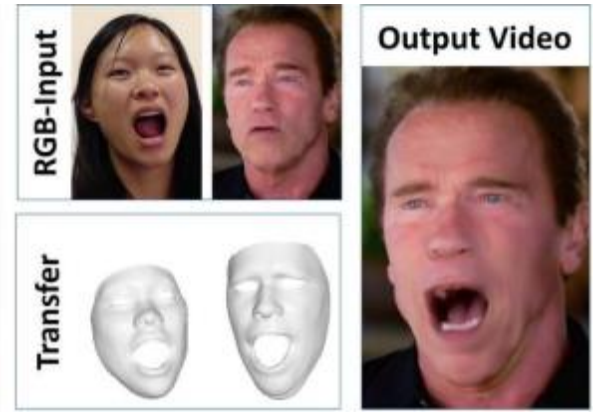
Hybrid Approach for 3D Head Reconstruction: Using Neural Networks and Visual Geometry

Oussema Bouafif [1,2] Bogdan Khomutenko [1] Mohamed Daoudi [2]

[1]MCQ-Scan [2]IMT Lille Douai

Lille, France

Introduction and Motivation



[Thies et al. 2016]

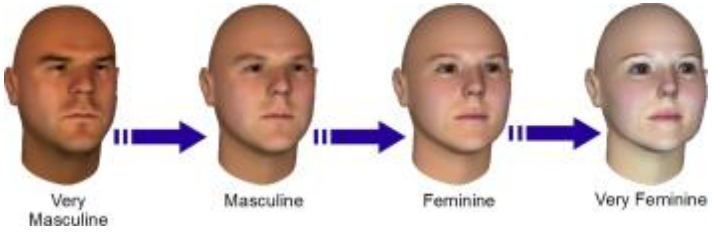
Facial Animation

[FotoNation.com]

Face Recognition

[Hu et al. 2017]

Avatar Digitization

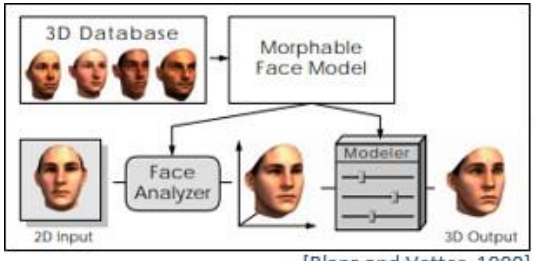


[Gilani et al. 2014]

Gender Classification

Related Work

Monocular 3D face reconstruction methods



[Blanz and Vetter, 1999]

3D Morphable Model (3DMM)

[Amberg et al., 2008]

Laser scans

[Booth et al., 2017]

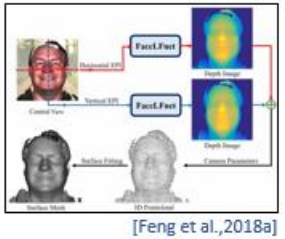
Landmarks

[Bas et al., 2016]

Edges

[Huber et al., 2015]

