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

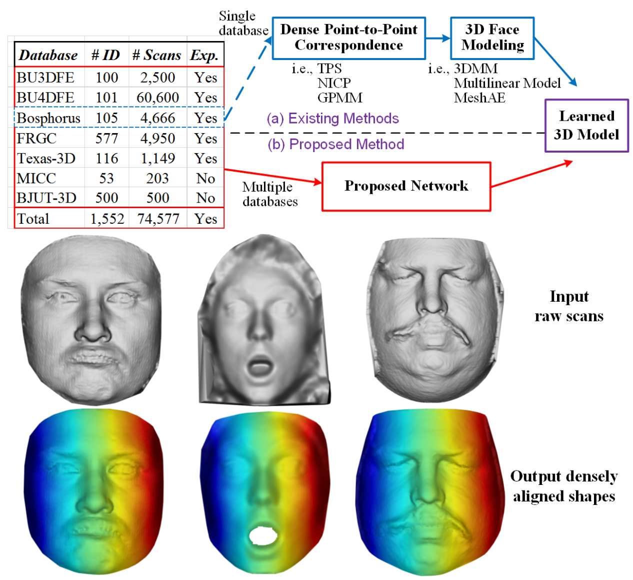
Abstract

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.



Database	*# ID*	*# Scans*	*Exp.*
BU3DFE	100	2,500	Yes
BU4DFE	101	60,600	Yes
Bosphorus	105	4,666	Yes
FRGC	577	4,950	Yes
Texas-3D	116	1,149	Yes
MICC	53	203	No
BJUT-3D	500	500	No
Total	1,552	74,577	Yes

Figure 1: Comparison between 3D face modeling of (a) existing methods and (b) our proposed method. Dense point-to-point correspondence is a pre-requisite for the existing 3D face modeling methods. Our proposed CNN-based approach learns face models *directly* from raw scans of *multiple* 3D face databases and establish dense point-to-point correspondence among all scans (best viewed in color). Despite the diversity of scans in resolution and expression, our model can express the fine level of details.

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 correspondence 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.

Table 1: Comparison of 3D face modeling from scans. 'Exp.' refers to whether learns the expression latent space, 'Corr.' refers to whether requires densely corresponded scans in training.

Method	Dataset	Lin./nonL.	#Subj.	Exp.	Corr.
BFM [48]	BFM	Linear	200	No	Yes
GPMMs [45]	BFM	Linear	200	Yes	Yes
LSFM [11, 12]	LSFM	Linear	9,663	No	Yes
LYHM [19]	LYHM	Linear	1,212	No	Yes
Multil. model [14]	FWH	Linear	150	Yes	Yes
FLAME [39]	CAESAR D3DFACS	Linear	3,800 10	Yes	Yes
VAE [4]	Proprietary	Nonlin.	20	No	Yes
MeshAE [51]	COMA	Nonlin.	12	No	Yes
Jiang *et al.* [32]	FWH	Nonlin.	150	Yes	Yes
Proposed	**7 datasets**	**Nonlin.**	**1,552**	**Yes**	**No**

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 multilinear 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 images, instead of 3D scans. Bagautdinov *et al.* [4] learn nonlinear face geometry representations directly from UV maps via a VAE. Ranjan *et al.* [51] introduce a convolution-

