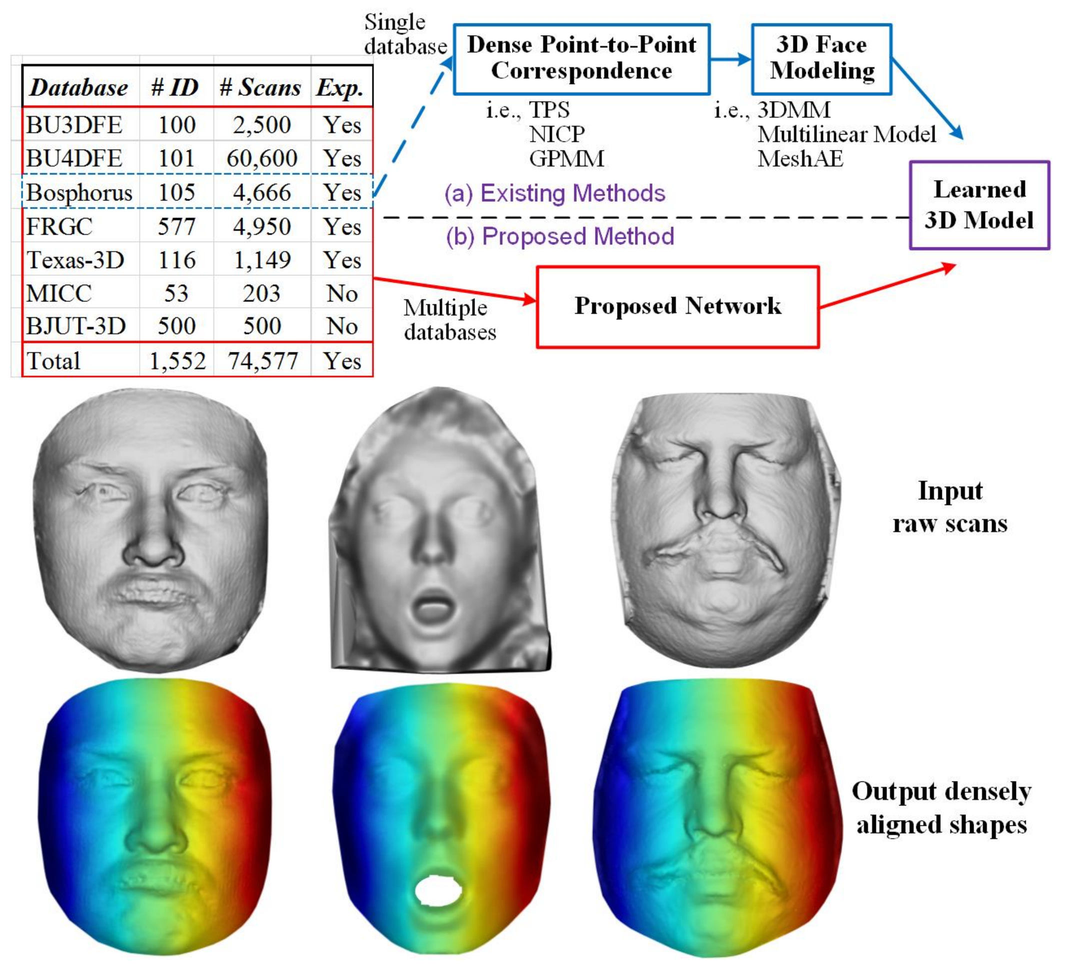
3D Face Modeling from Diverse Raw Scan Data

Feng Liu, Luan Tran and Xiaoming Liu
Department of Computer Science and Engineering, Michigan State University

ICCV 2019 Seoul, Korea

Problem & Contributions



- We jointly learn 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 which can leverage known correspondence prior from synthetic data.
- 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.

