3D Face Modeling From Diverse Raw Scan Data

Feng Liu, Luan Tran, Xiaoming Liu
Department of Computer Science and Engineering
Michigan State University, East Lansing MI 48824
{liufeng6, tranluan, liuxm}@msu.edu

Abstract

Traditional 3D face models learn a latent representation of faces using linear subspaces from limited scans of a single database. The main roadblock of building a large-scale face model from diverse 3D databases lies in the lack of dense correspondence among raw scans. To address these problems, this paper proposes an innovative framework to jointly learn a nonlinear face model from a diverse set of raw 3D scan databases and establish dense point-to-point correspondence among their scans. Specifically, by treating input scans as unorganized point clouds, we explore the use of PointNet architectures for converting point clouds to identity and expression feature representations, from which the decoder networks recover their 3D face shapes. Further, we propose a weakly supervised learning approach that does not require correspondence label for the scans. We demonstrate the superior dense correspondence and representation power of our proposed method, and its contribution to single-image 3D face reconstruction.

1. Introduction

Robust and expressive 3D face modeling is valuable for computer vision problems, e.g., 3D reconstruction [7,24,41,54] and face recognition [42,43,58], as well as computer graphics problems, e.g., character animation [15,31]. The state-of-the-art 3D face representations mostly adopt linear transformations [39,59,60], e.g., 3D Morphable Model (3DMM) or higher-order tensor generalizations [1,13,14,67], e.g., Blendshapes Model. However, these linear models fall short of capturing the nonlinear deformations such as high-frequency details and extreme expressions. Recently, with the advent of deep learning, there have been several attempts at using deep neural networks for nonlinear data-driven face modeling [4,32,51,65].

To model 3D face shapes, a large amount of high-quality 3D scans is required. The widely used 3DMM-based BFM2009 [48] is built from scans of merely 200 subjects in neutral expressions. Lack of expression may be compensated with expression bases from FaceWarehouse [14] or BD-3FE [70]. After more than a decade, almost all existing models use less than 300 training subjects. Such a small training set is far from adequate to describe the full variability of faces. Until recently, Booth et al. [11,12] build the first Large-Scale Face Model (LSFM) from neutral scans of 9,663 subjects. Unfortunately, with only the resultant linear 3DMM bases being released instead of the original scans, we cannot fully leverage this large database to explore different 3D modeling techniques.

In fact, there are many publicly available 3D face databases, as shown in Fig. 1. However, these databases are often used individually, rather than jointly to create large-scale face models. The main hurdle lies in the challenge
of estimating dense point-to-point correspondence for raw
scans, which allows these scans to be organized in the same
vector space, enabling analysis as a whole.

Dense point-to-point correspondence is one of the most
fundamental problems in 3D face modeling [22, 26], which
can be defined as in [22]: given two 3D faces $S$ and
$S'$, the correspondence should satisfy three perspectives:
i) $S$ and $S'$ have the same number of vertices; ii) the
corresponding points share the same semantic meaning; iii)
the corresponding points lie in the same local topological
triangle context. Prior dense correspondence methods [3,
7, 24, 47] lack either accuracy, robustness or automation.
Moreover, few of them have shown success on multiple
databases. Beyond of the data scale, the challenge of dense
correspondence for multiple databases is certainly escalated
over single database: the quality of scans is often inevitably
corrupted with artifacts (e.g., hair and eyebrows), missing
data and outliers; facial morphology varies significantly due
to expressions like mouth opening and closing; different
databases contain high variability on the resolution.

To address these challenges, we propose a novel
encoder-decoder to learn face models directly from
raw 3D scans of multiple diverse databases, as well as
establish dense correspondence among them. Our approach
provides: i) a PointNet-based encoder that learns nonlinear
identity and expression latent representations of 3D faces;
ii) a corresponding decoder capable of establishing dense
correspondence for scans with a variety of expressions
and resolutions; iii) the decoder can be plugged into
existing image-based encoders for 3D face reconstruction.
Specifically, by treating raw scans as unorganized point
clouds, we explore the use of PointNet [50] for converting
point clouds to identity and expression representations,
from which the decoder recovers their 3D face shapes.