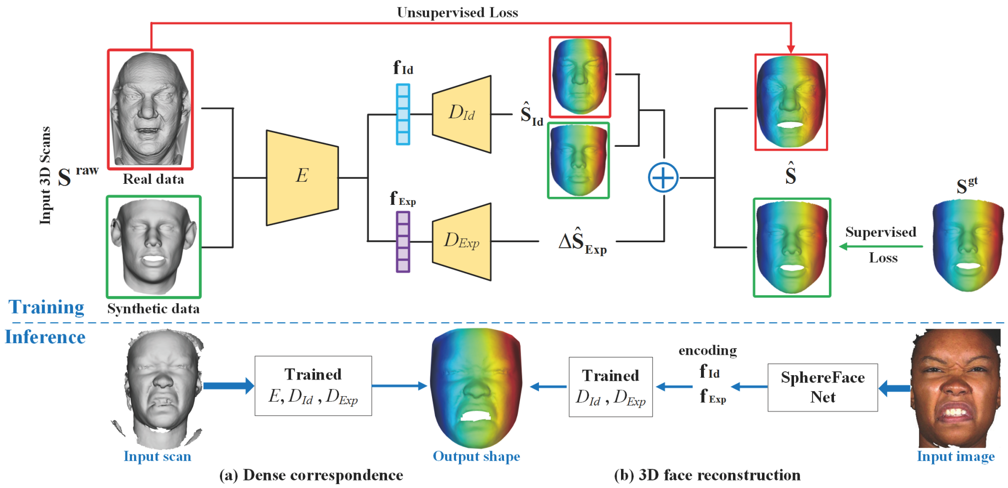


Figure 2: Overview of our 3D face modeling method. A mixture of synthetic and real data is used to train the encoder-decoder network with supervised (green) and unsupervised (red) loss. Our network can be used for 3D dense correspondence and 3D face reconstruction.

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 $\mathbf{S} \in \mathbb{R}^{n \times 3}$ is represented by concatenating its n vertex coordinates as,

$$\mathbf{S} = [x_1, y_1, z_1; x_2, y_2, z_2; \cdots; x_n, y_n, z_n]. \quad (1)$$

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

$$\mathbf{S} = \mathbf{S}_{Id} + \Delta\mathbf{S}_{Exp}, \quad (2)$$

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

Specifically, as shown in Fig. 2, we use two networks to decode shape component $\mathbf{S}_{Id}$ and $\Delta\mathbf{S}_{Exp}$ from the

Table 2: Summary of training data from related databases.

Database	#Subj.	#Neu.	#Sample	#Exp.	#Sample
BU3DFE [70]	100	100	1,000	2,400	2,400
BU4DFE [69]	101	>101	1,010	>606	2,424
Bosphorus [56]	105	299	1,495	2,603	2,603
FRGC [49]	577	3,308	6,616	1,642	1,642
Texas-3D [30]	116	813	1,626	336	336
MICC [5]	53	103	515	–	–
BJUT-3D [6]	500	500	5,000	–	–
Real Data	1,552	5,224	17,262	7,587	9,405
Synthetic Data	1,500	1,500	15,000	9,000	9,000

corresponding latent representations. Formally, given a set of raw 3D faces $\{\mathbf{S}_i^{raw}\}_{i=1}^N$, we learn an encoder E : $\mathbf{S}^{raw} \rightarrow \mathbf{f}_{Id}, \mathbf{f}_{Exp}$ 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 $\hat{\mathbf{S}}$.

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

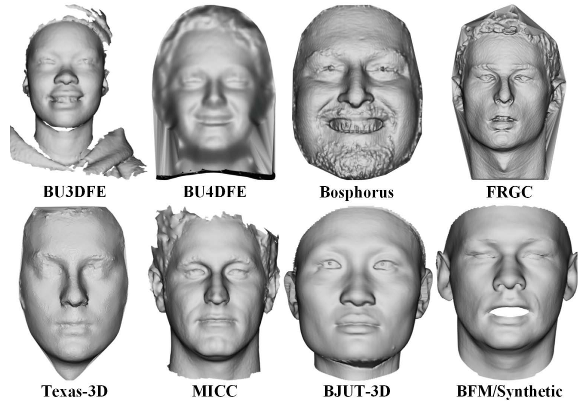


Figure 3: One sample from each of eight 3D face databases. They exhibit a wide variety of expressions and resolutions.

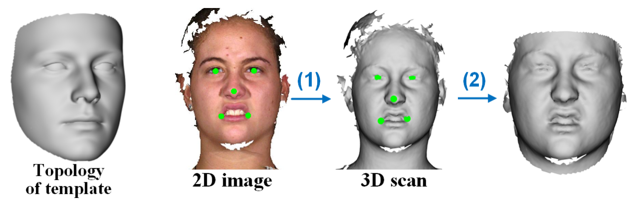


Figure 4: Preprocessing. (1) Automatic 3D landmark detection based on rendered images. (2) Guided by landmarks and predefined template, the 3D scan is alignment and cropping.

