbf{S}, (\mathbf{e}-\mathbf{B})\mathbf{S}\rangle \\ &= \langle \mathbf{S}, \mathbf{S}-\mathbf{B}(\mathbf{S})\rangle = \langle \mathbf{S}, \mathbf{S}-\mathcal{A}(\mathbf{S})\rangle.\end{aligned}$$

The last equality holds under denoting $\mathbf{B}(\mathbf{S})$ as $\mathcal{A}(\mathbf{S})$.

Inspired by the above generalized formulation, we define the energy loss $E(\mathbf{S},\mathbf{I})$ as

$$E(\mathbf{S},\mathbf{I}) = \frac{1}{2}\left\langle \mathbf{S}, \mathbf{S}-\mathcal{T}_E(\mathbf{S},\mathbf{I})\right\rangle.$$

Assuming $\mathcal{T}_E$ is differentiable and $\mathcal{T}_E(c\mathbf{S},\mathbf{I}) = c\mathcal{T}_E(\mathbf{S},\mathbf{I})$, the derivation of $\mathcal{T}_E$ about $\mathbf{S}$ can be calculated as follows:

$$\begin{aligned}&\nabla_{\mathbf{S}}\mathcal{T}_E(\mathbf{S},\mathbf{I}) \\ &= \lim_{\epsilon\to 0}\frac{\mathcal{T}_E(\mathbf{S}+\epsilon\mathbf{S})-\mathcal{T}_E(\mathbf{S})}{\epsilon\mathbf{S}} \\ &= \lim_{\epsilon\to 0}\frac{(1+\epsilon)\mathcal{T}_E(\mathbf{S})-\mathcal{T}_E(\mathbf{S})}{\epsilon\mathbf{S}} \\ &= \frac{\mathcal{T}_E(\mathbf{S})}{\mathbf{S}},\end{aligned}$$

which is equal to $\mathcal{T}_E(\mathbf{S}) = \nabla_{\mathbf{S}}E(\mathbf{S},\mathbf{I})\mathbf{S}$. Therefore, we have

$$\begin{aligned}&\nabla_{\mathbf{S}}L(\mathbf{S},\mathbf{I}) \\ &= \nabla_{\mathbf{S}}E(\mathbf{S},\mathbf{I}) + \nabla_{\mathbf{S}}R(\mathbf{S},\mathbf{I}) \\ &= \mathbf{S} - \frac{1}{2}\mathbf{S}\nabla_{\mathbf{S}}\mathcal{T}_E(\mathbf{S},\mathbf{I}) - \frac{1}{2}\nabla_{\mathbf{S}}\mathcal{T}_E(\mathbf{S},\mathbf{I}) + \nabla_{\mathbf{S}}R(\mathbf{S},\mathbf{I}) \\ &= \mathbf{S} - \mathcal{T}_E(\mathbf{S},\mathbf{I}) - \nabla_{\mathbf{S}}R(\mathbf{S},\mathbf{I}),\end{aligned}$$

which deduces the iterations of solving $\min_{\mathbf{S}} L(\mathbf{S},\mathbf{I})$ by gradient descent as (3):

$$\mathbf{S}_{t+1} = \mathbf{S}_t - \nabla_{\mathbf{S}}L(\mathbf{S},\mathbf{I}) = \mathcal{T}_E(\mathbf{S},\mathbf{I}) + \nabla_{\mathbf{S}}R(\mathbf{S},\mathbf{I}).$$

References

[1] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Trans. on Multimedia*, vol. 17, no. 11, pp. 2049–2058, 2015.

[2] C. Jing, Z. Dong, M. Pei, and Y. Jia, "Heterogeneous hashing network for face retrieval across image and video domains," *IEEE Trans. on Multimedia*, vol. 21, no. 3, pp. 782–794, 2018.

[3] S. Kenny, N. Mahmood, C. Honda, M. J. Black, and N. F. Troje, "Perceptual effects of inconsistency in human animations," *ACM Trans. on Applied Perception*, vol. 16, no. 1, p. 2, 2019.

[4] Y. He, J. Shi, C. Wang, H. Huang, J. Liu, G. Li, R. Liu, and J. Wang, "Semi-supervised skin detection by network with mutual guidance," in *Proc. IEEE Conf. Int. Conf. Comput. Vis.*, 2019, pp. 2111–2120.

[5] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset," *IEEE Trans. on Multimedia*, vol. 17, no. 6, pp. 804–815, 2015.

[6] X. Zhao, Y. Lin, and J. Heikkilä, "Dynamic texture recognition using volume local binary count patterns with an application to 2d face spoofing detection," *IEEE Trans. on Multimedia*, vol. 20, no. 3, pp. 552–566, 2017.