However, full supervision is often not available due to
the lack of ground-truth dense correspondence. Thus, we
propose a weakly-supervised approach with a mixture of
synthetic and real 3D scans. Synthetic data with topological
ground truth helps to learn a shape correspondence prior
in a supervised fashion, which allows us to incorporate
order-invariant loss functions, e.g., Chamfer distance [21],
for unsupervised training of real data. Meanwhile, a
surface normal loss retains the original high-frequency
details. For regularization, we use the edge length loss
to encourage the triangulation topology on the template
and the reconstructed point cloud to be the same. Finally,
a Laplacian regularization loss improves the performance
of mouth regions with extreme expressions. The above
strategies allow the network to learn from a large set of raw
3D scan databases without any label on correspondences.
In summary, the contributions of this work include:

- We propose a new encoder-decoder framework that for
the first time jointly learns face models directly from raw
scans of multiple 3D face databases and establishes dense
correspondences among all scans.
- We devise a weakly-supervised learning approach and
several effective loss functions for the proposed framework
that can leverage known correspondences from synthetic
data and relax the Chamfer distance loss for vertex corre-
spondence in an unsupervised fashion.
- We demonstrate the superiority of our nonlinear model
in preserving high-frequency details of 3D scans, providing
compact latent representation, and applications of single-
image 3D face reconstruction.

### 2. Related Work

#### 3D Face Modeling

Traditional 3DMMs [7, 8] model geometry variation from limited data via PCA. Paysan
et al. [48] build BFM2009, the publicly available morphable
model in neutral expression, which is extended to emotive
face shapes [2]. Gerig et al. [24, 45] propose the Gaussian
Process Morphable Models (GPMMs) and release a new
BFM2017. Facial expressions can also be represented with
higher-order generalizations. Vlasic et al. [67] use a mul-
tilinear tensor-based model to jointly represent the identity
and expression variations. FaceWarehouse (FWH) [14] is
a popular multilinear 3D face model. The recent FLAME
model [39] additionally models head rotation. However, all
these works adopt a linear space, which is over-constrained
and might not well represent high-frequency deformations.

Deep models have been successfully used for 3D face
fitting, which recovers 3D shape from 2D images [20, 33–
35, 52, 59, 62, 72]. However, in these works the linear model
is learned a-priori and fixed during fitting, unlike ours where
the nonlinear model is learned during training.

In contrast, applying CNN to learn more powerful 3D
face models has been largely overlooked. Recently, Tran
et al. [63, 64] learn to regress 3DMM representation, along
with the decoder-based models. SfSNet [57] learns shape,
albedo and lighting decomposition of a face, from 2D
nonlinear face geometry representations directly from UV
maps via a VAE. Ranjan et al. [51] introduce a convolution-
al mesh autoencoder to learn nonlinear variations in shape and expression. Note that [4, 51] train with no more than 20 subjects and encode the 3D data to a single latent vector. Jiang et al. [32] extend [51] to decompose a 3D face into identity and expression parts. Unlike our work, these three methods require densely corresponded 3D scans in training. We summarize the comparison in Tab. 1.

3D Face Dense Correspondence. As a fundamental shape analysis task, correspondence has been well studied in the literature. Shape correspondence, a.k.a. registration, alignment or simply matching [66], finds a meaningful mapping between two surfaces. The granularity of mapping varies greatly, from semantic parts [28, 53], group [17], to points [38]. Within this range, point-to-point correspondences for 3D face is the most challenging and strict one. In original 3DMM [7], the 3D face dense correspondence is solved with a regularized form of optical flow as a cylindrical image registration task. This is only effective in constrained settings, where subjects share similar ethnicities and ages. To overcome this limitation, Patel and Smith [47] use a Thin Plate Splines (TPS) [10] warp to register scans into a template. Alternatively, Amberg et al. [3] propose an optimal step Nonrigid Iterative Closest Point (NICP) for registering 3D shapes. Booth et al. [11, 12] quantitatively compare these three popular dense correspondence techniques in learning 3DMM. Additional extensions are also proposed [22, 24, 25, 71].

