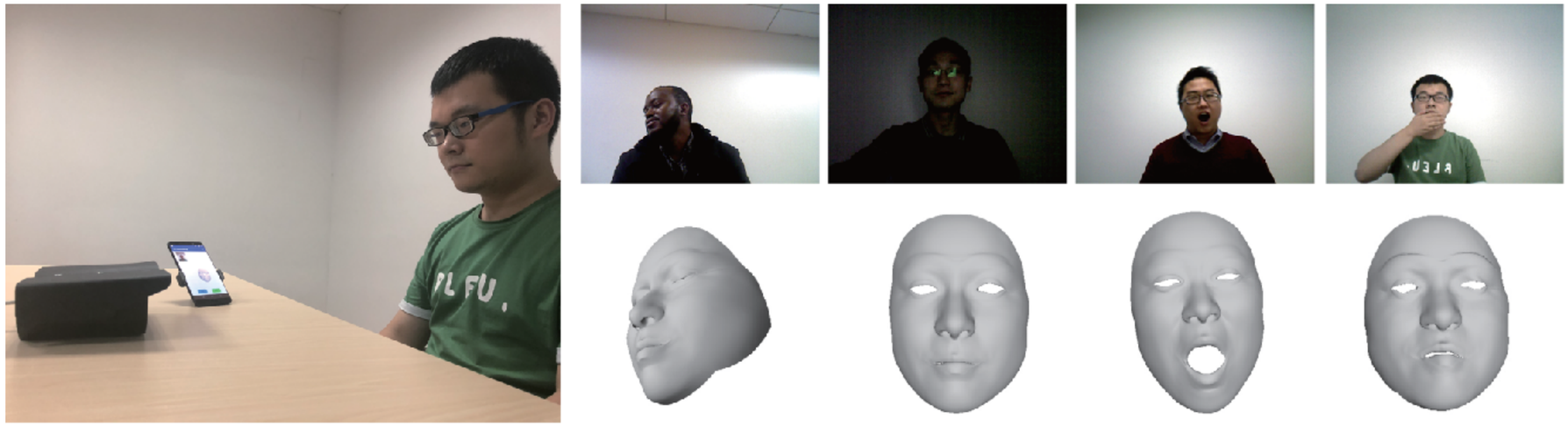
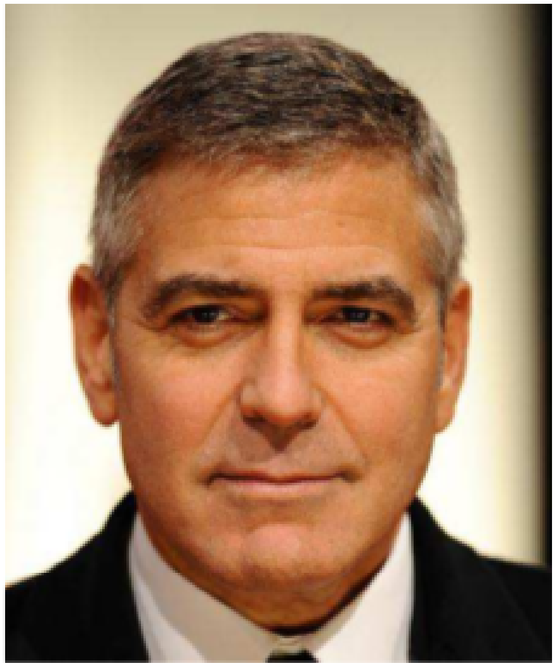
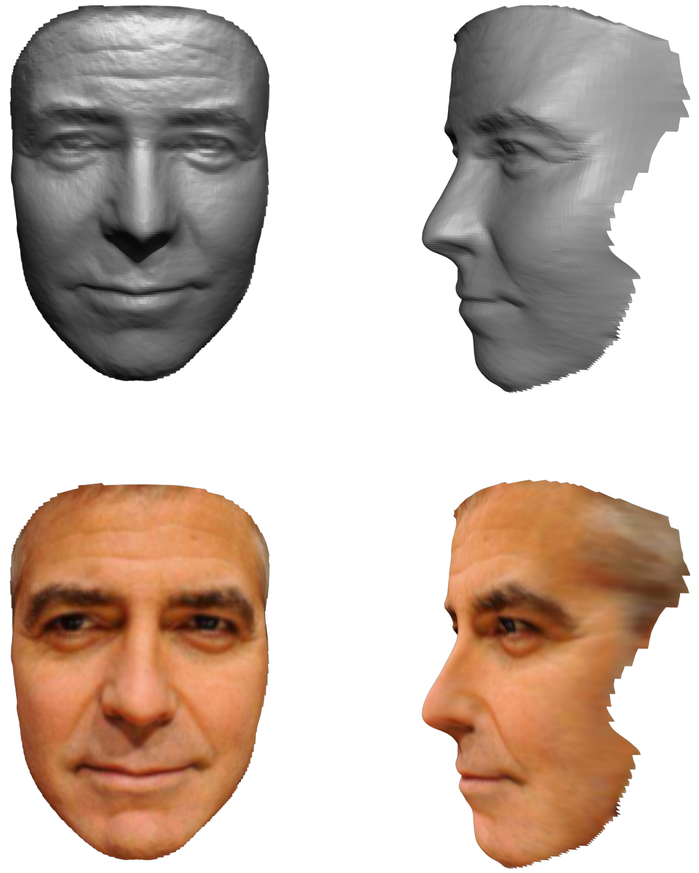
基于深度学习的三维人脸建模



