Imperial College London
Facesoft

GANFIT: Generative Adversarial Network Fitting for High Fidelity 3D Face Reconstruction

Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou
{b.gecer, s.ploumpis, s.zafeiriou}@imperial.ac.uk, drkotsia@gmail.com

Project Page

CVPR
LONG BEACH
CALIFORNIA
June 16-20, 2019

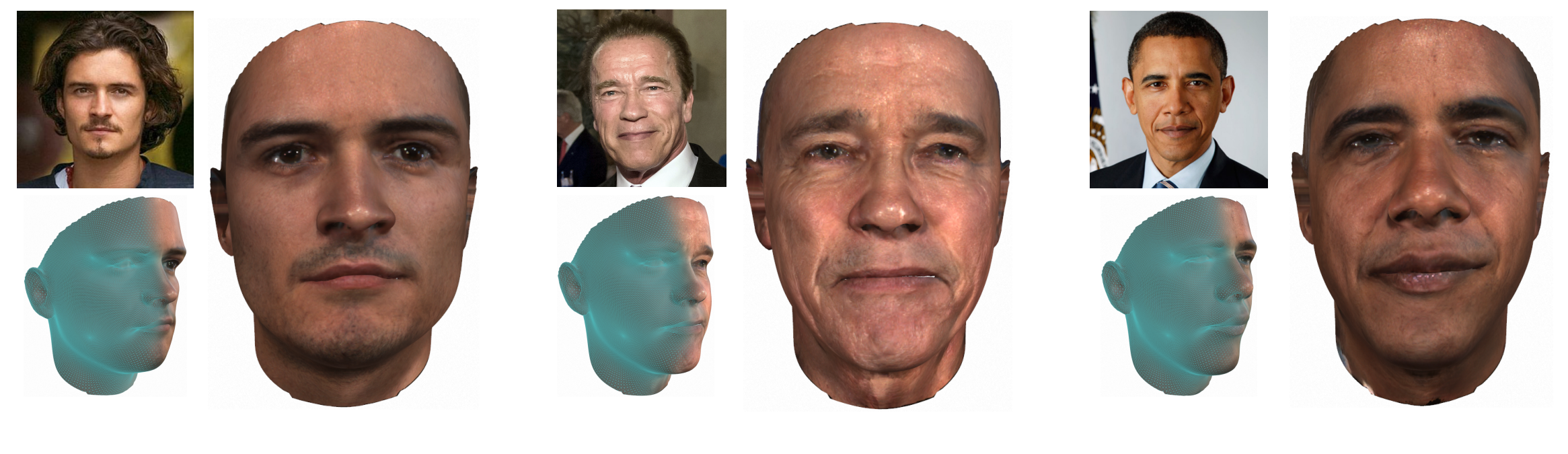


Figure 1: The proposed deep fitting approach can reconstruct high quality texture and geometry from a single image with precise identity recovery. The reconstructions in the figure and the rest of the paper are represented by a vector of size 700 floating points and rendered without any special effects. We would like to highlight that the depicted texture is reconstructed by our model and none of the features taken directly from the image.

Abstract

- In this paper, we **revisit** *optimization-based 3D face reconstruction* under a new perspective:
- Instead of linear models, we use a **GAN** trained with high-resolution UV maps as our *statistical representation* of the facial texture.
- Instead of primitive cost functions used in the literature based on low- and mid-level features (e.g., RGB values, edges, SIFT), we propose a *novel cost function* that is based on **deep face recognition network**.
- We replace physical *image formation* stage with a **differentiable renderer** to make use of first order derivatives (i.e., gradient descent).
- This is the first time that GANs are used for model fitting and the proposed approach shows identity preserving high fidelity 3D reconstructions in qualitative and quantitative experiments.

Project page: https://github.com/barisgecer/ganfit

References

[1] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018.

[2] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *CVPR*, 2018.

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

[4] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. 2018.

[5] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *CVPR*, 2018.

[6] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, Stefanos Zafeiriou, et al. 3d face morphable models "in-the-wild". In *CVPR*, 2017.