Local image Features



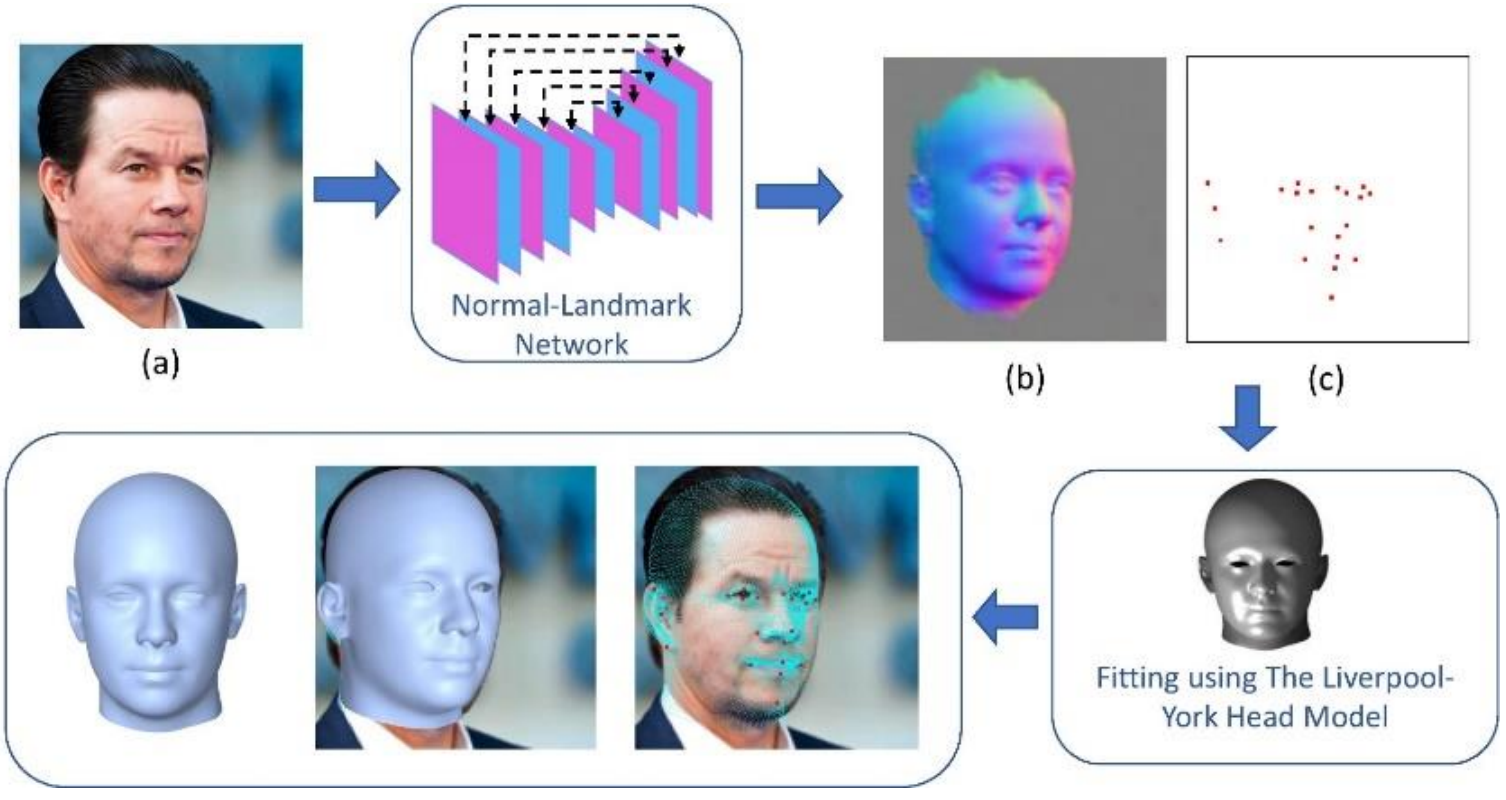
[Feng et al.,2018a]

[Sela et al., 2017]

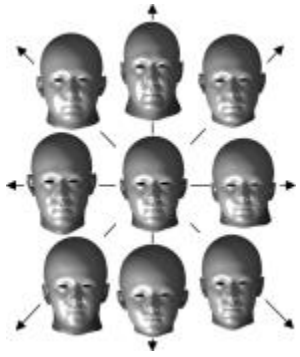
[Feng et al.,2018b]

CNNs

Proposed Method



Synthetic Data Generation



[Dai et al., 2017]

The Liverpool-York Head Model (LYHM)

$$X = X_0 + Wy$$

- X : 3D head,
- X_0 : Mean face shape
- W : Principal components,
- $y \sim N(0, 1)$: Coefficient vector shape.

Complete head model

- Textures
- Hair model (USC-HairSalon)
- Eyes model
- Glasses model

- Illumination conditions
- Shadows
- Poses
- Coefficients for PBM
- Background image

Synthetic Data Generation

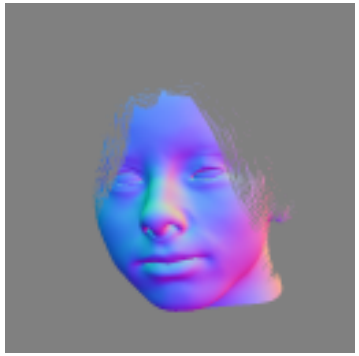
- Local geometric properties,
- Completely independent prediction,
- Translation and scale invariance