张举勇
中国科学技术大学

Reconstruction From Image





3D Face Reconstruction with Geometry Details from a Single Image

IEEE Trans on Image Processing

Preliminaries - Rendering Equation

- With geometry, albedo and lighting, we can render the image according to this equation:

$$C_S(p) = L^T \phi(n_p) \cdot \rho_p$$

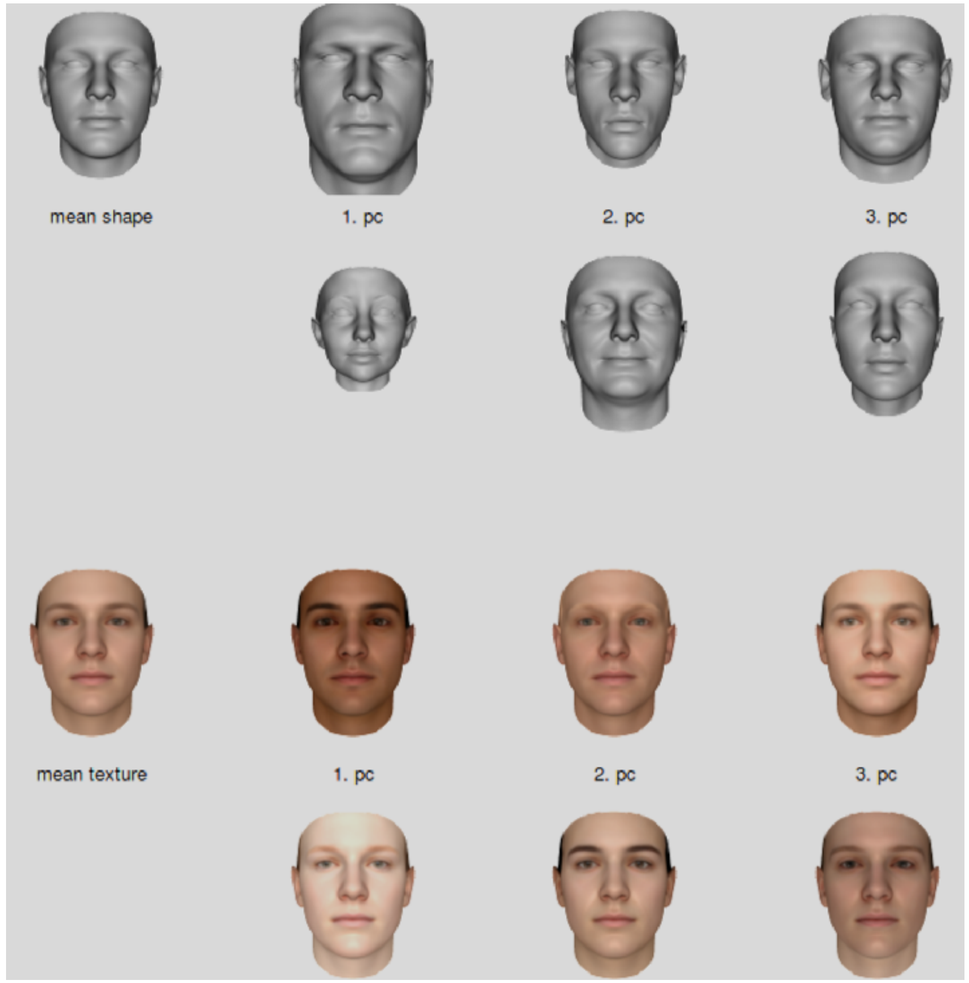
Image Lighting parameter Geometry Albedo

Preliminaries - 3D Face Representation

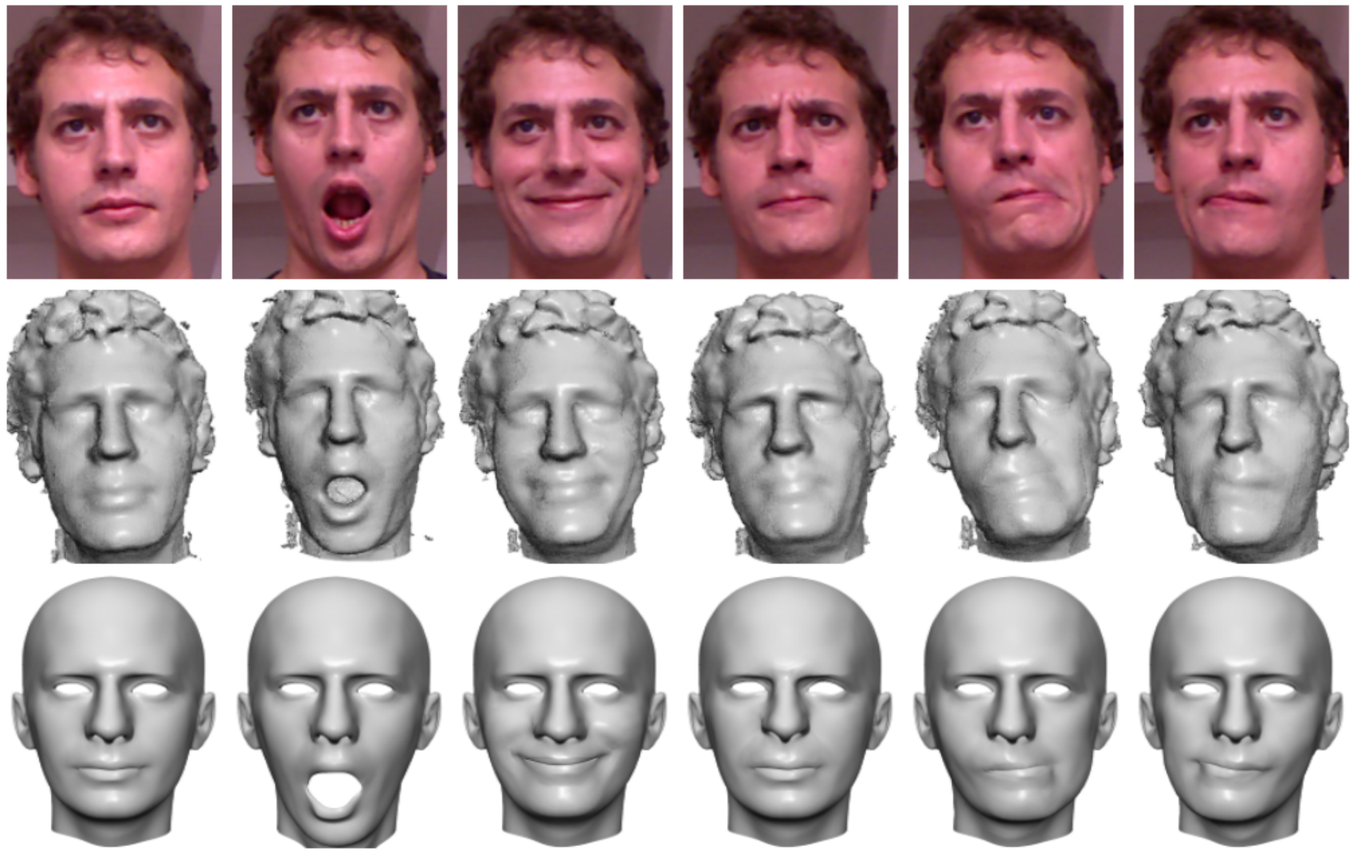
Mean Face Identity Expression Displacement

$$F = (\bar{F} + A_{\text{id}}\alpha_{\text{id}} + A_{\text{exp}}\alpha_{\text{exp}}) + F_{\text{disp}}$$