Overview of the Proposed Framework

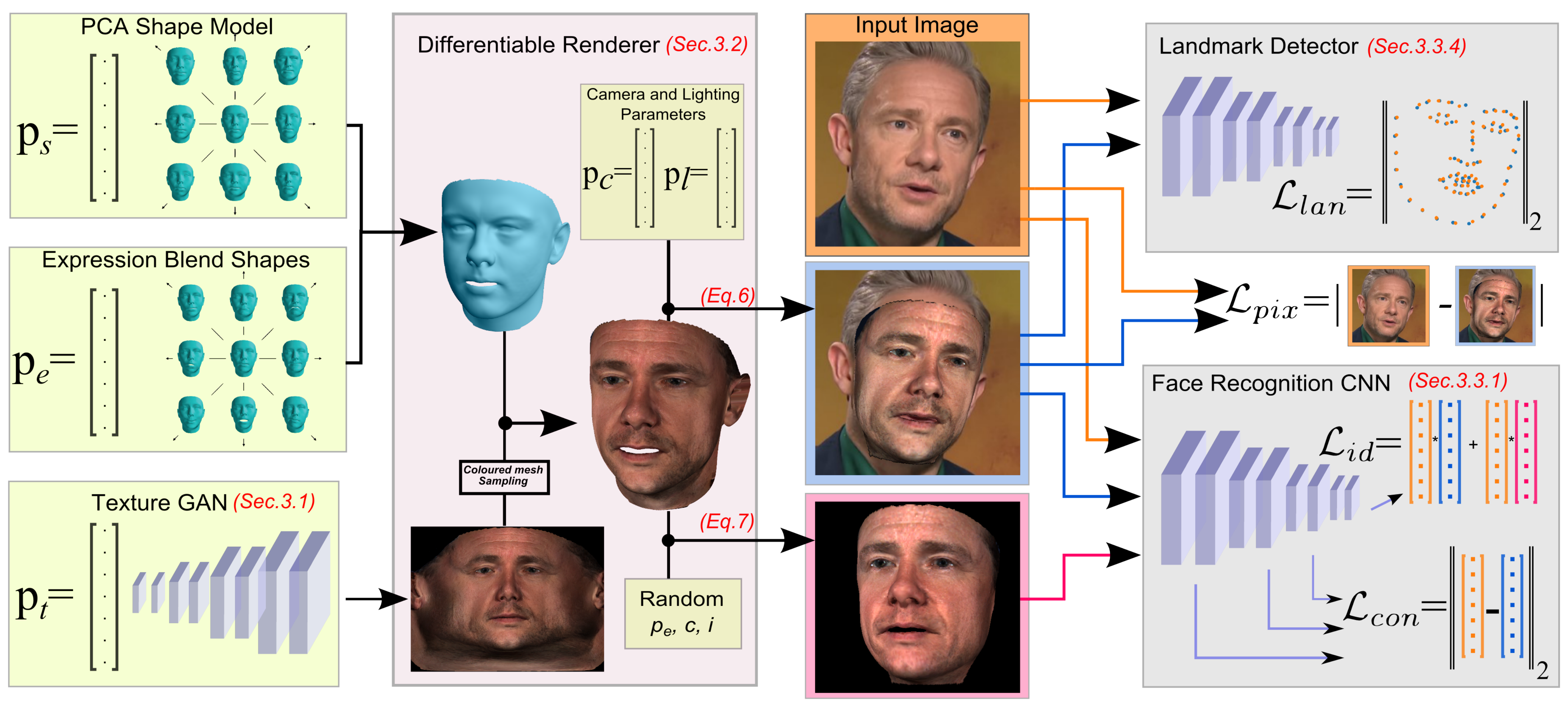


Figure 2: A 3D face reconstruction is rendered by a differentiable renderer (shown in purple). Cost functions are mainly formulated by means of identity features on a pretrained face recognition network (shown in gray) and they are optimized by flowing the error all the way back to the latent parameters ($\mathbf{p}_s, \mathbf{p}_e, \mathbf{p}_t, \mathbf{p}_c, \mathbf{p}_l$, shown in green) with gradient descent optimization. End-to-end differentiable architecture enables us to use computationally cheap and reliable first order derivatives for optimization thus making it possible to employ deep networks as a statistical model and as a cost function.