Complete head model

$$n_i = \frac{1}{k}\sum_{j=1}^{k} \omega_j \frac{[q_{i,j} - p_i] \times [q_{i,j+1} - p_i]}{\left\|[q_{i,j} - p_i] \times [q_{i,j+1} - p_i]\right\|},$$

$$\omega_j = \arccos\left(\frac{\langle q_{i,j} - p_i, q_{i,j+1} - p_i \rangle}{|q_{i,j} - p_i||q_{i,j+1} - p_i|}\right)$$

[Klasing et al., 2009]



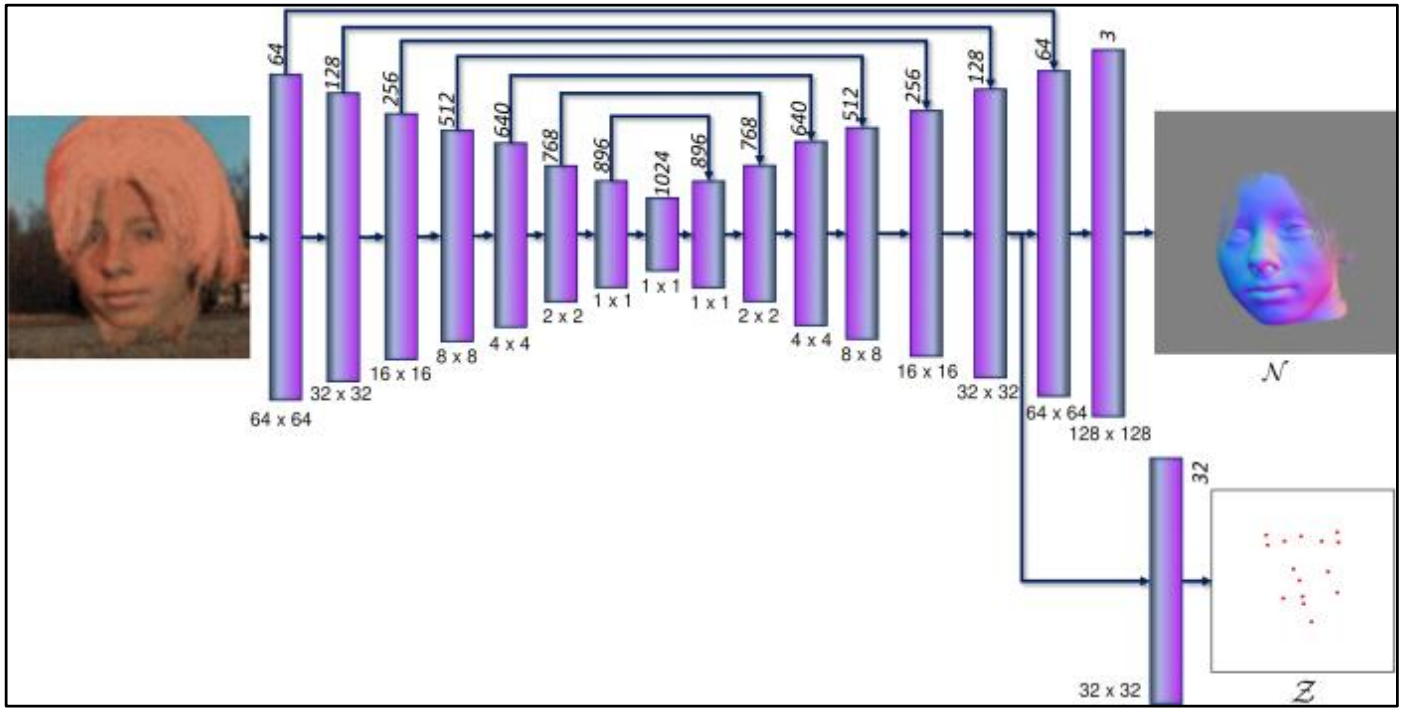
Normal surface

(*N*)

Landmarks map

(Z)