Parametric face models

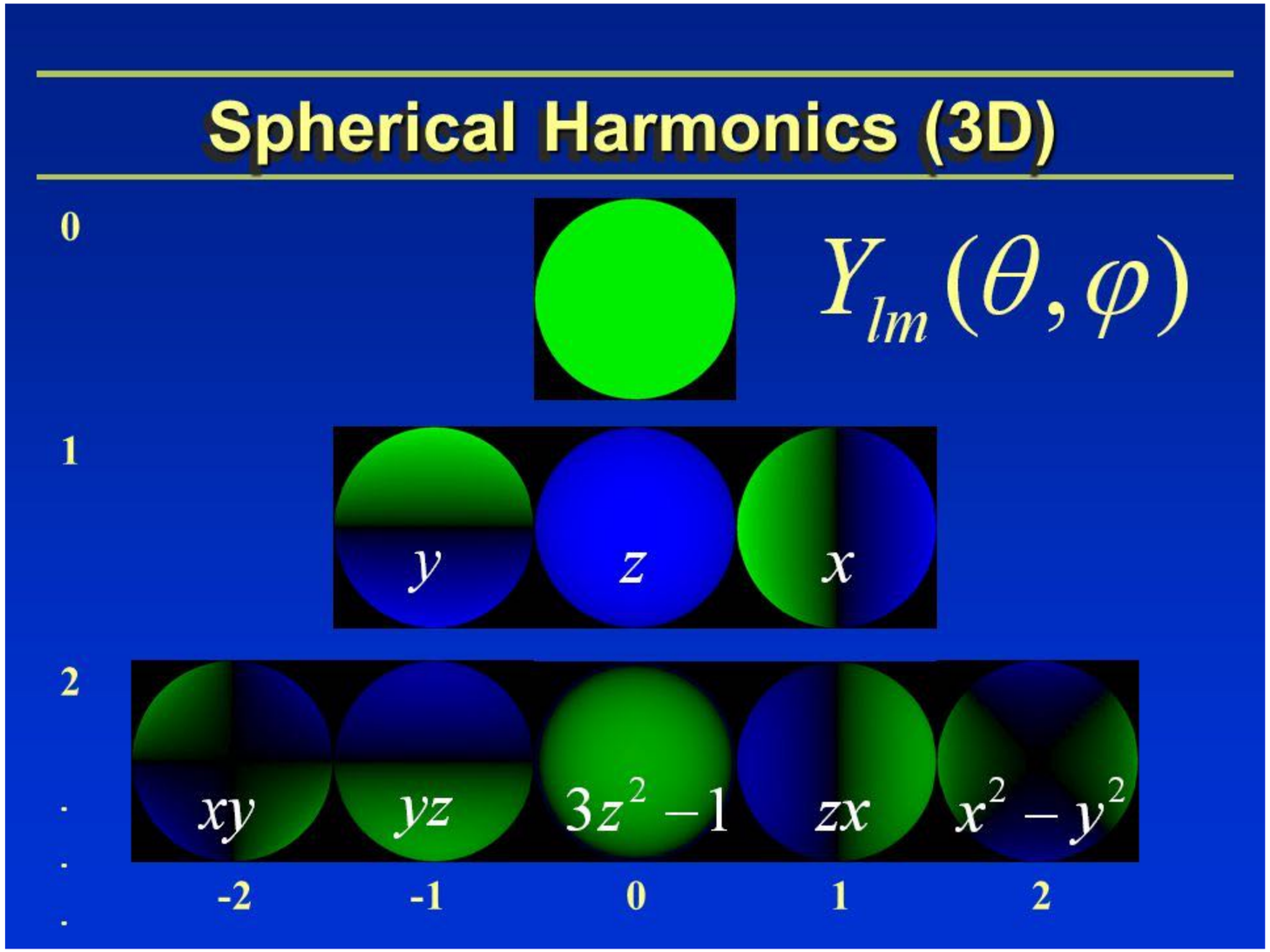


3DMM

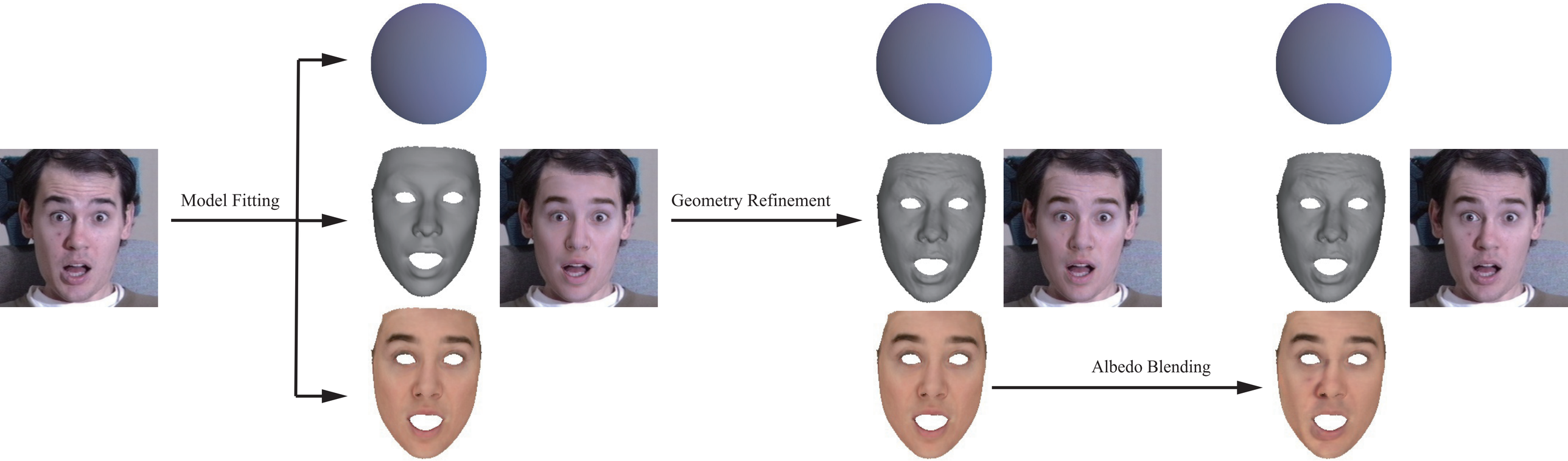


FaceWarehouse

Preliminaries - Lighting



Inverse Rendering Pipeline



Inverse Rendering - Coarse

$$\chi = \{\underbrace{\alpha_{\text{id}}, \alpha_{\text{exp}}}_{\text{Geometry}}, \underbrace{\alpha_{\text{alb}}}_{\text{Albedo}}, \underbrace{s, pitch, yaw, roll, t_x, t_y}_{\text{Pose}}, \underbrace{L}_{\text{Lighting}}\}$$

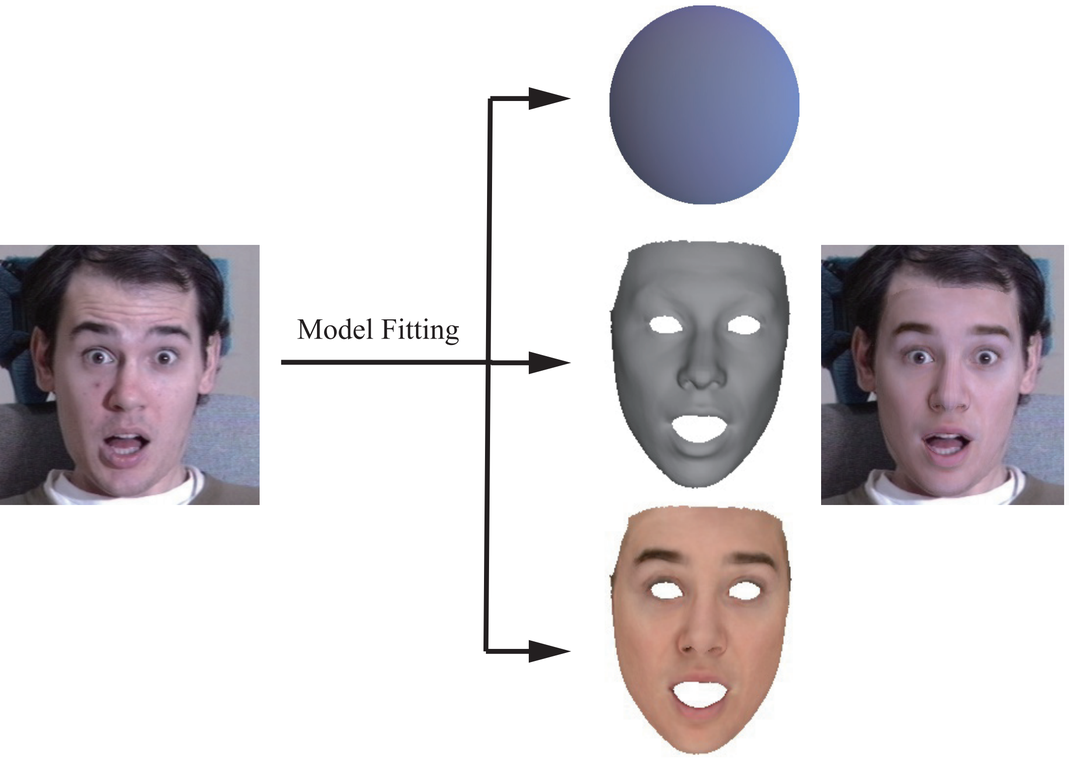
$$E(\chi) = E_{\text{con}} + w_{\text{lan}} E_{\text{lan}} + w_{\text{reg}} E_{\text{reg}}$$

$$E_{\text{con}}(\chi) = \frac{1}{|P|} \sum_{p \in P} \| C_S(p) - C_I(p) \|^2$$

$$E_{\text{lan}}(\chi) = \frac{1}{|\mathcal{F}|} \sum_{f_i \in \mathcal{F}} \| f_i - (\Pi R V_i + t) \|^2$$

$$E_{\text{reg}}(\chi) = \sum_{i=1}^{100} \left[\left(\frac{\alpha_{\text{id},i}}{\sigma_{\text{id},i}} \right)^2 + \left(\frac{\alpha_{\text{alb},i}}{\sigma_{\text{alb},i}} \right)^2 \right] + \sum_{i=1}^{79} \left(\frac{\alpha_{\text{exp},i}}{\sigma_{\text{exp},i}} \right)^2$$