Approach

- **Texture GAN ($\mathcal{G}(\mathbf{p}_t) : \mathbb{R}^{512} \rightarrow \mathbb{R}^{H \times W \times C}$) :** A Progressive Growing GAN [1] is trained with 10,000 high resolution textures as our texture model.

- **Differentiable Renderer :** Formation of reconstruction images are done by a differentiable renderer [2] to backpropagate the cost functions

$$\mathbf{I}^{\mathcal{R}} = \mathcal{R}(\mathbf{S}(\mathbf{p}_s, \mathbf{p}_e), \mathcal{P}(\mathcal{G}(\mathbf{p}_t)), \mathbf{p}_c, \mathbf{p}_l)$$
$$\hat{\mathbf{I}}^{\mathcal{R}} = \mathcal{R}(\mathbf{S}(\mathbf{p}_s, \hat{\mathbf{p}_e}), \mathcal{P}(\mathcal{G}(\mathbf{p}_t)), \hat{\mathbf{p}_c}, \hat{\mathbf{p}_l})$$

$\mathbf{p}_s, \mathbf{p}_e, \mathbf{p}_t, \mathbf{p}_x, \mathbf{p}_l$ are shape, expression, texture, camera and lighting parameters respectively

- **Cost Functions:**
  - **Identitity Loss:** $\mathcal{L}_{id} = 1 - \frac{\mathcal{F}^n(\mathbf{I}^0).\mathcal{F}^n(\mathbf{I}^{\mathcal{R}})}{||\mathcal{F}^n(\mathbf{I}^0)||_2||\mathcal{F}^n(\mathbf{I}^{\mathcal{R}})||_2}$
  - **Content Loss:** $\mathcal{L}_{con} = \sum_j^n \frac{||\mathcal{F}^j(\mathbf{I}^0) - \mathcal{F}^j(\mathbf{I}^{\mathcal{R}})||_2}{H_{\mathcal{F}^j} \times W_{\mathcal{F}^j} \times C_{\mathcal{F}^j}}$
  - **Pixel Loss:** $\mathcal{L}_{pix} = ||\mathbf{I}^0 - \mathbf{I}^{\mathcal{R}}||_1$
  - **Landmark Loss:** $\mathcal{L}_{lan} = ||\mathcal{M}(\mathbf{I}^0) - \mathcal{M}(\mathbf{I}^{\mathcal{R}})||_2$

- **Model Fitting:**

$$\min_{\mathbf{p}} \mathcal{E}(\mathbf{p}) = \lambda_{id}\mathcal{L}_{id} + \hat{\lambda}_{id}\hat{\mathcal{L}}_{id} + \lambda_{con}\mathcal{L}_{con} + \lambda_{pix}\mathcal{L}_{pix} + \lambda_{lan}\mathcal{L}_{lan} + \lambda_{reg}Reg(\{\mathbf{p}_{s,e}, \mathbf{p}_l\})$$

- **Fitting Multiple Images:** $\mathbf{p}_s = \sum_i^n \mathbf{p}_s^i, \mathbf{p}_t = \sum_i^n \mathbf{p}_t^i$