Network Structure

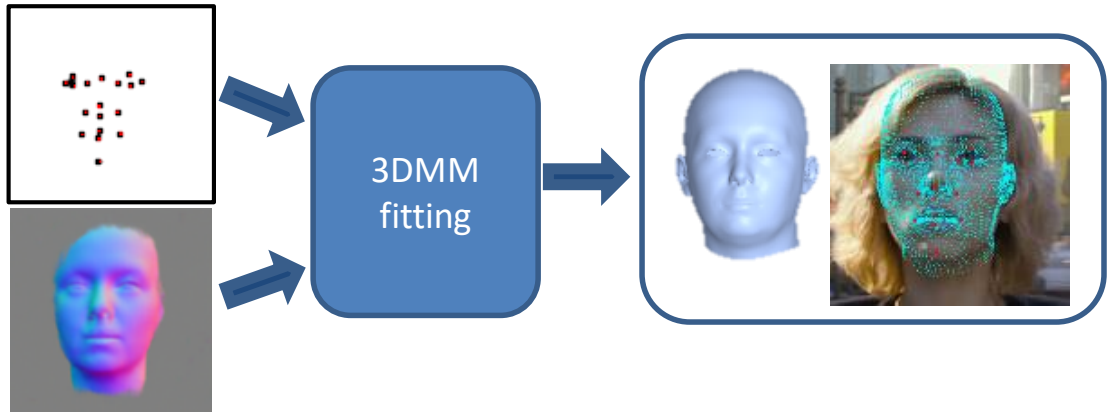


Inspired by [Su et al.,2018]

$$\text{Loss} = L_{\mathcal{N}} + L_{\mathcal{Z}}$$

$$L_{\mathcal{N}} = ||\mathcal{N}_{\text{GT}} - \mathcal{N}||_2^2 \quad , \quad L_{\mathcal{Z}} = ||\mathcal{Z}_{\text{GT}} - \mathcal{Z}||_2^2$$

3D Morphable Model fitting



3D Morphable Model

LYHM

$$X = X_0 + Wy$$

$X \in \mathrm{R}^{3 \times 11510} \rightarrow$ Generated 3D head
$X_0 \in \mathrm{R}^{3 \times 11510} \rightarrow$ Mean head
$\mathrm{W} \in \mathrm{R}^{3 \times 11510 \times 100} \rightarrow$ Principal components
$y \in \mathrm{R}^{100} \rightarrow$ Parameter vector

Morphable Model's Normals

$$\tilde{n}_i = \sum_{j=1}^{11510} \frac{(q_{i,j} - p_i) \times (q_{i,j+1} - p_i)}{\left\| (q_{i,j} - p_i) \times (q_{i,j+1} - p_i) \right\|}$$

$$n_i = \frac{\tilde{n}_i}{\|\tilde{n}_i\|}$$

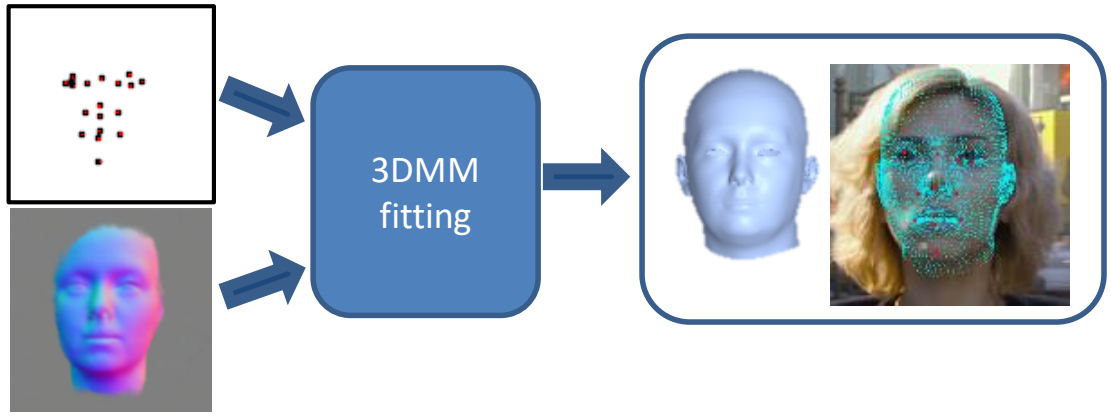
$n_i \rightarrow$ normal vector
$p_i \rightarrow$ vertex location
$\{q_{i,1}, q_{i,2}, \dots, q_{i,11510}\} \rightarrow$ adjacent vertices