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 correspondence 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 correspondence of the output shape and template, also to retain the original global and local information. The overall loss is:

$$\mathcal{L} = \mathcal{L}^{vt} + \lambda_1 \mathcal{L}^{normal} + \lambda_2 \mathcal{L}^{edge}, \quad (3)$$

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 $\mathbf{S}^{gt}$ (and $\hat{\mathbf{S}}$) and predefined triangle topology, we can easily compute the corresponding surface normal $\mathbf{n}^{gt}$ (and $\hat{\mathbf{n}}$) and edge length $\mathbf{e}^{gt}$ (and $\hat{\mathbf{e}}$). Therefore, for vertex loss, we can use $\mathcal{L}_1$ loss $\mathcal{L}^{vt}(\hat{\mathbf{S}}, \mathbf{S}^{gt})$=$||\mathbf{S}^{gt} - \hat{\mathbf{S}}||_1$. We measure the normal loss by cosine similarity distance $\mathcal{L}^{normal}(\hat{\mathbf{n}}, \mathbf{n}^{gt}) = \frac{1}{n}\sum_i(1 - \mathbf{n}_i^{gt} \cdot \hat{\mathbf{n}}_i)$. If the predicted normal has a similar orientation as the ground truth, the dot-product $\mathbf{n}_i^{gt} \cdot \hat{\mathbf{n}}_i$ 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,

$$\mathcal{L}^{edge}(\hat{\mathbf{S}}, \mathbf{S}^{gt}) = \frac{1}{\#E} \sum_{(i,j)\in E} \left| \frac{\left\|\hat{\mathbf{S}}_i - \hat{\mathbf{S}}_j\right\|}{\left\|\mathbf{S}_i^{gt} - \mathbf{S}_j^{gt}\right\|} - 1 \right|, \quad (4)$$

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 $\mathcal{L}^{vt}(\hat{\mathbf{S}}, \mathbf{S}^{raw})$ between the input scans $\mathbf{S}^{raw}$ and the predicted $\hat{\mathbf{S}}$,

$$\mathcal{L}^{vt}(\hat{\mathbf{S}}, \mathbf{S}^{raw}) = \sum_{p\in\hat{\mathbf{S}}} \min_{q\in\mathbf{S}^{raw}} \|p - q\|_2^2 + \sum_{q\in\mathbf{S}^{raw}} \min_{p\in\hat{\mathbf{S}}} \|p - q\|_2^2, \quad (5)$$

where p is a vertex in the predicted shape, q is a vertex in the input scan. When $\min_{q\in\mathbf{S}^{raw}} \|p - q\|_2^2 > \epsilon$ or $\min_{p\in\hat{\mathbf{S}}} \|p - q\|_2^2 > \epsilon$, 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, $\mathcal{L}^{normal}(\hat{\mathbf{n}}, \mathbf{n}_{(q)}^{raw})$, where q is the closest vertex for p that is found when calculating the Chamfer distance, and $\mathbf{n}_{(q)}^{raw}$ 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 $\mathcal{L}^{normal}(\hat{\mathbf{n}}, \mathbf{n}_{(q)}^{raw})$ always find the closet vertex q in $\mathbf{S}^{raw}$.

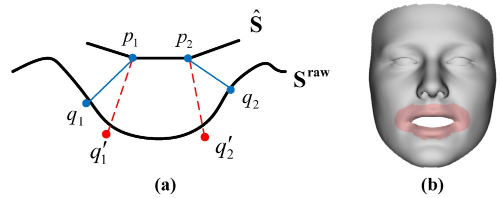


Figure 5: (a) q_i is the closet vertex of p_i and q_i' is computed by the normal ray scheme. (b) the predefined mouth region.

The disadvantage of this closest vertex scheme is that the counterpart q_i is not necessary the true target for correspondence in high-curvature regions (Fig. 5 (a)). Therefore, the loss is not capable of capturing high-frequency details in $\mathbf{S}^{raw}$. To remedy this issue, as suggested in [46], we consider the normal ray method which computes the closest point of intersection of the normal ray originated from p_i with $\mathbf{S}^{raw}$. As shown in Fig. 5 (a), the normal ray in sharp regions would find a better counterpart q_i'. At the early stage of the training process, we use the closet vertex scheme which is computationally more efficient. When the loss gets saturated, we switch to use the normal ray scheme.