[7] B. Chen, C. Jung, and Z. Zhang, "Variational fusion of time-of-flight and stereo data for depth estimation using edge-selective joint filtering," *IEEE Trans. on Multimedia*, vol. 20, no. 11, pp. 2882–2890, 2018.

[8] D. A. Forsyth and J. Ponce, *Computer vision: a modern approach*. Prentice Hall Professional Technical Reference, 2002.

[9] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *SIGGRAPH*, 1999, pp. 187–194.

[10] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, "A 3D morphable model learnt from 10,000 faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5543–5552.

[11] F. Liu, D. Zeng, Q. Zhao, and X. Liu, "Joint face alignment and 3D face reconstruction," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 545–560.

[12] C. Cao, Q. Hou, and K. Zhou, "Displaced dynamic expression regression for real-time facial tracking and animation," *ACM Trans. on Graphics*, vol. 33, no. 4, p. 43, 2014.

[13] C. Cao, Y. Weng, S. Lin, and K. Zhou, "3d shape regression for real-time facial animation," *ACM Trans. on Graphics*, vol. 32, no. 4, p. 41, 2013.

[14] A. T. Tran, T. Hassner, I. Masi, and G. Medioni, "Regressing robust and discriminative 3D morphable models with a very deep neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1493–1502.

[15] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman, "Unsupervised training for 3d morphable model regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8377–8386.

[16] M. Tatarchenko, A. Dosovitskiy, and T. Brox, "Multi-view 3D models from single images with a convolutional network," in *Proc. Eur. Conf. Comput. Vis.*, no. 1, 2016, pp. 231–257.

[17] M. Feng, S. Z. Gilani, Y. Wang, and A. Mian, "3d face reconstruction from light field images: A model-free approach," in *Proc. Eur. Conf. Comput. Vis.*, 2018.

[18] L. Tran and X. Liu, "Nonlinear 3d face morphable model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7346–7355.

[19] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," in *Proc. Eur. Conf. Comput. Vis.*, 2018.

[20] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt, "Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018.

[21] P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W. Christmas, M. Ratsch, and J. Kittler, "A multiresolution 3D morphable face model and fitting framework," in *Int. Jounal of comput. Vis.*, 2016.

[22] S. Romdhani, V. Blanz, and T. Vetter, "Face identification by fitting a 3D morphable model using linear shape and texture error functions," *Proc. Eur. Conf. Comput. Vis.*, pp. 3–19, 2006.

[23] S. Romdhani and T. Vetter, "Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, 2005, pp. 986–993.

[24] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, S. Zafeiriou *et al.*, "3D face morphable models in-the-wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[25] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schönborn, and T. Vetter, "Morphable face models-an open framework," in *IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 75–82.

[26] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large pose 3D face reconstruction from a single image via direct volumetric cnn regression," in *Proc. IEEE Conf. Int. Conf. Comput. Vis.*, 2017, pp. 1031–1039.

[27] P. Dou, S. K. Shah, and I. A. Kakadiaris, "End-to-end 3D face reconstruction with deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 5, 2017.

[28] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 00, 2016, pp. 146–155.

[29] E. Richardson, M. Sela, and R. Kimmel, "3D face reconstruction by learning from synthetic data," in *Fourth International Conference on 3d Vision*, 2016, pp. 460–469.

[30] A. Tewari, M. Zollhofer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt, "Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1274–1283.

[31] E. Richardson, M. Sela, R. Or-El, and R. Kimmel, "Learning detailed face reconstruction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, 2017, pp. 5553–5562.

[32] Y. Guo, J. Zhang, J. Cai, B. Jiang, and J. Zheng, "Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.

[33] M. Castelán and J. Van Horebeek, "3D face shape approximation from intensities using partial least squares," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[34] M. Sela, E. Richardson, and R. Kimmel, "Unrestricted facial geometry reconstruction using image-to-image translation," in *Proc. IEEE Conf. Int. Conf. Comput. Vis.*, 2017, pp. 1585–1594.

[35] J. Ren, Y. Hu, Y.-W. Tai, C. Wang, L. Xu, W. Sun, and Q. Yan, "Look, listen and learna multimodal lstm for speaker identification," in *Thirtieth AAAI Conf. on Artif. Intell.*, 2016.

[36] C. Poultney, S. Chopra, Y. L. Cun *et al.*, "Efficient learning of sparse representations with an energy-based model," in *Proc. Advances in Neural Inf. Process. Systems*, 2007, pp. 1137–1144.