Related Works

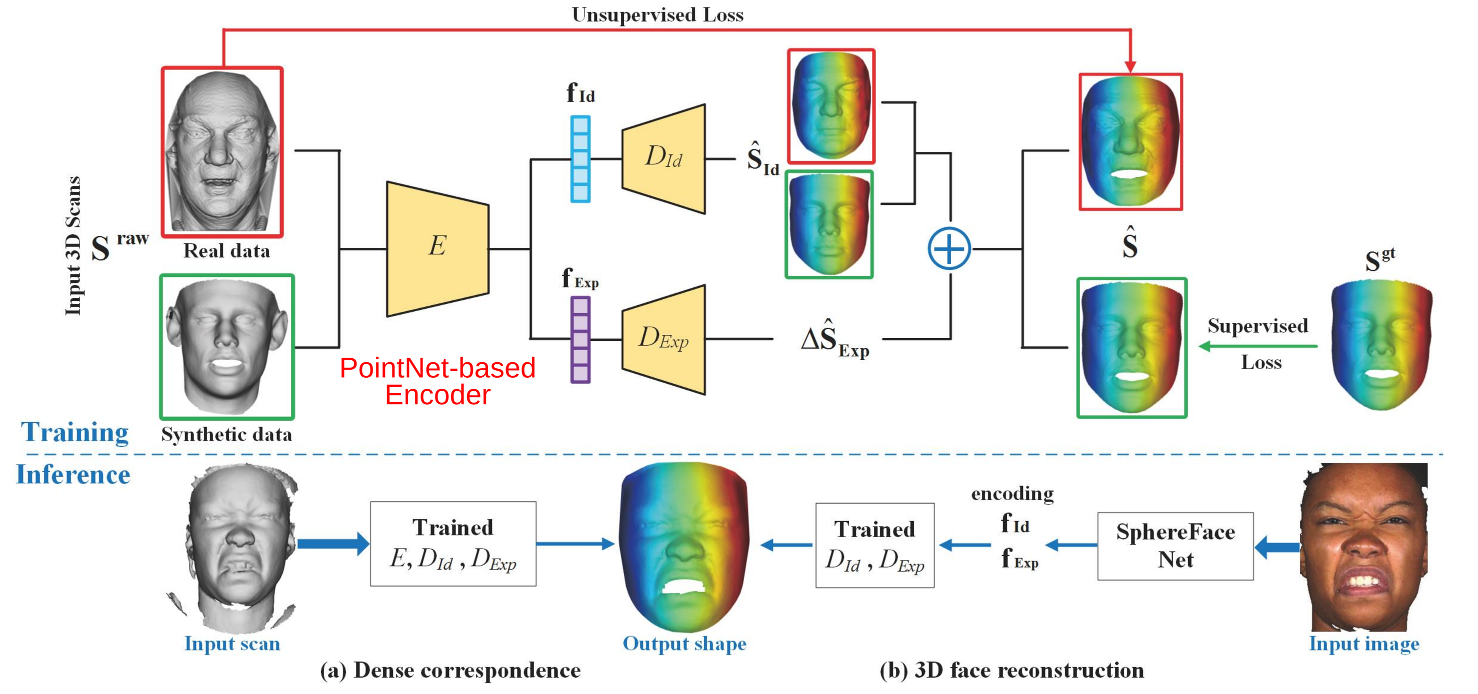
Method	Dataset	Lin./nonL.	#Subj.	Exp.	Corr.
BFM	BFM	Linear	200	No	Yes
GPMMs	BFM	Linear	200	Yes	Yes
LSFM	LSFM	Linear	9,663	No	Yes
LYHM	LYHM	Linear	1,212	No	Yes
Multil. model	FWH	Linear	150	Yes	Yes
FLAME	CAESAR D3DFACS	Linear	3,800 10	Yes	Yes
VAE	Proprietary	Nonlin.	20	No	Yes
MeshAE	COMA	Nonlin.	12	No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

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.

[1] 3D Face Modeling from Diverse Raw Scan Data. ICCV 2019. (Oral presentation)