Inverse Rendering - Details

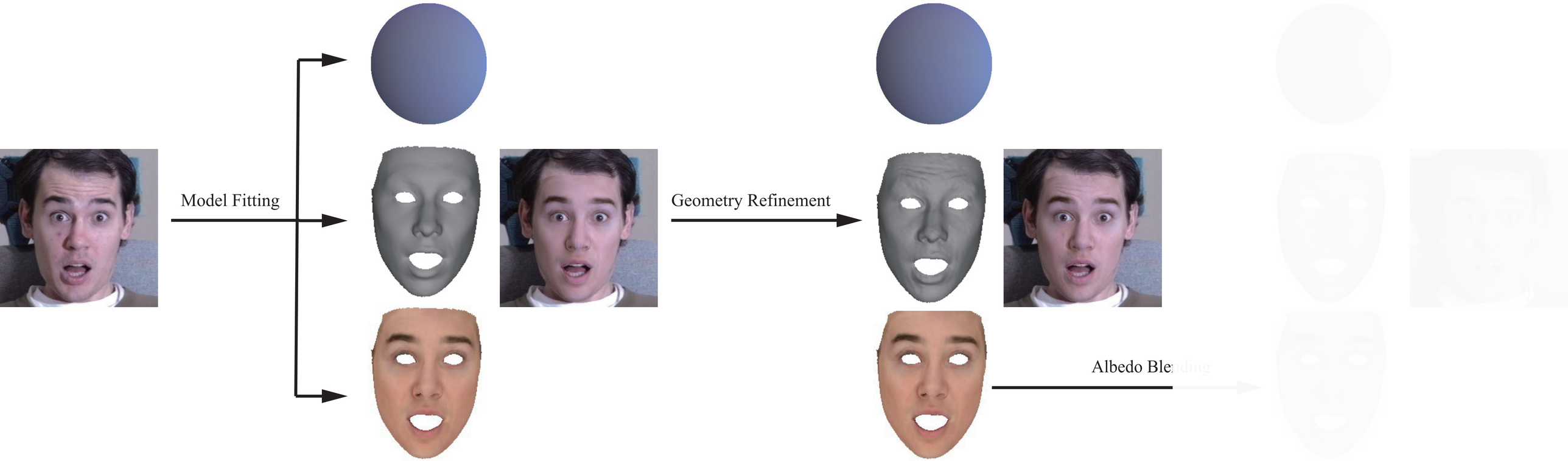


Inverse Rendering - Details

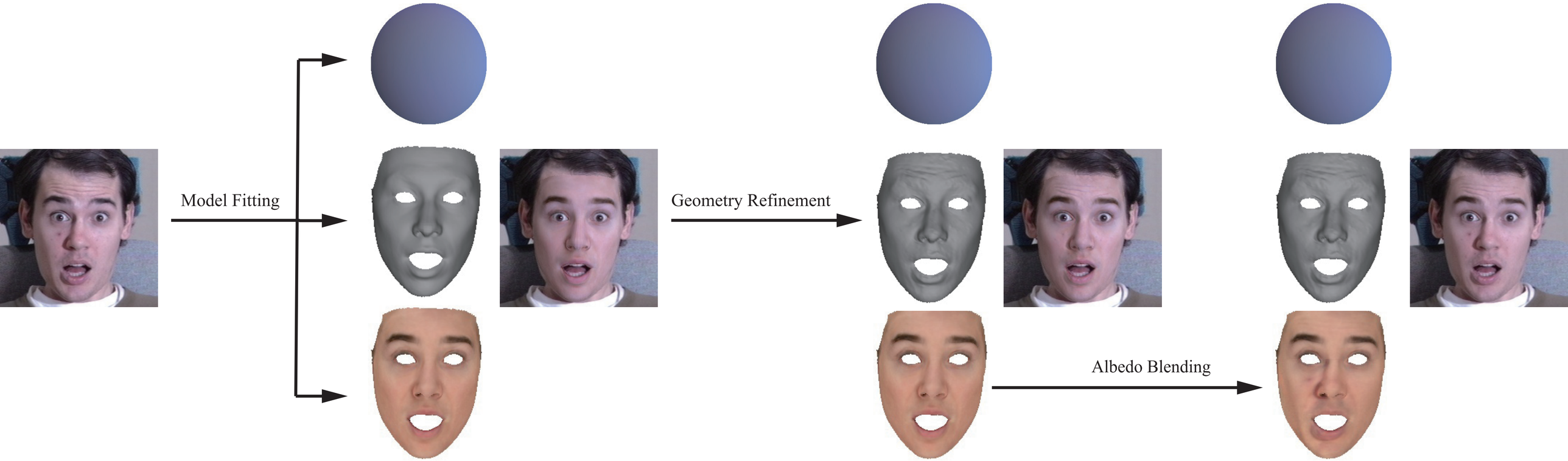
$$E(\mathbf{d}) = E_{\text{con}} + \mu_1 \|\mathbf{d}\|_2^2 + \mu_2 \|\mathbf{L}\mathbf{d}\|_1$$