[37] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," *arXiv preprint arXiv:1609.03126*, 2016.

[38] W. A. S. Oswald Aldrian, "Inverse rendering of faces with a 3D morphable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1080–1093, 2013.

[39] X. Fan, R. Liu, Z. Luo, Y. Li, and Y. Feng, "Explicit shape regression with characteristic number for facial landmark localization," *IEEE Trans. on Multimedia*, vol. 20, no. 3, pp. 567–579, 2018.

[40] S. Z. Li, "Markov random field models in computer vision," in *Proc. Eur. Conf. Comput. Vis.* Springer, 1994, pp. 361–370.

[41] C. Wang, Y. Guo, J. Zhu, L. Wang, and W. Wang, "Video object co-segmentation via subspace clustering and quadratic pseudo-boolean optimization in an mrf framework," *IEEE Trans. MultiMedia*, vol. 16, no. 4, pp. 903–916, 2014.

[42] S. Xie, R. Girshick, P. Dollr, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *arXiv preprint arXiv:1611.05431*, 2016.

[43] P. J. Phillips, "Face recognition grand challenge," in *Biometric Consortium Conference*, 2004.

[44] B. Amberg, S. Romdhani, and T. Vetter, "Optimal step nonrigid icp algorithms for surface registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[45] Y. Sun, W. Zuo, and M. Liu, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019.

[46] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *Advanced video and signal based surveillance*. Ieee, 2009, pp. 296–301.

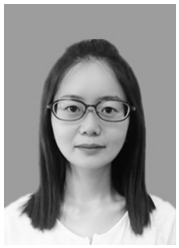
[47] F. Liu, D. Zeng, J. Li, and Q. Zhao, "Cascaded regressor based 3D face reconstruction from a single arbitrary view image," *arXiv preprint arXiv:1509.06161*, 2015.

[48] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2930–2940, 2013.

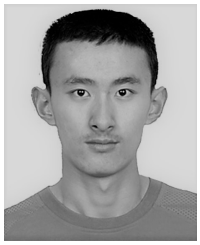
[49] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 679–692.

[50] R. Liu, S. Cheng, Y. He, X. Fan, Z. Lin, and Z. Luo, "On the convergence of learning-based iterative methods for nonconvex inverse problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

Xin Fan received the B.E. and Ph.D. degrees at Xian Jiaotong University, Xian, China, in 1998 and 2004, respectively. He was with Oklahoma State University, Stillwater, and the University of Texas Southwestern Medical Center, Dallas, from 2006 to 2009, as a post-doctoral research fellow. He joined Dalian University of Technology, Dalian, China, in 2009, where he is currently a full professor. He won the 2015 IEEE ICME Best Student Award as the corresponding author, and two papers were selected as the Finalist of the Best Paper Award at ICME 2017. His current research interests include image processing and machine vision.



Shichao Cheng received the PhD degree in mathematics from the Dalian University of Technology, Dalian, China, in 2019, and the B.E. degree in Mathematics and Applied Mathematics from Henan Normal University. Xinxiang, China, in 2013. She is currently a lecturer in Hangzhou Dianzi University. Her research interests include computer vision, machine learning and optimization.



Kang Huyan received the M.S. and B.E. degrees in software engineering at Dalian University of Technology. His main research interest is image processing.

Minjun Hou received the B.E. degree in Software Engineering from Dalian University of Technology, China, in 2017. Now she is studying as a postgraduate student in Dalian University of Technology. Her research interests include computer vision, image processing and machine learning.

Risheng Liu (M'12-) received the BSc and PhD degrees both in mathematics from the Dalian University of Technology in 2007 and 2012, respectively. He was a visiting scholar in the Robotic Institute of Carnegie Mellon University from 2010 to 2012. He served as Hong Kong Scholar Research Fellow at the Hong Kong Polytechnic University from 2016 to 2017. He is currently an associate professor with the Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Internal School of Information and Software Technology, Dalian University of Technology. His research interests include machine learning, optimization, computer vision and multimedia. He was a co-recipient of the IEEE ICME Best Student Paper Award in both 2014 and 2015. Two papers were also selected as Finalist of the Best Paper Award in ICME 2017.

Zhongxuan Luo received the B.S. degree in Computational Mathematics from Jilin University, China, in 1985, the M.S. degree in Computational Mathematics from Jilin University in 1988, and the PhD degree in Computational Mathematics from Dalian University of Technology, China, in 1991. He has been a full professor of the School of Mathematical Sciences at Dalian University of Technology since 1997. His research interests include computational geometry and computer vision.