For many expressive scans, some tongue points have been recorded when the mouth is open, which are hard to establish correspondence. To address this issue, on top of the losses in Eqn. 3, we add a mouth region Laplacian regularization loss $\mathcal{L}^{lap} = \|L\mathbf{S}_{mouth}\|_2$ to maintain relative location between neighboring vertices. Here L is the discrete Laplace-Beltrami operator and $\mathbf{S}_{mouth}$ is the mouth region vertex as predefined in Fig. 5 (b). See [36] for details on Laplacian regularization loss.

3.4. Implementation Detail

Encoder/Decoder Network We employ the PointNet [50] architecture as the base encoder. As shown in Fig. 2, the encoder takes the 1024-dim output of PointNet and appends two parallel fully-connected (FC) layers to generate identity and expression latent representations. We set l_{Id}=l_{Exp}=512. The decoders are two-layer MLP networks, whose numbers of inputs and outputs are respectively $\{l_{Id}(l_{Exp}), 1024\}$ (ReLU), $\{1024, n\times3\}$.

Training Process We train our encoder-decoder network in three phases. First, we train the encoder and identity decoder with neutral examples. Then, we fix the identity decoder and train the expression decoder with expression examples. Finally, the end-to-end joint training is conducted. In the first two phases, we start the training with only synthetic data. When the loss gets saturated (usually in 10 epochs), we continue training using a mixture of synthetic and real data for another 10 epochs. We optimize the networks via Adam with an initial learning rate of 0.0001. The learning rate is decreased by half every 5 epochs. We

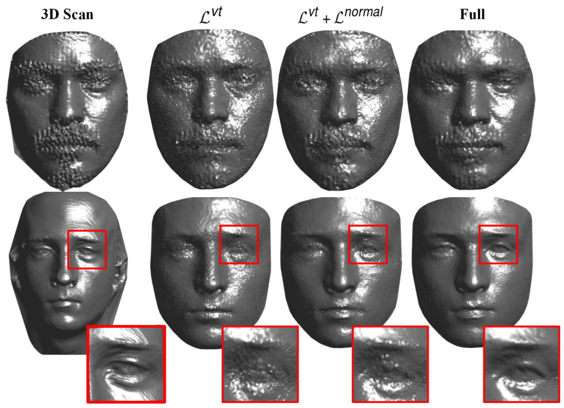


Figure 6: Qualitative results reflecting the contribution of loss components. The first column is the input scan. Column 2-4 show the reconstructed shapes with different loss combination.

explore the different batch sizes including strategies such as 50% synthetic and 50% real data per batch, and find the optimal batch size to be 1. λ_1 and λ_2 control the influence of regularizations against $\mathcal{L}^{vt}$. They are both set to 1.6×10^{-4} in our experiments. We set ϵ=0.001 and the weight of $\mathcal{L}^{lap}$ is 0.005.

3.5. Single-Image Shape Inference

As shown in Fig. 2, our identity and expression shape decoders can be used for image-to-shape inference. Specifically, we employ a SOTA face feature extraction network SphereFace [44] as the base image encoder. This network consists of 20 convolutional layers and FC layer, and takes the 512-dim output of the FC layer as the face representation. We append another two parallel FC layers to generate the identity and expression latent representations, respectively. Here, we use the raw scans from the 7 real databases to render images as our training samples. With the learnt ground truth identity and expression latent codes, we employ a $\mathcal{L}_1$ latent loss to fine-tune this image encoder. Since the encoder excels in face feature extraction and latent loss has strong supervision, the encoder is fine-tuned for 100 epochs with the batch size of 32.

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 [43] 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.