$$E_{\text{con}}(\chi) = \frac{1}{|P|} \sum_{p \in P} \|C_S(p) - C_I(p)\|^2$$

Inverse Rendering - Geometry



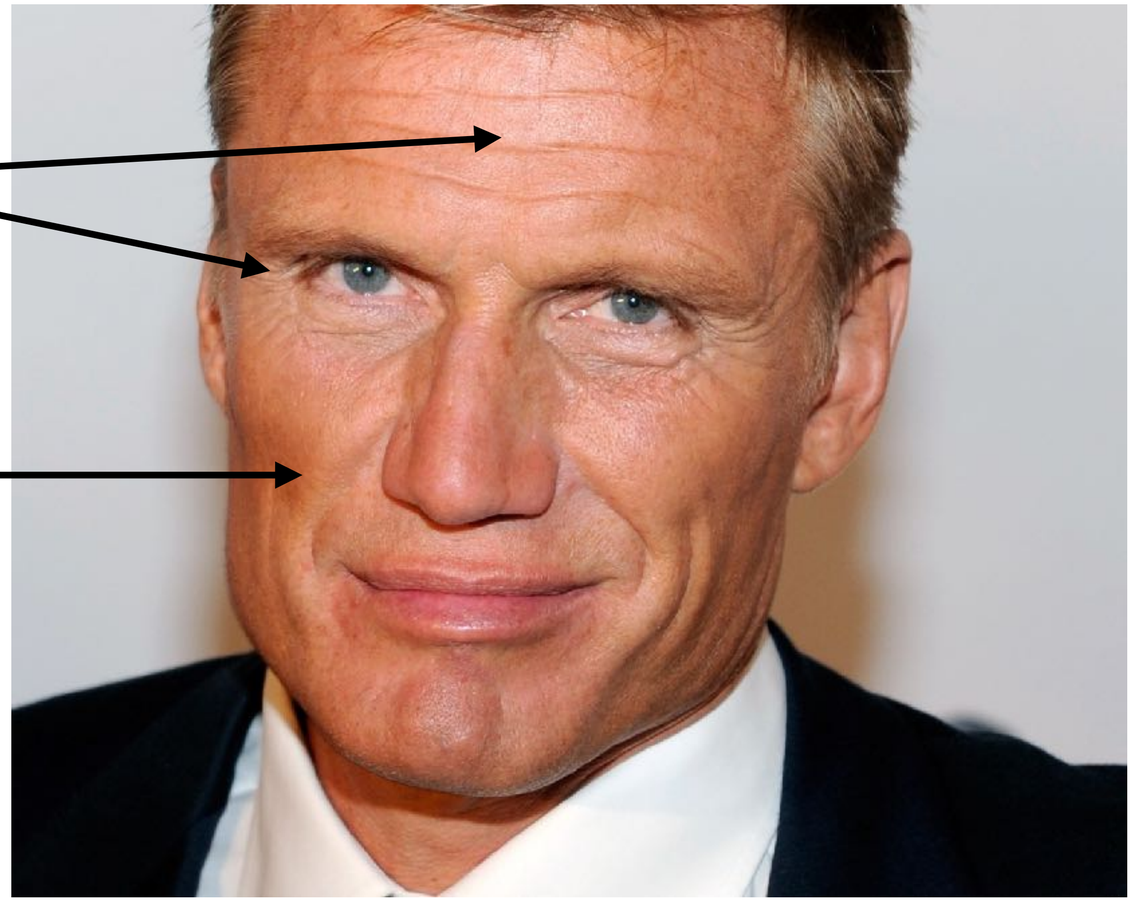
Inverse Rendering - Albedo



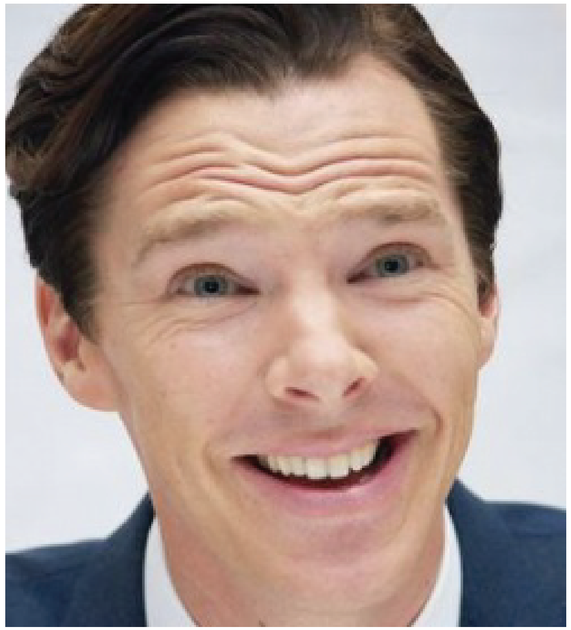
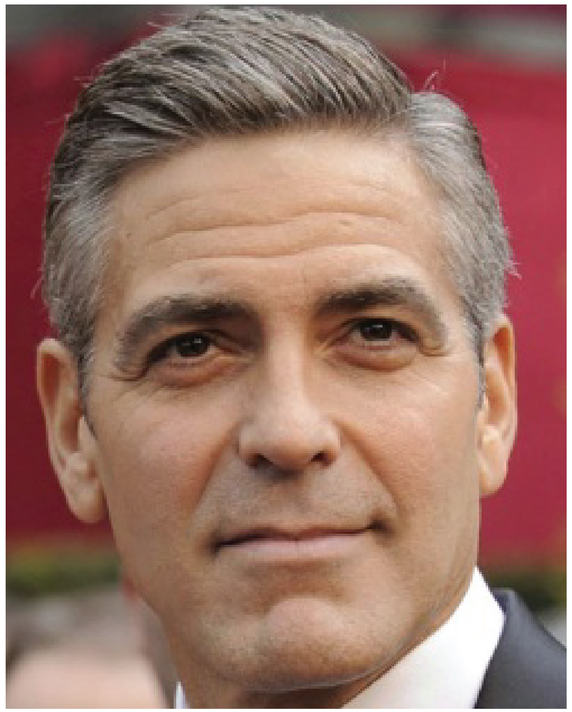
Inverse Rendering - Albedo

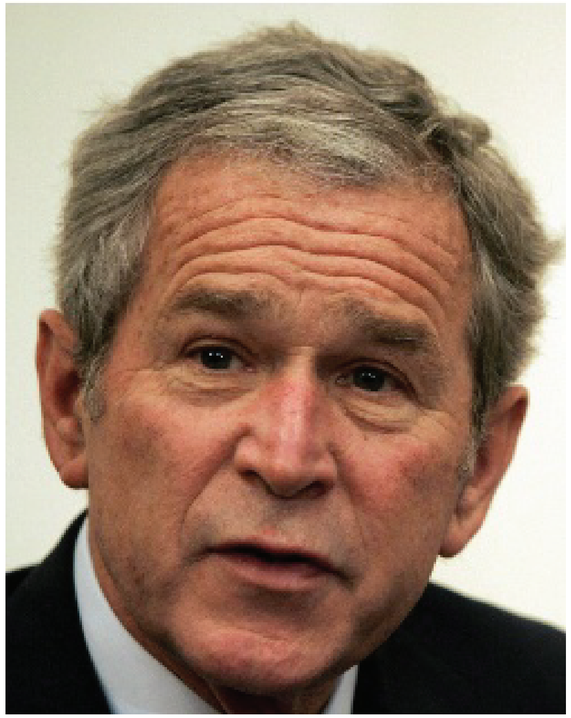
$$\rho_f = \frac{C_I(p)}{L^T \phi(n_p)}$$

$$\beta \rho_c + (1 - \beta)\rho_f$$



Input Images



Coarse Results



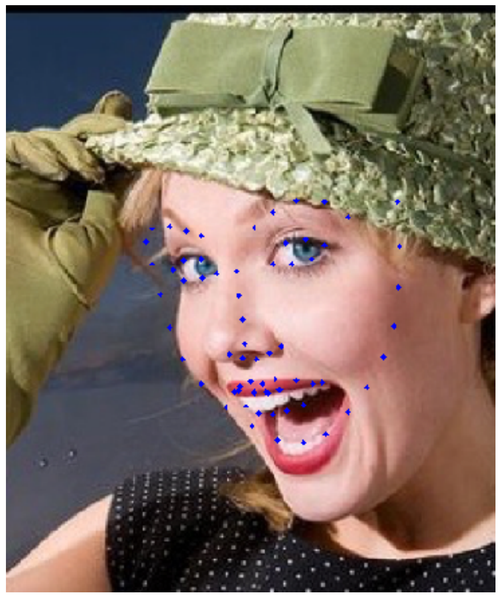

Fine Results




Limitation of Inverse Rendering

- The total computation time is 8s on a desktop with a quad-core Intel CPU i7, 4GB RAM and NVIDIA GTX 1070 GPU.