Qualitative Results

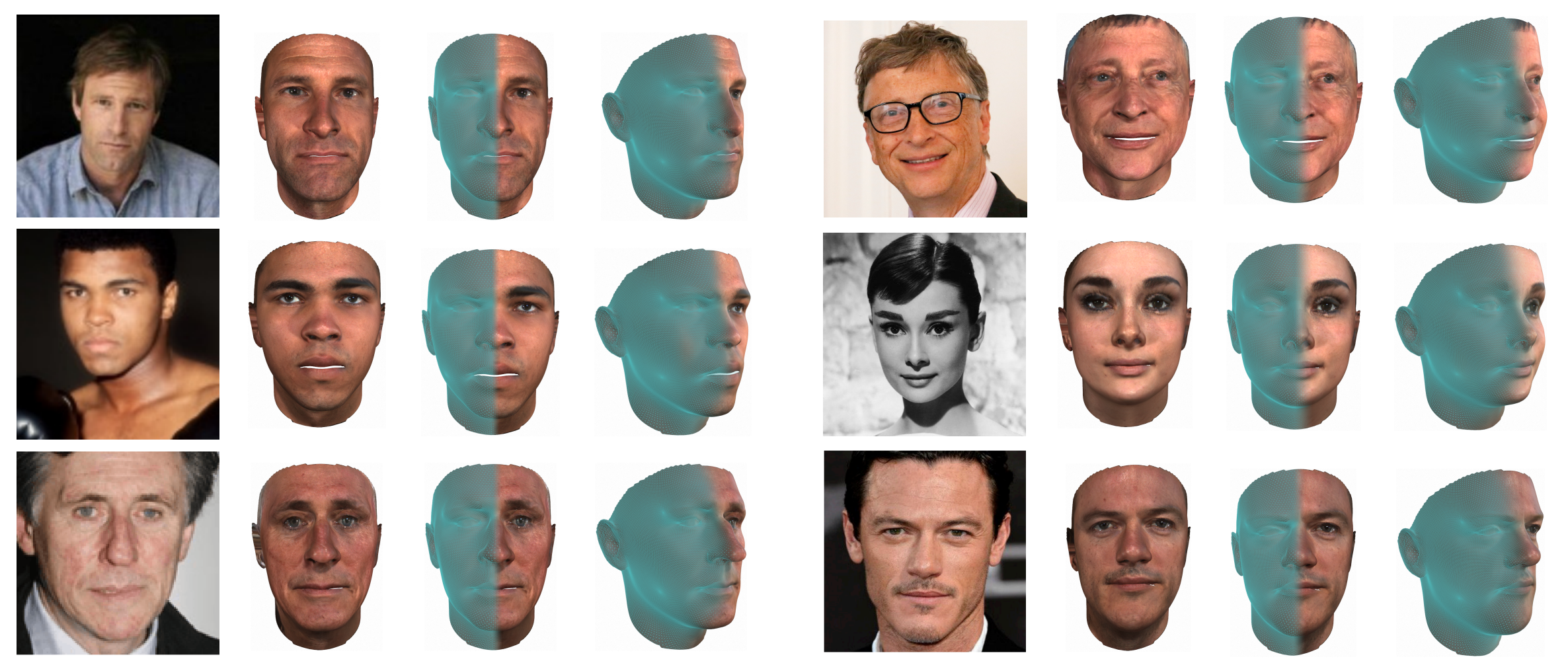


Figure 3: Our approach is robust to occlusion (e.g., glasses), low resolution and black-white in the photos and generalizes well with ethnicity, gender and age. The reconstructed textures are very well at capturing high frequency details of the identities; likewise, the reconstructed geometries from 3DMM are surprisingly good at identity preservation thanks to the identity features used, e.g. crooked nose at bottom-left, dull eyes at bottom-right and chin dimple at top-left