Projection Model

$$a = \frac{1}{z} K R (p + t)$$

$\mathrm{a} \in \mathrm{R}^2 \rightarrow$ projected point
$\mathrm{R} \in \mathrm{SO}(3) \rightarrow$ rotation matrix
$\mathrm{z} \rightarrow$ normalization of a 3D point
$\mathrm{K} \rightarrow$ projection matrix $\begin{pmatrix} f & 0 & u_0 \\ 0 & f & v_0 \end{pmatrix}$
$\mathrm{t} \in \mathrm{R}^3 \rightarrow$ translation vector

3D Morphable Model fitting



Parametric Regression

$$E = \lambda_N \mathbf{E}_N + \lambda_Z \mathbf{E}_Z + \lambda_P \mathbf{E}_P$$

$$E_N = \frac{1}{2} \sum_{i=1}^{11510} \|N(a_i) - n_i\|^2$$

$N(a_i) \rightarrow$ interpolated normals
$n_i \rightarrow$ normal vector

The normals loss

$$E_Z = \frac{1}{2} \sum_{j=1}^{11510} \|z_j - b_j\|^2$$

$z_j \rightarrow$ detected landmarks
$b_j \rightarrow$ projections of the corresponding subset of 3D vertices

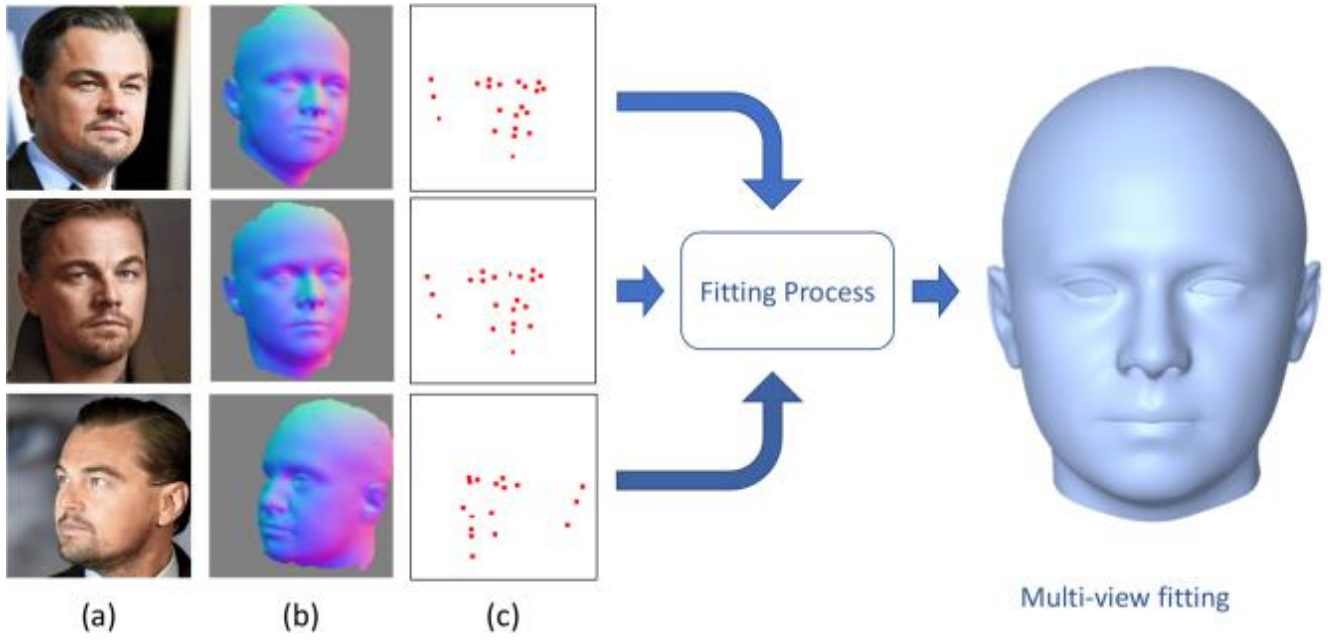
The landmarks loss

$$E_P = \frac{1}{2} y^T C_y^{-1} y$$

$C_y \rightarrow$ covariance matrix with the diagonal $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_{100}^2)$
$b_j \rightarrow$ projections of the corresponding subset of 3D vertices