Many algorithms [1, 9, 27] treat dense correspondence as a 3D-to-3D model fitting problem. E.g., [9] propose a multilinear groupwise model for 3D face correspondence to decouple identity and expression variations. Abrevayaemph et al. [1] propose a 3D face autoencoder with a CNN-based depth image encoder and multilinear model as a decoder for 3D face fitting. However, these methods require 3D faces with an initial correspondence as input and the correspondence problem is considered in the restrictive space expressed by the model. Although insightful and useful, a chicken-and-egg problem still remains unsolved [22].

To summarize, prior work tackle the problems of 3D face modeling, and 3D face dense correspondence separately. However, dense correspondence is a prerequisite for modeling. If the correspondence has errors, they will accumulate and propagate to 3D modeling. Therefore, these two problems are highly relevant and our framework for the first time tackles them simultaneously.

3. Proposed Method

This section first introduces a composite 3D face shape model with latent representations. We then present the mixture training data and our encoder-decoder network. We finally provide implementation details and face reconstruction inference. Figure 2 depicts the overview of our method.

3.1. Problem Formulation

In this paper, the output 3D face scans are represented as point clouds. Each densely aligned 3D face $S \in \mathbb{R}^{n \times 3}$ is represented by concatenating its $n$ vertex coordinates as,

$$S = [x_1, y_1, z_1; x_2, y_2, z_2; \cdots; x_n, y_n, z_n].$$

(1)

We assume that a 3D face shape is composed of identity and expression deformation parts,

$$S = S_{Id} + \Delta S_{Exp},$$

(2)

where $S_{Id}$ is the identity shape and $\Delta S_{Exp}$ is expression difference. Since the identity and expression spaces are independent, we further assume these two parts can be described by respective latent representations, $f_{Id}$ and $f_{Exp}$.

Specifically, as shown in Fig. 2, we use two networks to decode shape component $S_{Id}$ and $\Delta S_{Exp}$ from the
corresponding latent representations. Formally, given a set of raw 3D faces \( \{ \mathbf{S}_i^{raw} \}_{i=1}^N \), we learn an encoder \( E : \mathbf{S}_i^{raw} \rightarrow \mathbf{f}_{id} \), \( \mathbf{f}_{Exp} \) that estimates the identity and expression shape parameters \( \mathbf{f}_{id} \in \mathbb{R}^{l_{id}}, \mathbf{f}_{Exp} \in \mathbb{R}^{l_{Exp}} \), an identity shape decoder \( D_{id} : \mathbf{f}_{id} \rightarrow \mathbf{S}_{id} \), and an expression shape decoder \( D_{Exp} : \mathbf{f}_{Exp} \rightarrow \mathbf{S}_{Exp} \) that decode the shape parameters to a 3D shape estimation \( \mathbf{S} \).

Recent attempts to encode 3D face shape in deep learning include point clouds, depth map [1], UV map based mesh [4], and mesh surface [32, 51]. Point clouds are a standard and popular 3D face acquisition format used by Kinect, iPhone’s face ID and structured light scanners. We thus design a deep encoder-decoder framework to directly consume unorganized point sets as input and output densely corresponded 3D shapes. Before providing the algorithm details, we first introduce the real and synthetic training data served for the weakly-supervised learning.

### 3.2. Training Data

To learn a robust and highly variable 3D face model, we construct training data of seven publicly available 3D databases with a wide variety of identity, age, ethnicity, expression and resolution, listed in Tab. 2. However, for these real scans, there are no associated ground-truth on dense correspondence. Recently, some 3D databases are released such as 4DFAB [16], Multi-Dim [40] and UHDB 3D [61,68]. While including them may increase the amount of training data, they do not provide new types of variations beyond the seven databases. We do not use the occlusion and pose (self-occlusion) data of Bosphorus database, since extreme occlusion or missing data would break semantic correspondence consistency of 3D faces. For BU4DFE database, we manually select one neutral and 24, expression scans per subject. To keep the balance between real and synthetic data, we use BFM2009 to synthesize 3D faces of 1,500 subjects, and use 3DDFA [73] expression model to generate 6 random expressions for each subject. Figure 3 shows one example scan from each of the eight databases.