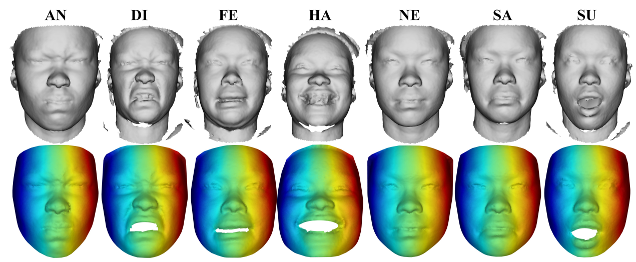


Figure 7: Raw scans (top) and their reconstructions with color-coded dense correspondences (bottom), for one BU3DFE subject in seven expressions: angry (AN), disgust (DI), fear (FE), happy (HA), neutral (NE), sad (SA), and surprise (SU).

To better evaluate the correspondence accuracy, prior 3D face correspondence works [22, 24, 55] adopt a ***semantic landmark error***. With the pre-labeled landmarks on the template face, it is easy to find p 3D landmarks $\{\hat{l}_i\}_{i=1}^p$ with the same indexes on the estimated shape. By comparing with manual annotations $\{l_i^*\}_{i=1}^p$, we can compute the semantic landmark error by $\frac{1}{p}\sum_{i=1}^p \left\| l_i^* - \hat{l}_i \right\|$. 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

Table 3: Comparison of the mean and standard deviation of semantic landmark error (mm) on BU3DFE.

Face Region	NICP [3]	Bolkart *et al.* [9]	Salaza *et al.* [55]	GPMMs [24]	Proposed (**out**)	Proposed (**in**)	Relative Impr.
Left Eyebrow	7.49±2.04	8.71±3.32	6.28±3.30	4.69±4.64	6.25±2.58	**4.18±1.62**	10.9%
Right Eyebrow	6.92±2.39	8.62±3.02	6.75±3.51	5.35±4.69	4.57±3.03	**3.97±1.70**	25.8%
Left Eye	3.18±0.76	3.39±1.00	3.25±1.84	3.10±3.43	2.00±1.32	**1.72±0.84**	44.5%
Right Eye	3.49±0.80	4.33±1.16	3.81±2.06	3.33±3.53	2.88±1.29	**2.16±0.82**	35.1%
Nose	5.36±1.39	5.12±1.89	3.96±2.22	3.94±2.58	4.33±1.24	**3.56±1.08**	9.6%
Mouth	5.44±1.50	5.39±1.81	5.69±4.45	**3.66±3.13**	4.45±2.02	4.17±1.70	-13.9%
Chin	12.40±6.15	11.69±6.39	7.22±4.73	11.37±5.85	7.47±3.01	**6.80±3.24**	5.8%
Left Face	12.49±5.51	15.19±5.21	18.48±8.52	12.52±6.04	12.10±4.06	**9.48±3.42**	24.1%
Right Face	13.04±5.80	13.77±5.47	17.36±9.17	10.76±5.34	13.17±4.54	**10.21±3.07**	5.1%
Avg.	7.56±3.92	8.49±4.29	8.09±5.75	6.52±3.86	6.36±3.92	**5.14±3.03**	21.2%

Table 4: Comparison of semantic landmark error (mean+STD in mm) on FRGC v2.0. The landmarks are defined in [22].

Landmark	Creusot *et al.* [18]	Gilant *et al.* [74]	Fan *et al.* [22]	Proposed
ex(L)	5.87±3.11	4.50±2.97	2.62±1.54	**1.79±1.01**
en(L)	4.31±2.44	3.12±2.09	2.53±1.66	**1.61±0.97**
n	4.20±2.07	3.63±2.02	**2.43±1.36**	2.69±1.43
ex(R)	6.00±3.03	3.74±2.79	2.60±1.71	**2.00±0.95**
en(R)	4.29±2.03	2.73±2.14	2.49±1.65	**1.82±0.93**
prn	3.35±2.00	2.68±1.48	**2.11±1.17**	2.36±1.37
Ch(L)	5.47±3.45	5.31±2.05	2.93±2.14	**2.58±2.61**
Ch(R)	5.64±3.58	4.38±2.08	2.84±2.17	**2.60±2.58**
ls	4.23±3.21	3.31±2.65	**2.35±2.86**	2.75±2.77
li	5.46±3.29	4.02±3.80	4.35±3.93	**4.02±3.97**
Avg.	4.88±0.91	3.74±0.83	2.73±0.62	**2.42±0.70**

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 baselines are from their papers. As shown in Tab. 4, our method improves the SOTA [22] by 11.4%, and preserves