The shape prior loss

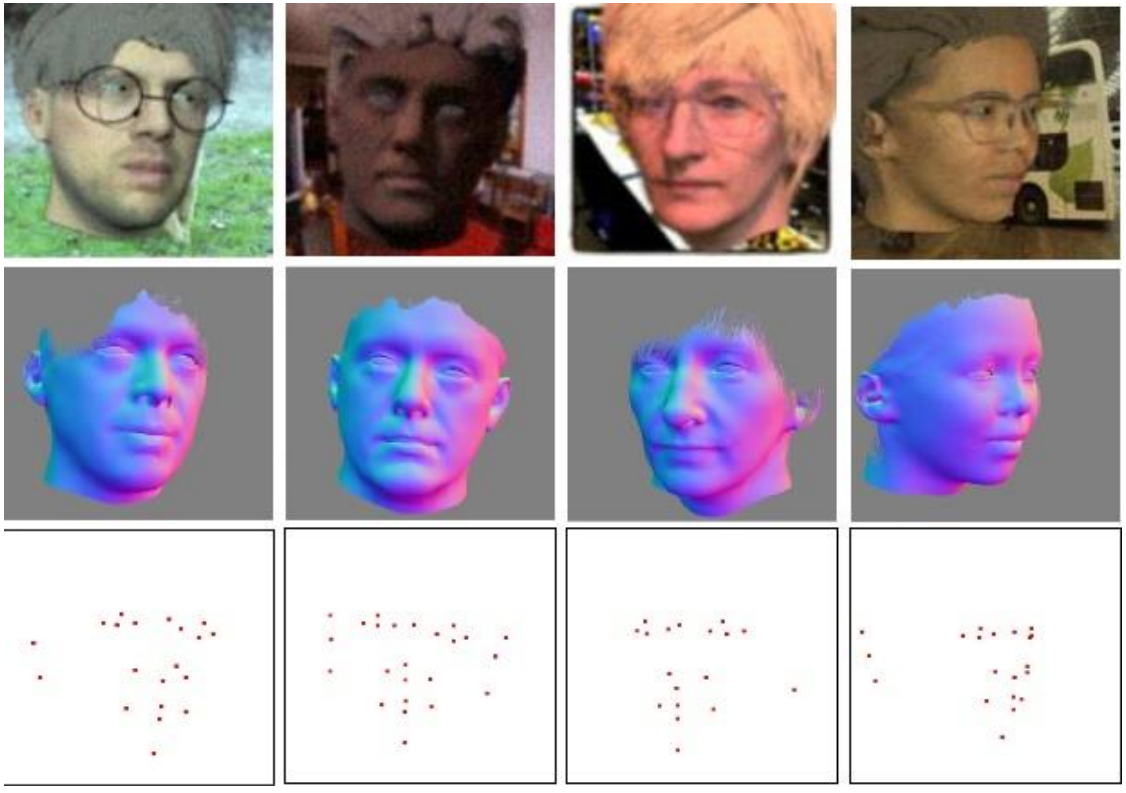
Multi-view 3D Face Reconstruction



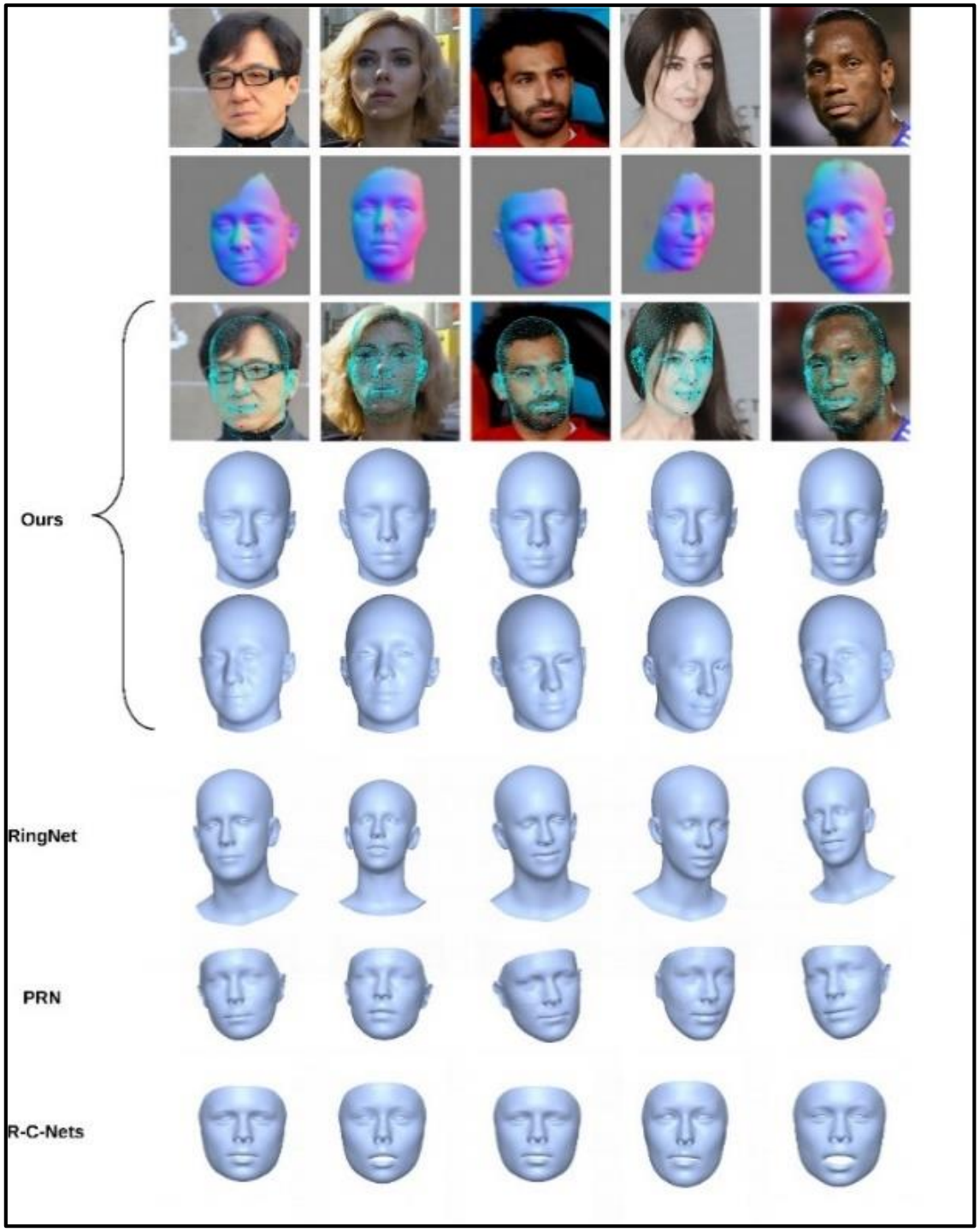
Experimental Results

Training Data set :

- 60,000 facial images (30,000 for males and also for females).
- 3000 epochs, learning rate : $1e^{-5}$, 32 : batch size, *RMSpropoptimizer*
- random blur effect, gaussian noise
- λ_N=1, λ_Z=0.8 , λ_P=0.4,