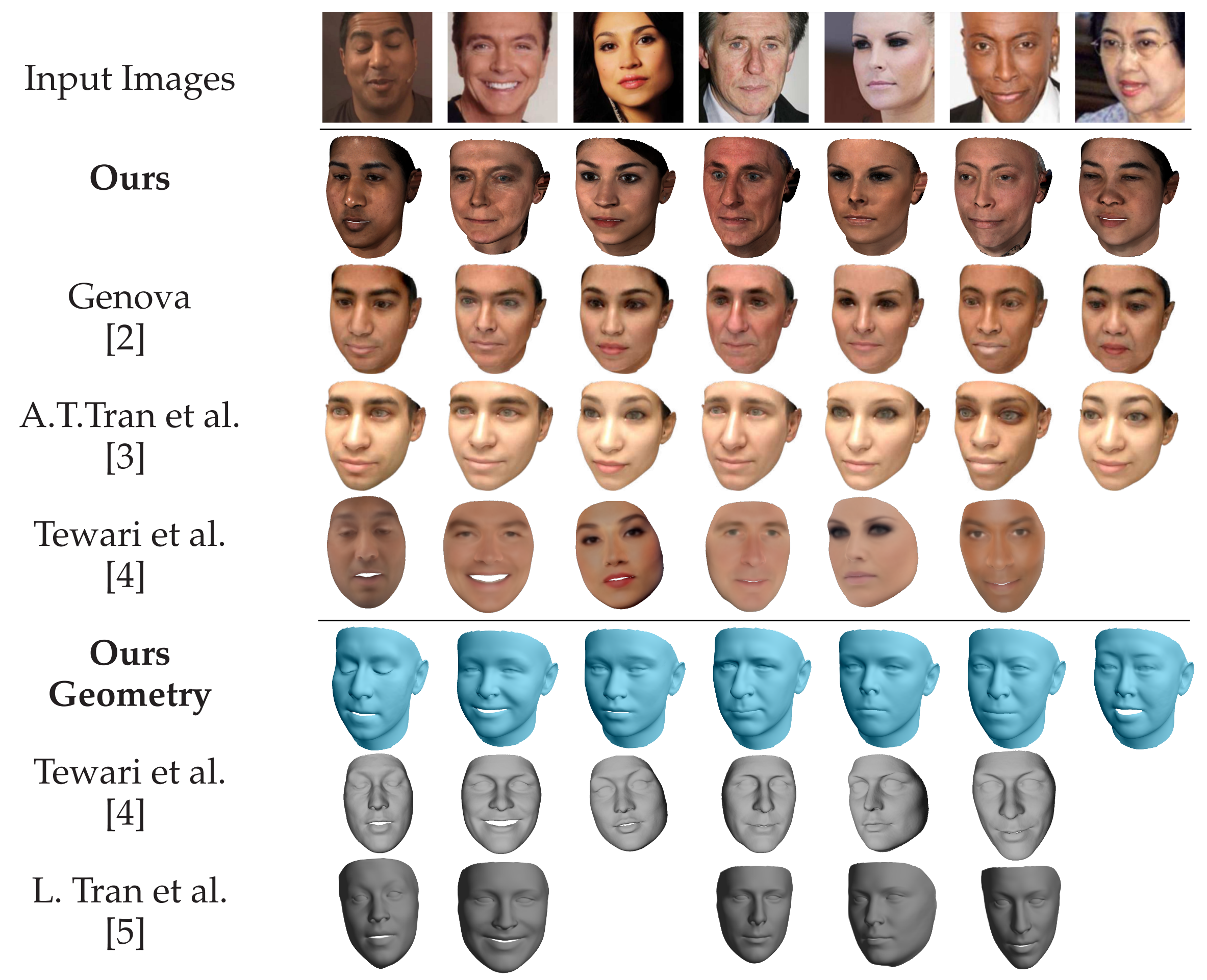
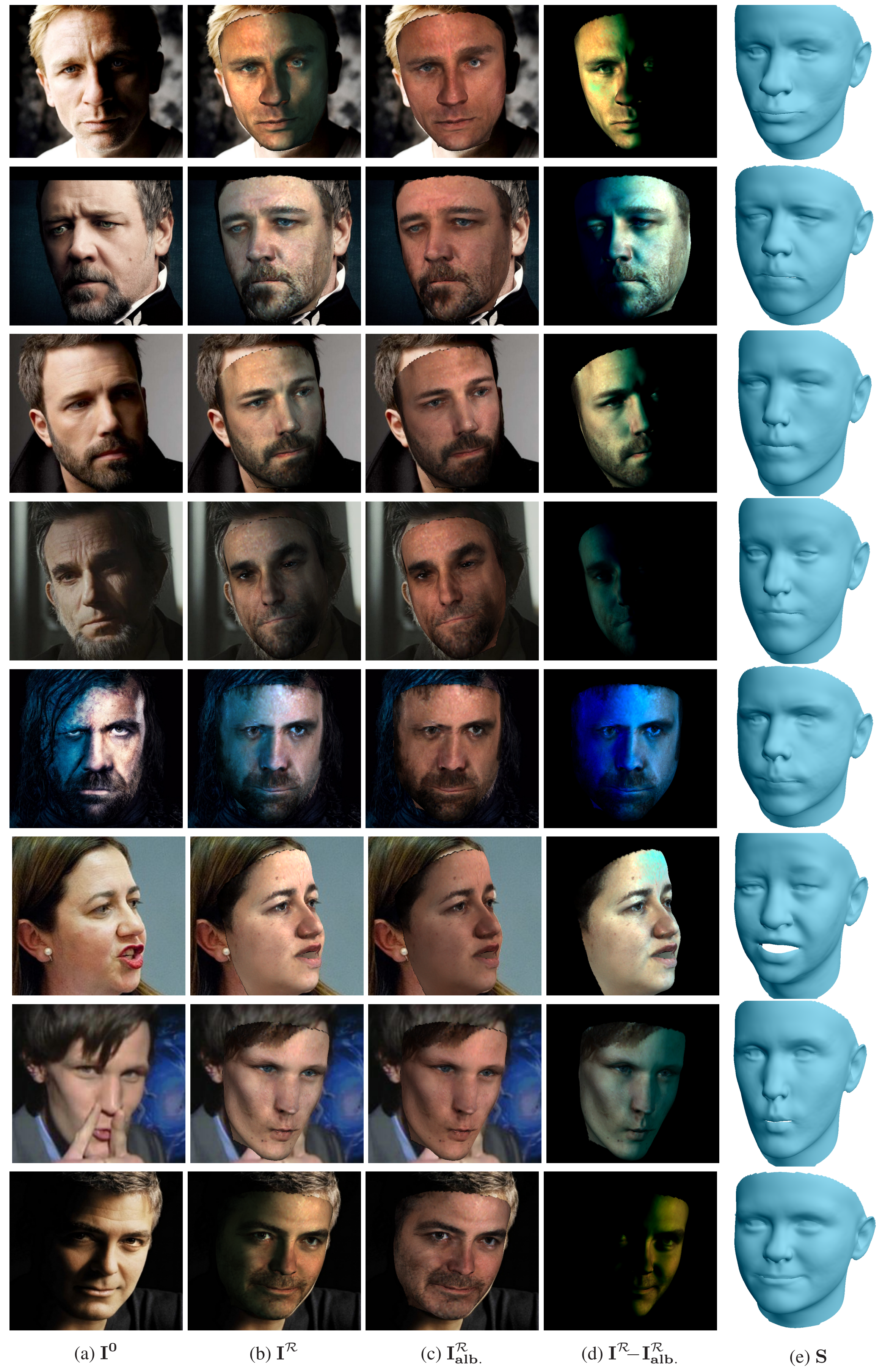