Table 5: 3D scan reconstruction comparison (per-vertex error, mm). l_{Id} denotes the dimension of latent representation.

l_{Id}	40	80	160
Linear 3DMM [48]	1.669	1.450	1.253
Nonlinear 3DMM [64]	1.440	1.227	1.019
Proposed	**1.258**	**1.107**	**0.946**

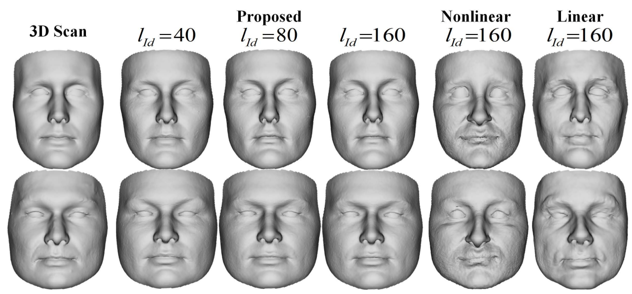


Figure 8: Shape representation power comparison. Our reconstructions closely match the face shapes and the higher-dim latent spaces can capture more local details.

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.

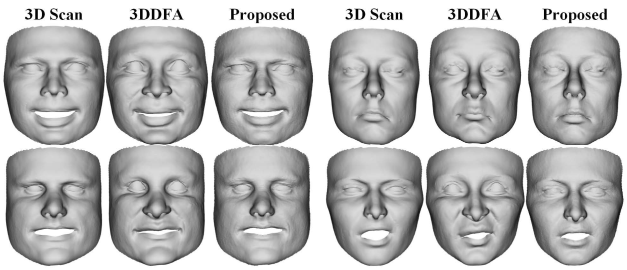


Figure 9: Expression representation power comparison. Our results better match the expression deformations than 3DDFA.

Table 6: Evaluation metric comparison on two databases.

Metric	BFM	BU3DFE
Per-vertex fitting error	$0.572mm$	$1.065mm$
Per-vertex error	$0.946mm$	–
Semantic landmark error	$1.493mm$	$5.140mm$

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

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

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

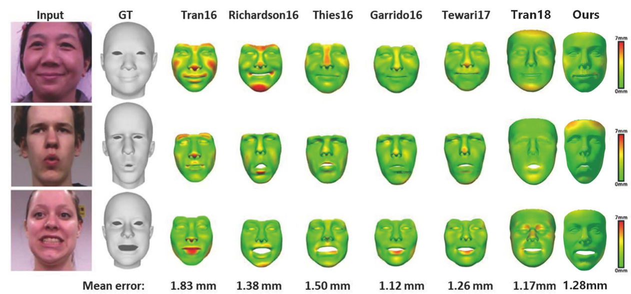


Figure 10: Quantitative evaluation of single-image 3D face reconstruction on samples of FaceWarehouse database.

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.

Acknowledgment Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-18-1-0330. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Victoria Fernández Abrevaya, Stefanie Wuhrer, and Edmond Boyer. Multilinear autoencoder for 3D face model learning. In *WACV*, 2018. 1, 3, 4

[2] Brian Amberg, Reinhard Knothe, and Thomas Vetter. Expression invariant 3D face recognition with a morphable model. In *FG*, 2008. 2

[3] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *CVPR*, 2007. 2, 3, 6, 7

[4] Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. Modeling facial geometry using compositional VAEs. In *CVPR*, 2018. 1, 2, 3, 4

[5] Andrew D Bagdanov, Alberto Del Bimbo, and Iacopo Masi. The florence 2D/3D hybrid face dataset. In *J-HGBU workshop*. ACM, 2011. 4

[6] Yin Baocai, Sun Yanfeng, Wang Chengzhang, and Ge Yun. BJUT-3D large scale 3D face database and information processing. *Journal of Computer Research and Development*, 6:020, 2009. 4

[7] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, 1999. 1, 2, 3

[8] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3D morphable model. *TPAMI*, 25(9):1063–1074, 2003. 2

[9] Timo Bolkart and Stefanie Wuhrer. A groupwise multilinear correspondence optimization for 3D faces. In *ICCV*, 2015. 3, 6, 7