**Preprocessing and data augmentation** As visualized in Fig. 4, we first predefine a template of 3D face topology consisting of \( n = 29,495 \) vertices and 58,366 triangles, which is manually cropped from BFM mean shape. Then, we normalize the template into a unit sphere. The original synthetic examples contain 53,215 vertices, after removing points on the tongue. For synthetic examples, we crop their face region with same topological triangulation as the template, perform the same normalization, and denote this resultant 3D face set with ground-truth triangulation as \( \{ \mathbf{S}_i^{gt} \}_{i=1}^M \) whose number of vertices is also \( n \).

Since raw scans are acquired from different distances, orientations or sensors, their point clouds exhibit enormous variations in pose and scale. Thus, before feeding them to our network, we apply a similarity transformation to align raw scans to the template by using five 3D landmarks. Following [11], we detect 2D landmarks on the corresponding rendered images, from which we obtain 3D landmarks by back-projection (Fig. 4 (1)). After alignment, the points outside the unit sphere are removed. Finally, we randomly sample \( n \) points as the input \( \mathbf{S}_i^{input} \in \mathbb{R}^{n \times 3} \). If the vertex number is less than \( n \), we apply interpolating subdivision [37] before sampling. As in Tab. 2, we perform data augmentation for neutral scans by repeating random sampling several times so that each subject has 10 neutral training scans. Note that the above preprocessing is also applied to synthetic data, except that their 3D landmarks are provided by BFM. As a result, the point ordering of both input raw and synthetic data is random.

### 3.3. Loss Function

This encoder-decoder architecture is trained end-to-end. We define three kinds of losses to constrain the correspon-
dence of the output shape and template, also to retain the original global and local information. The overall loss is:

\[ \mathcal{L} = \mathcal{L}^{vt} + \lambda_1 \mathcal{L}^{normal} + \lambda_2 \mathcal{L}^{edge}, \]  

(3)

where the vertex loss \( \mathcal{L}^{vt} \) is to constrain the location of mesh vertices, normal loss \( \mathcal{L}^{normal} \) is to enforce the consistency of surface normals, and edge length loss is to preserve the topology of 3D faces.

Here, we consider two training scenarios: synthetic and real data. Supervision is typically available for the synthetic data with ground truth (supervised case), but real scans are obtained without correspondence label (unsupervised case).

**Supervised loss** In the supervised case, given the shape \( S^{gt} \) (and \( \hat{S} \)) and predefined triangle topology, we can easily compute the corresponding surface normal \( \hat{n}^{gt} \) (and \( n \)) and edge length \( e^{gt} \) (and \( \hat{e} \)). Therefore, for vertex loss, we can use \( L_1 \) loss \( \mathcal{L}^{vt}(S, S^{gt}) = ||S^{gt} - S||_1 \). We measure the normal loss by cosine similarity distance \( \mathcal{L}^{normal}(\hat{n}, n^{gt}) = \frac{1}{n} \sum_i (1 - \hat{n}^{gt} \cdot n_i) \). If the predicted normal has a similar orientation as the ground truth, the dot-product \( \hat{n}^{gt} \cdot n_i \) will be close to 1 and the loss will be small, and vice versa. The third term \( \mathcal{L}^{edge} \) encourages the ratio between edges length in the predicted shape and ground truth to be close to 1. Following [28], edge length loss is defined as,

\[ \mathcal{L}^{edge}(\hat{S}, S^{gt}) = \frac{1}{|E|} \sum_{(i,j) \in E} \left| \frac{||S_i - S_j||}{||S^{gt}_i - S^{gt}_j||} - 1 \right|, \]  

(4)

where \( E \) is the fixed edge graph of the template.

**Unsupervised loss** In the case where the correspondences between the template and real scans are not available, we still optimize the reconstructions, but regularize the deformations toward correspondence. For reconstruction, we use the Chamfer distance as the \( \mathcal{L}^{vt}(\hat{S}, S^{raw}) \) between the input scans \( S^{raw} \) and the predicted \( \hat{S} \).

\[ \mathcal{L}^{vt}(\hat{S}, S^{raw}) = \frac{1}{|S^{raw}|} \sum_{p \in S^{raw}} \min_{q \in \hat{S}} ||p - q||_2^2 + \frac{1}{|\hat{S}|} \sum_{q \in \hat{S}} \min_{p \in S^{raw}} ||p - q||_2^2, \]  

(5)

where \( p \) is a vertex in the predicted shape, \( q \) is a vertex in the input scan. When \( \min_{p \in S^{raw}} ||p - q||_2^2 > \epsilon \) or \( \min_{p \in S^{raw}} ||p - q||_2^2 > \epsilon \), we treat \( q \) as a flying vertex and the error will not be counted.