- It might fail for challenging cases like large pose face images.



Landmarks

Result

Optimization → CNN

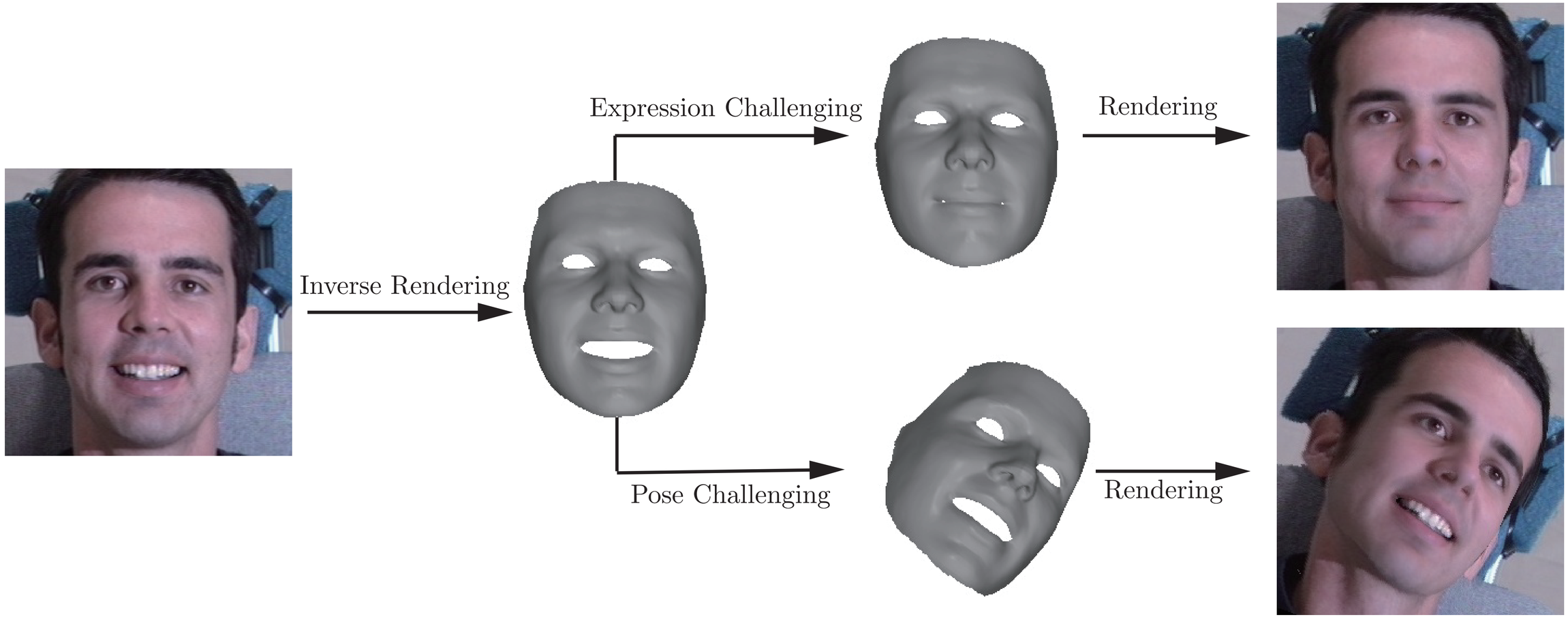
CNN-based Real-time Dense Face Reconstruction with
Inverse-rendered Photo-realistic Face Images
IEEE Trans on PAMI, 2018

Proposed Solution

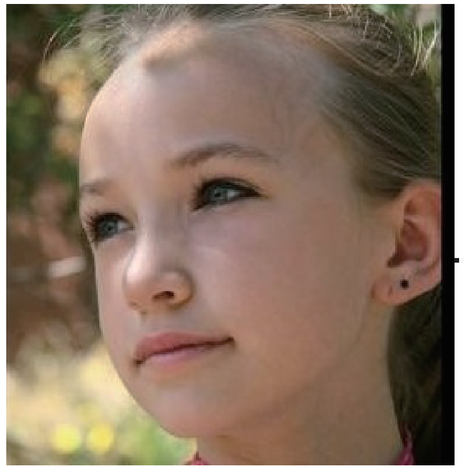
- Synthesize large-scale training pairs including input image and output 3D face models.
- A two layers network: coarse network to train the 3DMM parameters, and fine network to train the depth displacement.
- Do data augmentation such that the network is robust to challenging cases.