[10] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *TPAMI*, 11(6):567–585, 1989. 3

[11] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3D morphable models. *IJCV*, 126(2-4):233–254, 2018. 1, 2, 3, 4, 6

[12] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3D morphable model learnt from 10,000 faces. In *CVPR*, 2016. 1, 2, 3

[13] Alan Brunton, Timo Bolkart, and Stefanie Wuhrer. Multilinear wavelets: A statistical shape space for human faces. In *ECCV*, 2014. 1

[14] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3D facial expression database for visual computing. *TVCG*, 20(3):413–425, 2014. 1, 2, 8

[15] Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. Real-time facial animation with image-based dynamic avatars. *TOG*, 35(4):126:1–126:12, 2016. 1

[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

[17] Timothy F Cootes, Stephen Marsland, Carole J Twining, Kate Smith, and Christopher J Taylor. Groupwise diffeomorphic non-rigid registration for automatic model building. In *ECCV*, 2004. 3

[18] Clement Creusot, Nick Pears, and Jim Austin. A machine-learning approach to keypoint detection and landmarking on 3D meshes. *IJCV*, 102(1-3):146–179, 2013. 7

[19] Hang Dai, Nick Pears, William AP Smith, and Christian Duncan. A 3D morphable model of craniofacial shape and texture variation. In *ICCV*, 2017. 2

[20] Pengfei Dou, Shishir K. Shah, and Ioannis A. Kakadiaris. End-to-end 3D face reconstruction with deep neural networks. In *CVPR*, 2017. 2

[21] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3D object reconstruction from a single image. In *CVPR*, 2017. 2

[22] Zhenfeng Fan, Xiyuan Hu, Chen Chen, and Silong Peng. Dense semantic and topological correspondence of 3D faces without landmarks. In *ECCV*, 2018. 2, 3, 6, 7

[23] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of personalized 3D face rigs from monocular video. *TOG*, 35(3):28, 2016. 8

[24] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-An open framework. In *FG*, 2018. 1, 2, 3, 6, 7

[25] Syed Zulqarnain Gilani, Ajmal Mian, and Peter Eastwood. Deep, dense and accurate 3D face correspondence for generating population specific deformable models. *Pattern Recognition*, 69:238–250, 2017. 3

[26] Syed Zulqarnain Gilani, Ajmal Mian, Faisal Shafait, and Ian Reid. Dense 3D face correspondence. *TPAMI*, 40(7):1584–1598, 2018. 2

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

[28] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. 3D-CODED: 3D correspondences by deep deformation. In *ECCV*, 2018. 3, 5

[29] Yudong Guo, Juyong Zhang, Jianfei Cai, Boyi Jiang, and Jianmin Zheng. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *TPAMI*, 2018. 8

[30] Shalini Gupta, Kenneth R Castleman, Mia K Markey, and Alan C Bovik. Texas 3D face recognition database. In *SSIAI*, 2010. 4

[31] Xiaoguang Han, Chang Gao, and Yizhou Yu. DeepSketch2Face: a deep learning based sketching system for 3D face and caricature modeling. *TOG*, 36(4):126, 2017. 1

[32] Zi-Hang Jiang, Qianyi Wu, Keyu Chen, and Juyong Zhang. Disentangled representation learning for 3D face shape. In *CVPR*, 2019. 1, 2, 3, 4, 8

[33] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via CNN-based dense 3D model fitting. In *CVPR*, 2016. 2

[34] Amin Jourabloo and Xiaoming Liu. Pose-invariant face alignment via CNN-based dense 3D model fitting. *International Journal of Computer Vision*, 124(2):187–203, 2017. 2

[35] Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. Pose-invariant face alignment with a single cnn. In *ICCV*, 2017. 2

[36] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 5

[37] Leif Kobbelt. 3-subdivision. In *SIGGRAPH*, 2000. 4

[38] Artiom Kovnatsky, Michael M Bronstein, Alexander M Bronstein, Klaus Glashoff, and Ron Kimmel. Coupled quasi-harmonic bases. *Computer Graphics Forum*, 32(24):439–448, 2013. 3

[39] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *TOG*, 36(6):194, 2017. 1, 2, 8