Proposed Method

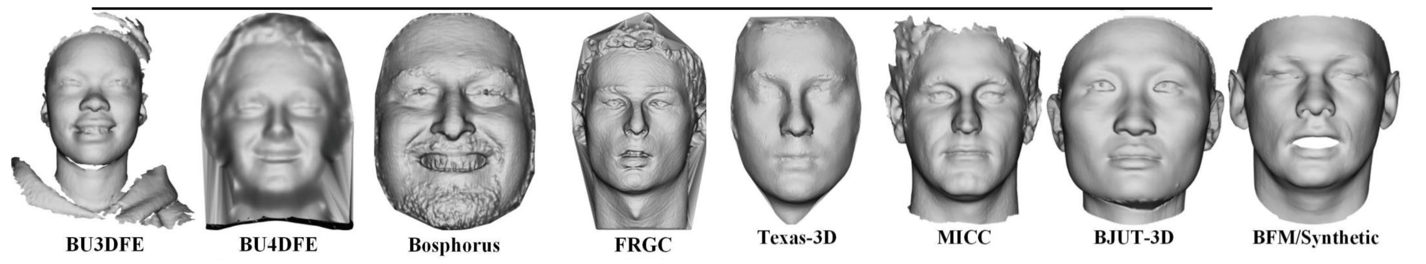
Overall Architecture



Training Data

Summary of training data from related databases.

Database	#Subj.	#Neu.	#Sample	#Exp.	#Sample
BU3DFE	100	100	1,000	2,400	2,400
BU4DFE	101	>101	1,010	>606	2,424
Bosphorus	105	299	1,495	2,603	2,603
FRGC	577	3,308	6,616	1,642	1,642
Texas-3D	116	813	1,626	336	336
MICC	53	103	515	–	–
BJUT-3D	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



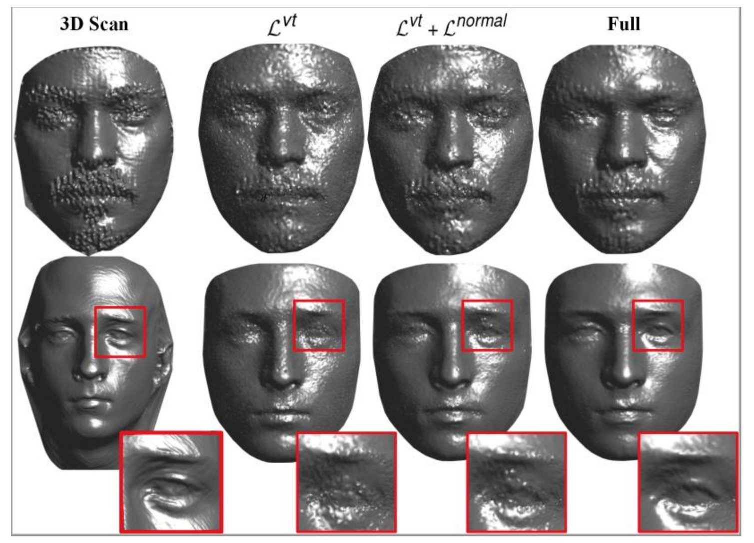
Loss Function

•**Overall Loss**

$$\mathcal{L} = \mathcal{L}^{vt} + \lambda_1 \mathcal{L}^{normal} + \lambda_2 \mathcal{L}^{edge}$$

$$\mathcal{L}^{normal}(\hat{\mathbf{n}}, \mathbf{n}) = \frac{1}{n}\sum_i (1 - \mathbf{n}_i \cdot \hat{\mathbf{n}}_i)$$

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



•**Supervised**

$$\mathcal{L}^{vt} = \|\mathbf{S}^{gt} - \hat{\mathbf{S}}\|_1$$

$$\mathcal{L}^{normal} = \mathcal{L}^{normal}(\hat{\mathbf{n}}, \mathbf{n}^{gt})$$

$$\mathcal{L}^{edge} = \mathcal{L}^{edge}(\hat{\mathbf{S}}, \mathbf{S}^{gt})$$

•**Unsupervised**

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

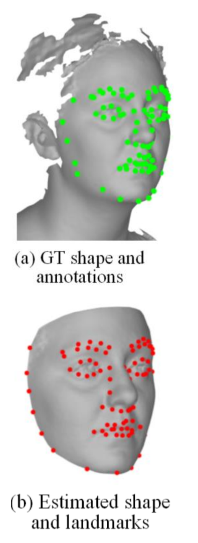
$$\mathcal{L}^{normal} = \mathcal{L}^{normal}(\hat{\mathbf{n}}, \mathbf{n}^{raw}_{(q)}) \qquad \mathcal{L}^{edge} = \mathcal{L}^{edge}(\hat{\mathbf{S}}, \mathbf{S}^{raw}_{(q)})$$