In this unsupervised case, we further define loss on the surface normal to characterize high-frequency properties, \( \mathcal{L}^{normal}(\hat{n}, n^{raw}_q) \), where \( q \) is the closest vertex for \( p \) that is found when calculating the Chamfer distance, and \( n^{raw}_q \) is the observed normal from the real scan. For the edge length loss, \( \mathcal{L}^{edge} \) is defined the same as Eqn. 4.

**Refine** In the unsupervised case, the normal loss \( \mathcal{L}^{normal}(\hat{n}, n^{raw}_q) \) always find the closest vertex \( q \) in \( S^{raw} \).
The ideal evaluation metric for presentation power, and single-image face reconstruction. dense correspondence accuracy, shape and expression representation is.

4. Experimental Results

Experiments are conducted to evaluate our method in dense correspondence accuracy, shape and expression representation power, and single-image face reconstruction.

Evaluation Metric The ideal evaluation metric for 3D shape analysis is per-vertex error. However, this metric is not applicable to evaluating real scans due to the absence of dense correspondence ground truth. An alternative metric is per-vertex fitting error, which has been widely used in 3D face reconstruction and 3D face-to-face fitting, e.g., LSFM [11], GPMMs [45]. The per-vertex fitting error is the distance between every vertex of the test shape and the nearest-neighbor vertex of the corresponding estimated shape. Generally, the value of this error could be very small due to the nearest-neighbor search. Thus, it sometimes can not faithfully reflect the accuracy of dense correspondence.

To better evaluate the correspondence accuracy, prior 3D face correspondence works [22, 44, 55] adopt a semantic landmark error. With the pre-labeled landmarks on the template face, it is easy to find p 3D landmarks \( \{l_i\}_{i=1}^{p} \) with the same indexes on the estimated shape. By comparing with manual annotations \( \{\hat{l}_i\}_{i=1}^{p} \), we can compute the semantic landmark error by \( \frac{1}{p} \sum_{i=1}^{p} \| l_i - \hat{l}_i \| \). Note that, this error is normally much larger than per-vertex fitting error due to inconsistent and imprecise annotations. Tab. 6 compares these three evaluation metrics.

4.1 Ablation Study

We qualitatively evaluate the function of each loss component. As seen in Fig. 6, only using vertex loss severely impairs the surface smoothness and local details; adding surface normal loss preserves the high-frequency details. adding edge length term refines the local triangle topology. These results demonstrate that all the loss components presented in this work contribute to the final performance.

4.2 Dense Correspondence Accuracy

We first report the correspondence accuracy on BU3DFE database. BU3DFE contains one neutral and six expression scans with four levels of strength, for each of 100 subjects. Following the same setting in [55] and GPMMs [24], we use all \( p = 83 \) landmarks of all neutral scans and expression scans in the highest level for evaluation. Specifically, the landmarks of the estimated shape are compared to the manually annotated landmarks that are provided with BU3DFE. We compare with four state-of-the-art dense correspondence methods, NICP [3], Bolkart et al. [9], Salazar et al. [55], and GPMMs [24]. Among them, NICP has been widely used for constructing neutral morphable model such as BFM 2009 and LSFM. For a fair comparison, we re-implement NICP with extra landmark constraint so that it can establish dense correspondence for expressive

Figure 6: Qualitative results reflecting the contribution of loss components. The first column is the input scan. Column 2-4 show the reconstructed shapes with different loss combination.

Figure 7: Raw scans (top) and their reconstructions with color-coded dense correspondences (bottom), for one BU3DFE subject in seven expressions: angry (AN), disgust (DI), fear (FE), happy (HA), neutral (NE), sad (SA), and surprise (SU).
3D scans. For the other three methods, we report results from their papers. Both Salazar et al. [55] and Bolkart et al. [9] are multilinear model based 3D face fitting method. GPMMs [24] is a recent Gaussian process registration based method. Note these four baselines do require labeled 3D landmarks as input, while our method does not.