[40] Feng Liu, Jun Hu, Jianwei Sun, Yang Wang, and Qijun Zhao. Multi-dim: A multi-dimensional face database towards the application of 3D technology in real-world scenarios. In *IJCB*, 2017. 4

[41] Feng Liu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Joint face alignment and 3D face reconstruction. In *ECCV*, 2016. 1

[42] Feng Liu, Qijun Zhao, Xiaoming Liu, and Dan Zeng. Joint face alignment and 3D face reconstruction with application to face recognition. *PAMI*, 2018. 1

[43] Feng Liu, Ronghang Zhu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Disentangling features in 3D face shapes for joint face reconstruction and recognition. In *CVPR*, 2018. 1, 6

[44] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017. 6

[45] Marcel Lüthi, Thomas Gerig, Christoph Jud, and Thomas Vetter. Gaussian process morphable models. *TPAMI*, 40(8):1860–1873, 2018. 2, 6

[46] Gang Pan, Xiaobo Zhang, Yueming Wang, Zhenfang Hu, Xiaoxiang Zheng, and Zhaohui Wu. Establishing point correspondence of 3D faces via sparse facial deformable model. *TIP*, 22(11):4170–4181, 2013. 5

[47] Ankur Patel and William AP Smith. 3D morphable face models revisited. In *CVPR*, 2009. 2, 3

[48] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D face model for pose and illumination invariant face recognition. In *AVSS*, 2009. 1, 2, 7

[49] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *CVPR*, 2005. 4, 7

[50] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, 2017. 2, 5

[51] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3D faces using convolutional mesh autoencoders. In *ECCV*, 2018. 1, 2, 3, 4, 8

[52] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, 2017. 2, 8

[53] Emanuele Rodolà, Luca Cosmo, Michael M Bronstein, Andrea Torsello, and Daniel Cremers. Partial functional correspondence. *Computer Graphics Forum*, 36(1):222–236, 2017. 3

[54] Joseph Roth, Yiying Tong, and Xiaoming Liu. Adaptive 3D face reconstruction from unconstrained photo collections. *PAMI*, 39(11):2127–2141, 2016. 1

[55] Augusto Salazar, Stefanie Wuhrer, Chang Shu, and Flavio Prieto. Fully automatic expression-invariant face correspondence. *Machine Vision and Applications*, 25(4):859–879, 2014. 6, 7

[56] Arman Savran, Neşe Alyüz, Hamdi Dibeklioğlu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun. Bosphorus database for 3D face analysis. In *Biometrics and Identity Management*, pages 47–56. 2008. 4

[57] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. SfSNet: Learning shape, reflectance and illuminance of facesin the wild. In *CVPR*, 2018. 2

[58] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 1

[59] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Pérez, and Christian Theobalt. MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *CVPR*, 2017. 1, 2, 8

[60] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *TOG*, 34(6):1–14, 2015. 1

[61] George Toderici, Georgios Evangelopoulos, Tianhong Fang, Theoharis Theoharis, and Ioannis A Kakadiaris. UHDB11 database for 3D-2D face recognition. In *PSIVT*, 2013. 4

[62] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gerard Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *CVPR*, 2017. 2, 8

[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

[66] Oliver Van Kaick, Hao Zhang, Ghassan Hamarneh, and Daniel Cohen-Or. A survey on shape correspondence. *Computer Graphics Forum*, 30(6):1681–1707, 2011. 3

[67] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. *TOG*, 24(3):426–433, 2005. 1, 2

[68] Yuhang Wu, Shishir K Shah, and Ioannis A Kakadiaris. Rendering or normalization? An analysis of the 3D-aided pose-invariant face recognition. In *ISBA*, 2016. 4

[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

[71] Chao Zhang, William AP Smith, Arnaud Dessein, Nick Pears, and Hang Dai. Functional faces: Groupwise dense correspondence using functional maps. In *CVPR*, 2016. 3

[72] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3D solution. In *CVPR*, 2016. 2

[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

[74] Syed Zulqarnain Gilani, Faisal Shafait, and Ajmal Mian. Shape-based automatic detection of a large number of 3D facial landmarks. In *CVPR*, 2015. 7