Figure 4: Comparison of our qualitative results with other state-of-the-art methods in MoFA-Test dataset. Rows 2-5 show comparison with textured geometry and rows 6-8 compare only shapes.



(a) $\mathbf{I}^0$ (b) $\mathbf{I}^{\mathcal{R}}$ (c) $\mathbf{I}^{\mathcal{R}}_{\mathbf{alb.}}$ (d) $\mathbf{I}^{\mathcal{R}} - \mathbf{I}^{\mathcal{R}}_{\mathbf{alb.}}$ (e) $\mathbf{S}$

Figure 5: Results under more challenging conditions, i.e. strong illuminations, self-occlusions and facial hair. (a) Input image, (b) Estimated fitting overlayyed including illumination estimation, (c) Overlayyed fitting without illumination, (d) Pixel-wise intensity difference of (b) to (c), (e) Estimated shape mesh

Quantitative Experiments

Method	*Cooperative* Mean	*Cooperative* Std.	*Indoor* Mean	*Indoor* Std.	*Outdoor* Mean	*Outdoor* Std.
Tran *et al.* [3]	1.93	0.27	2.02	0.25	1.86	0.23
Booth *et al.* [6]	1.82	0.29	1.85	0.22	1.63	0.16
Genova *et al.* [2]	1.50	0.13	1.50	0.11	1.48	0.11
Ours	**0.95**	**0.107**	**0.94**	**0.106**	**0.94**	**0.106**

Table 1: Accuracy results for the meshes on the MICC Florence Dataset using point-to-plane distance.