To further evaluate the generalization ability of the proposed method for new scan data, we conduct two series of experiments: (i) training using data from BU3DFE database, denoted as Proposed (in), and (ii) training using data outside BU3DFE database, denoted as Proposed (out).

As shown in Tab. 3, the Proposed (in) setting significantly reduces errors by at least 21.2\% w.r.t. the best baseline. These results demonstrate the superiority of the proposed method in dense correspondence. The error of Proposed (out) setting shows a small increase, but is still lower than the baselines. The relatively high semantic landmark error is attributed by the imprecise manual annotations, especially on the semantic ambiguity contour, i.e., Chin, Left Face and Right Face. Some example dense correspondence results are shown in Fig. 7 and Supp.

We further compare semantic landmark error with the very recent SOTA correspondence method [22], which is an extension of ICP-based method, on the high-resolution FRGC v2.0 database [49]. We also compare with two 3D landmark localization works [18, 74]. Following the same setting in [22], we compute the mean and standard deviation of \( p = 10 \) landmarks for 4,007 scans. The results of the baseline methods are from their papers. As shown in Tab. 4, our method improves the SOTA [22] by 11.4\%, and preserves high-frequency details for high-resolution 3D models (see Fig. 7 of Supp.). The landmark errors are much smaller than BU3DFE since the annotations used here are more accurate than BU3DFE’s. Thanks to the offline training process, our method is two order of magnitude faster than the existing dense correspondence methods: 0.26s (2ms with GPU) vs. 57.48s of [3] vs. 164.60s of [22].

### 4.3. Representation Power

**Identity shape** We compare the capabilities of the proposed 3D face models with linear and nonlinear 3DMMs on BFM. The BFM database provides 10 test face scans, which are not included in the training set. As these scans are already established dense correspondence, we use the per-vertex error for evaluation. For fair comparison, we train different models with different latent space sizes. As shown in Tab. 5, the proposed model has smaller reconstruction error than the linear or nonlinear models. Also, the proposed models are more compact. They can achieve similar performances as linear and nonlinear models whose latent spaces sizes are doubled. Figure 8 shows the visual quality of three models’ reconstruction.
Expression shape  We compare the expression representation power of the proposed 3D face models with 3DDFA expression model [73], a 29-dim model originated from FaceWarehouse [14]. We use a 79-dim expression model from [29] to randomly generate an expression difference with Gaussian noise for each BFM test sample. Those data are treated as the test set. For a fair comparison, we train a model with the same expression latent space size ($l_{Exp}=29$). Our model has significantly smaller per-vertex error than 3DDFA: 1.424mm vs. 2.609mm. Figure 9 shows the visual quality of four scans’ reconstructions.

Shape representation on BU3DFE and BFM Table 6 compares the shape expressiveness of our model with the three different metrics. Following the setting in Tab. 5, we further calculate the per-vertex fitting error and semantic landmark error ($p=51$) for BFM test samples. We also provide the per-vertex fitting error for the BU3DFE reconstructions in Tab 3. From Tab. 6, compared to the ideal per-vertex error, semantic landmark error is much larger while per-vertex fitting error is smaller.

Shape representation on COMA We further evaluate our shape representation on a large-scale COMA database [51]. For a fair comparison with FLAME [39] and Jiang et al. [32], we follow the same setting as [32], and set our latent vector size as 4 for identity and 4 for expression. As in Tab. 7, our method shows better shape representation compared to SOTA methods. While MeshAE [51] achieves a smaller error (1.160mm), comparing ours with it is not fair, as it has the advantage of encoding 3D faces into a single vector without decomposing into identity and expression. Also, their mesh convolution requires densely corresponded 3D scans as input.

4.4. Single-image 3D Face Reconstruction

With the same setting in [59], we quantitatively compare our single-image shape inference with prior works on nine subjects (180 images) of the FaceWarehouse database. Visual and quantitative comparisons are shown in Fig. 10. We achieve on-par results with nonlinear 3DMM [64], Tewari [59] and Garrido et al. [23], while surpassing all other CNN-based regression methods [52, 62].