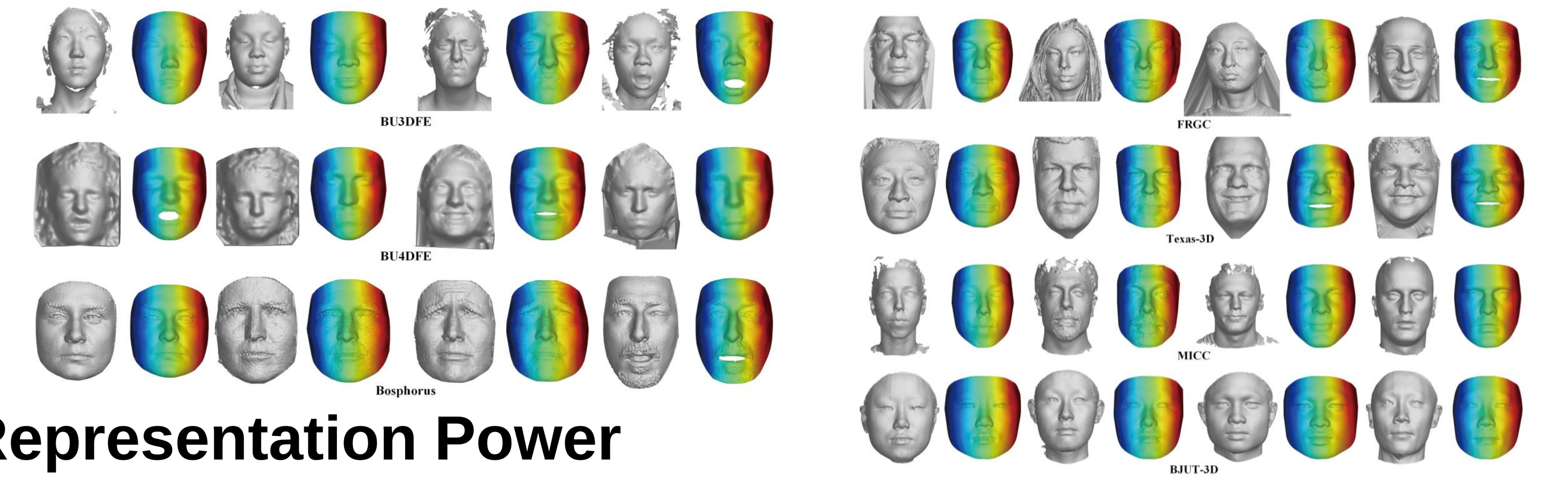
Experimental Results

Dense Correspondence Accuracy

Comparison of the semantic landmark error (*mm*) on BU3DFE.

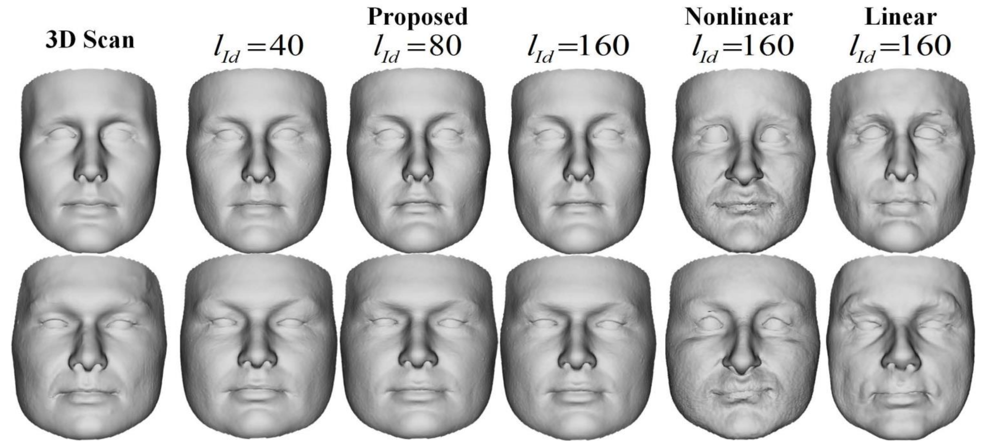
Face Region	NICP	Bolkart *et al.*	Salaza *et al.*	GPMMs	Proposed (**out**)	Proposed (**in**)	Relative Impr.
Left Eyebrow	7.49±2.04	8.71±3.32	6.28±3.30	4.69±4.64	6.25±2.58	**4.18±1.62**	10.9%
Right Eyebrow	6.92±2.39	8.62±3.02	6.75±3.51	5.35±4.69	4.57±3.03	**3.97±1.70**	25.8%
Left Eye	3.18±0.76	3.39±1.00	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%