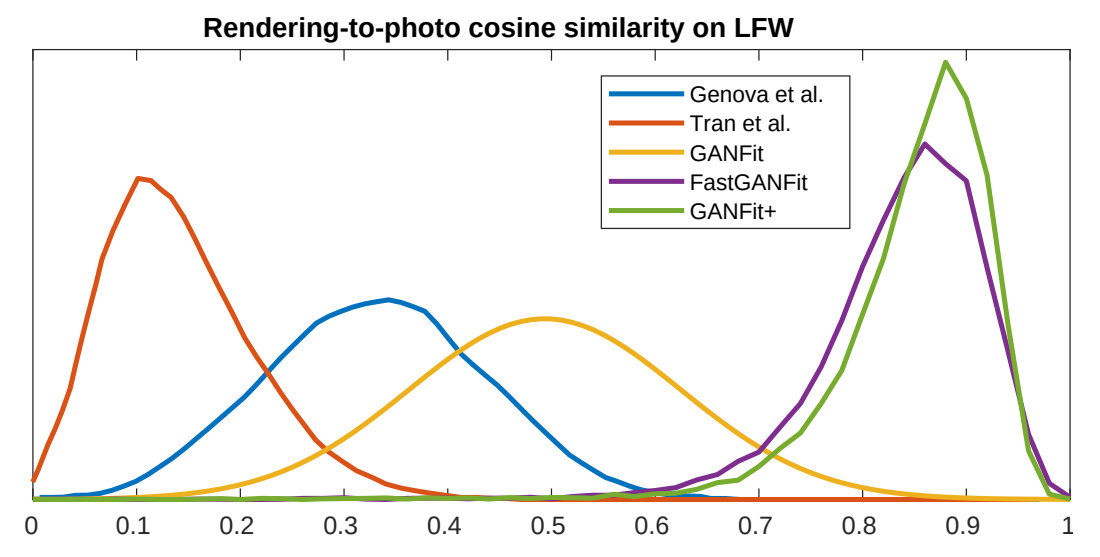


Figure 6: Cosine similarity distributions of rendered and real images LFW based on VGG-Face.

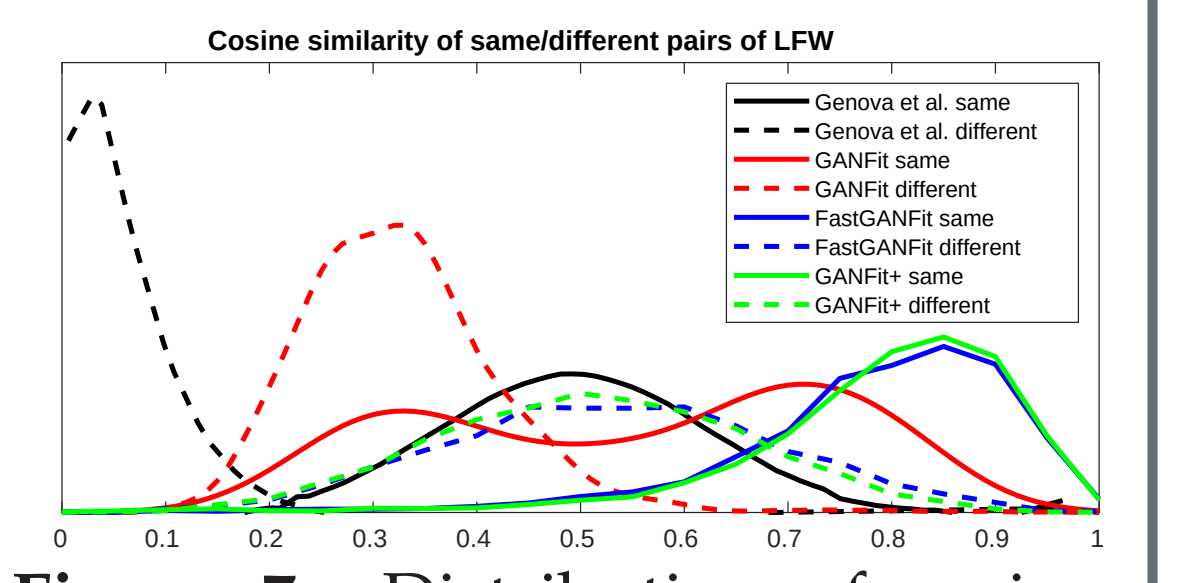


Figure 7: Distribution of cosine similarity of same/different pairs of LFW