5. Conclusions

This paper proposes an innovative encoder-decoder to jointly learn a robust and expressive face model from a diverse set of raw 3D scan databases and establish dense correspondence among all scans. By using a mixture of synthetic and real 3D scan data with an effective weakly-supervised learning-based approach, our network can preserve high-frequency details of 3D scans. The comprehensive experimental results show that the proposed method can effectively establish point-to-point dense correspondence, achieve more representation power in identity and expression, and is applicable to 3D face reconstruction.

Table 6: Evaluation metric comparison on two databases.

<table>
<thead>
<tr>
<th>Metric</th>
<th>BFM</th>
<th>BU3DFE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Per-vertex fitting error</td>
<td>0.572mm</td>
<td>1.065mm</td>
</tr>
<tr>
<td>Per-vertex error</td>
<td>0.946mm</td>
<td>1.149mm</td>
</tr>
<tr>
<td>Semantic landmark error</td>
<td>1.493mm</td>
<td>5.140mm</td>
</tr>
</tbody>
</table>

Table 7: Comparison (per-vertex error, mm) with state-of-the-art 3D face modeling methods on COMA database.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Proposed</th>
<th>Jiang et al. [32]</th>
<th>FLAME [39]</th>
</tr>
</thead>
<tbody>
<tr>
<td>barieith</td>
<td>1.609</td>
<td>1.695</td>
<td>2.002</td>
</tr>
<tr>
<td>checks</td>
<td>1.561</td>
<td>1.706</td>
<td>2.011</td>
</tr>
<tr>
<td>eyebrow</td>
<td>1.400</td>
<td>1.475</td>
<td>1.862</td>
</tr>
<tr>
<td>high smile</td>
<td>1.556</td>
<td>1.714</td>
<td>1.960</td>
</tr>
<tr>
<td>lips back</td>
<td>1.532</td>
<td>1.752</td>
<td>2.047</td>
</tr>
<tr>
<td>lips up</td>
<td>1.529</td>
<td>1.747</td>
<td>1.983</td>
</tr>
<tr>
<td>mouth down</td>
<td>1.362</td>
<td>1.655</td>
<td>2.029</td>
</tr>
<tr>
<td>mouth extreme</td>
<td>1.442</td>
<td>1.551</td>
<td>2.028</td>
</tr>
<tr>
<td>mouth middle</td>
<td>1.383</td>
<td>1.757</td>
<td>2.043</td>
</tr>
<tr>
<td>mouth open</td>
<td>1.381</td>
<td>1.393</td>
<td>1.894</td>
</tr>
<tr>
<td>mouth side</td>
<td>1.502</td>
<td>1.748</td>
<td>2.090</td>
</tr>
<tr>
<td>mouth up</td>
<td>1.426</td>
<td>1.528</td>
<td>2.067</td>
</tr>
<tr>
<td>Avg</td>
<td>1.474</td>
<td>1.643</td>
<td>1.615</td>
</tr>
</tbody>
</table>

Figure 9: Expression representation power comparison. Our results better match the expression deformations than 3DDFA.

Figure 10: Quantitative evaluation of single-image 3D face reconstruction on samples of FaceWarehouse database.
References


[16] Shiyang Cheng, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. 4DFAB: A large scale 4D database for facial expression analysis and biometric applications. In CVPR, 2018. 4


[27] Carl Martin Grewe and Stefan Zachow. Fully automated and highly accurate dense correspondence for facial surfaces. In ECCV, 2016. 3


[31] Xiaolong Han, Chang Gao, and Yizhou Yu. DeepS-ketch2Face: a deep learning based sketching system for 3D face and caricature modeling. TOG, 36(4):126, 2017. 1


[63] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3D face morphable model. In CVPR, 2019. 2

[64] Luan Tran and Xiaoming Liu. Nonlinear 3D face morphable model. In CVPR, 2018. 2, 7, 8

[65] Luan Tran and Xiaoming Liu. On learning 3D face morphable model from in-the-wild images. TPAMI, 2019. doi:10.1109/TPAMI.2019.2927975. 1


[69] Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. A high-resolution 3D dynamic facial expression database. In FG, 2008. 4

[70] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3D facial expression database for facial behavior research. In FG, 2006. 1, 4


[73] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In CVPR, 2015. 4, 8