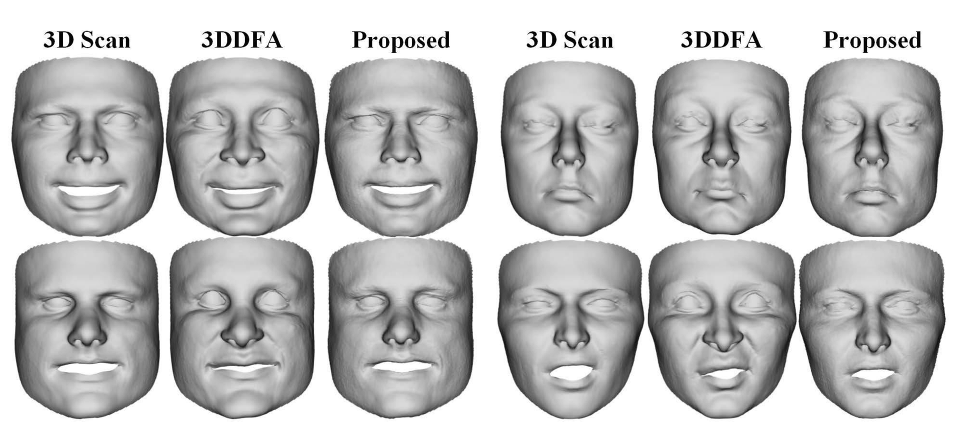
Representation Power

- Identity shape



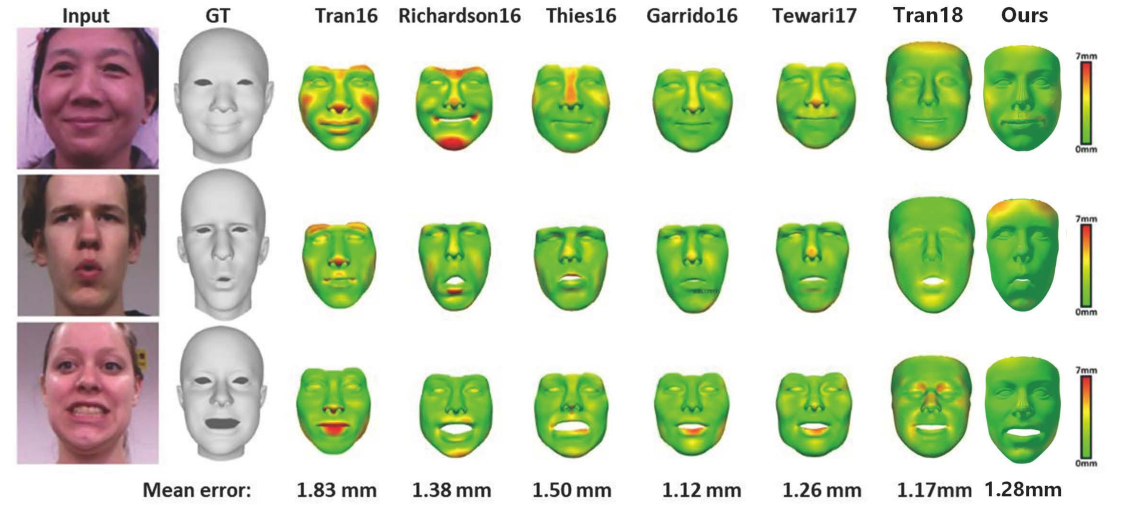
Error 1.258 1.107 **0.946** 1.253 1.019

- Expression shape



Error 2.609 (3DDFA) **1.424** (Proposed)

Application - 3D Face Reconstruction



Efficiency comparison.

Method	Time (s)
NICP	57.48
Fan *et al.*	164.60
Proposed (CPU)	0.26
Proposed (GPU)	2.19×10^{-3}

Conclusions

We propose 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.