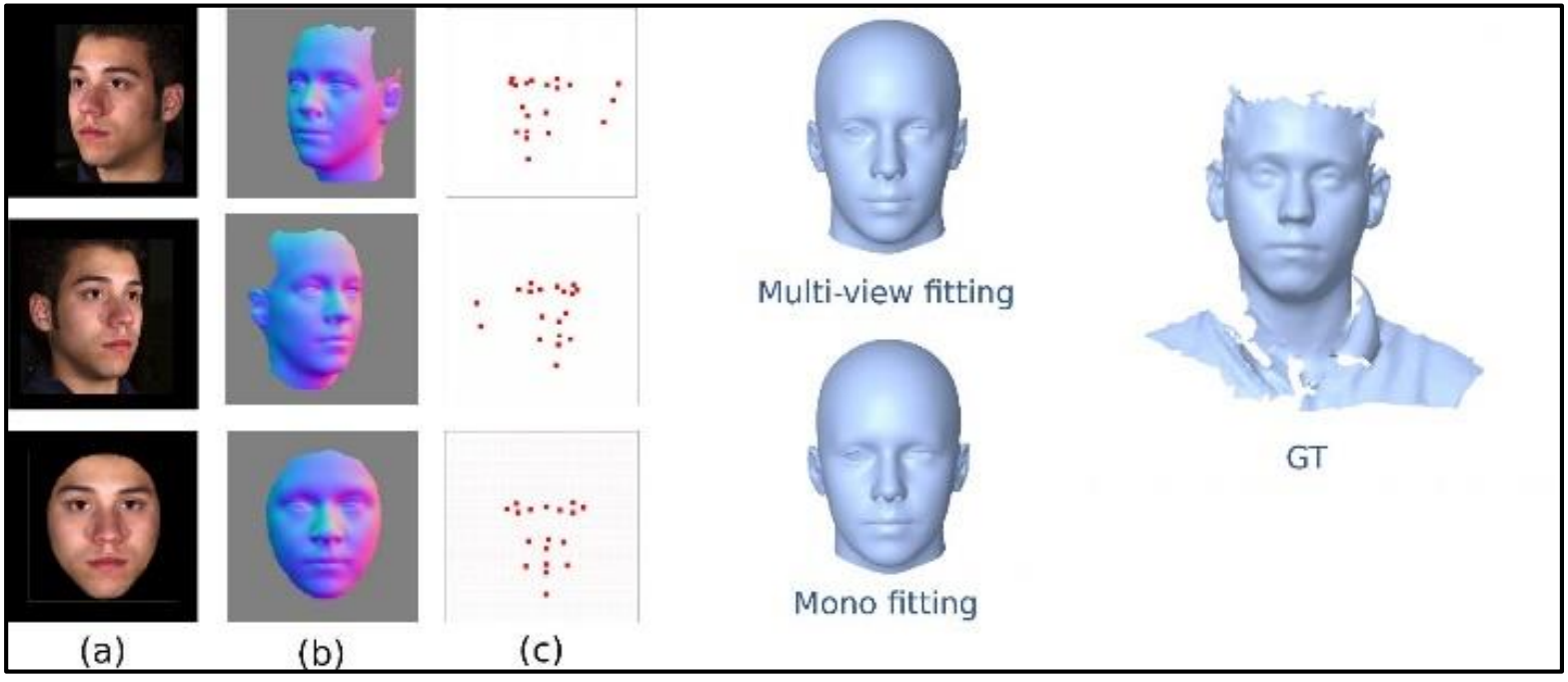


Evaluation



Evaluation

Stereo vs Mono Fitting



- (Multi-view fitting): 3D head reconstruction using all images in the same fitting process.
- (Mono fitting) : 3D head exploiting only the frontal image in the fitting process (third row).
- (GT) : the ground-truth of the 3D head mesh.

Final Evaluation

Point-to-plane Root Mean Square Error (RMSE)

RMSE → $\epsilon = \sqrt{\frac{\sum_{i=1}^{N}((p_i - q_i).n_i)^2}{N}}$

N → number of vertices

P_i → vertex of the reference model

q_i → nearest neighbor from the reconstructed model

n_i → normal vector of the reference model

Method	Ours (Mono)	Ours (Multi)	RingNet [3]	PRN [4]	R-C-Nets [5]
RMSE	1.74±0.44	1.67±0.43	1.90±0.49	1.86±0.47	**1.60±0.41**
μ	2.21	2.17	3.42	1.83	**1.64**
σ	1.08	**1.04**	1.58	1.70	1.69
m̃	2.09	2.05	3.23	1.43	**1.27**
$\delta_{90\%}$	3.61	3.53	5.60	3.46	**3.00**

Absolute depth error

Quantitative comparison on the BU-3DFE [2] dataset. Lowers are better

[2] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in7th international conference on automatic face and gesture recognition (FGR06). IEEE,2006, pp. 211–216.

[3] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black, "Learning to regress 3D face shape and expression from an image without 3D supervision" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7763–7772.

[4] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," in Proceedings of the European Conference on Computer Vision (ECCV),2018, pp. 534–551.

[5] . Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019

Conclusion

- Hybrid 3D face reconstruction approach *(machine learning and geometric)*. Reconstruction from a single or multiple images
- Trained only on synthetic data but generalizes on real world data.
- Evaluation :
- Effectiveness of using the *(Z)* map.
  - Performance on *BU3D-FE* dataset.
  - New error calculation method *(Point-to-Plane)*.
  -

- Limitation :
  - No facial expressions and a limited age range (LYHM).
  - Synthetic data can have unrealistic features.

Thank you !