Pipeline

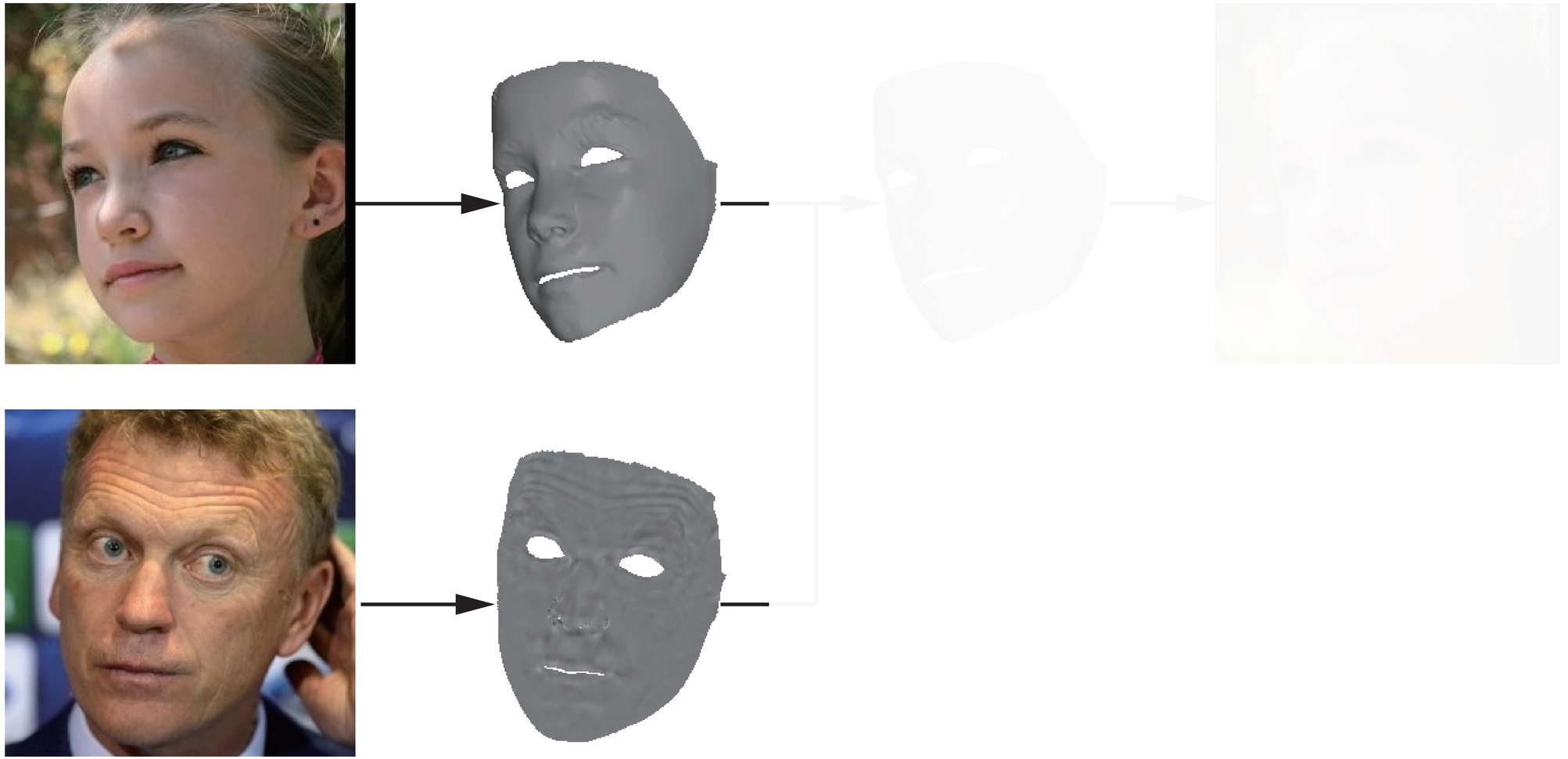
Data Augmentation - Coarse



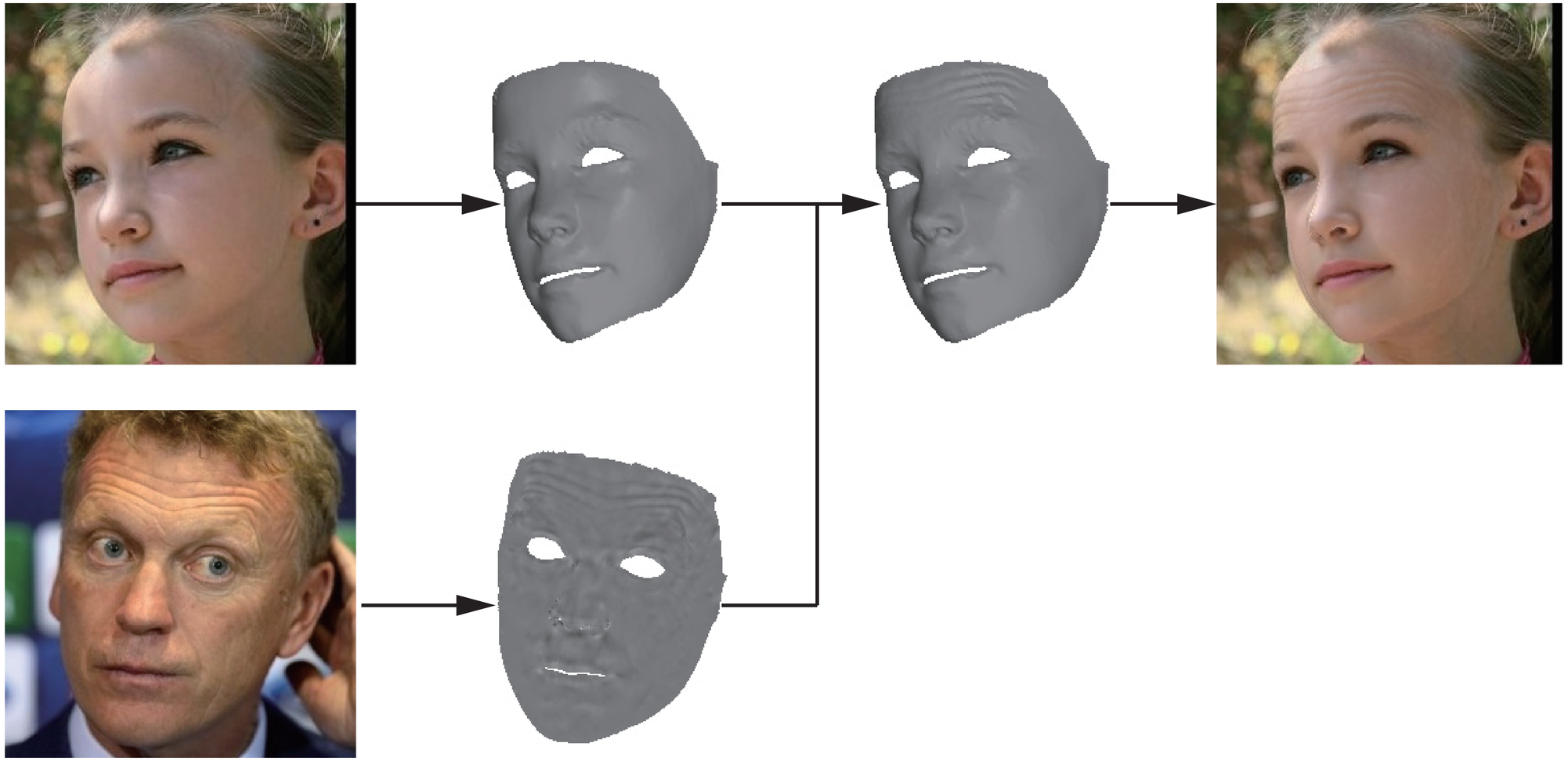
Data Augmentation - Fine



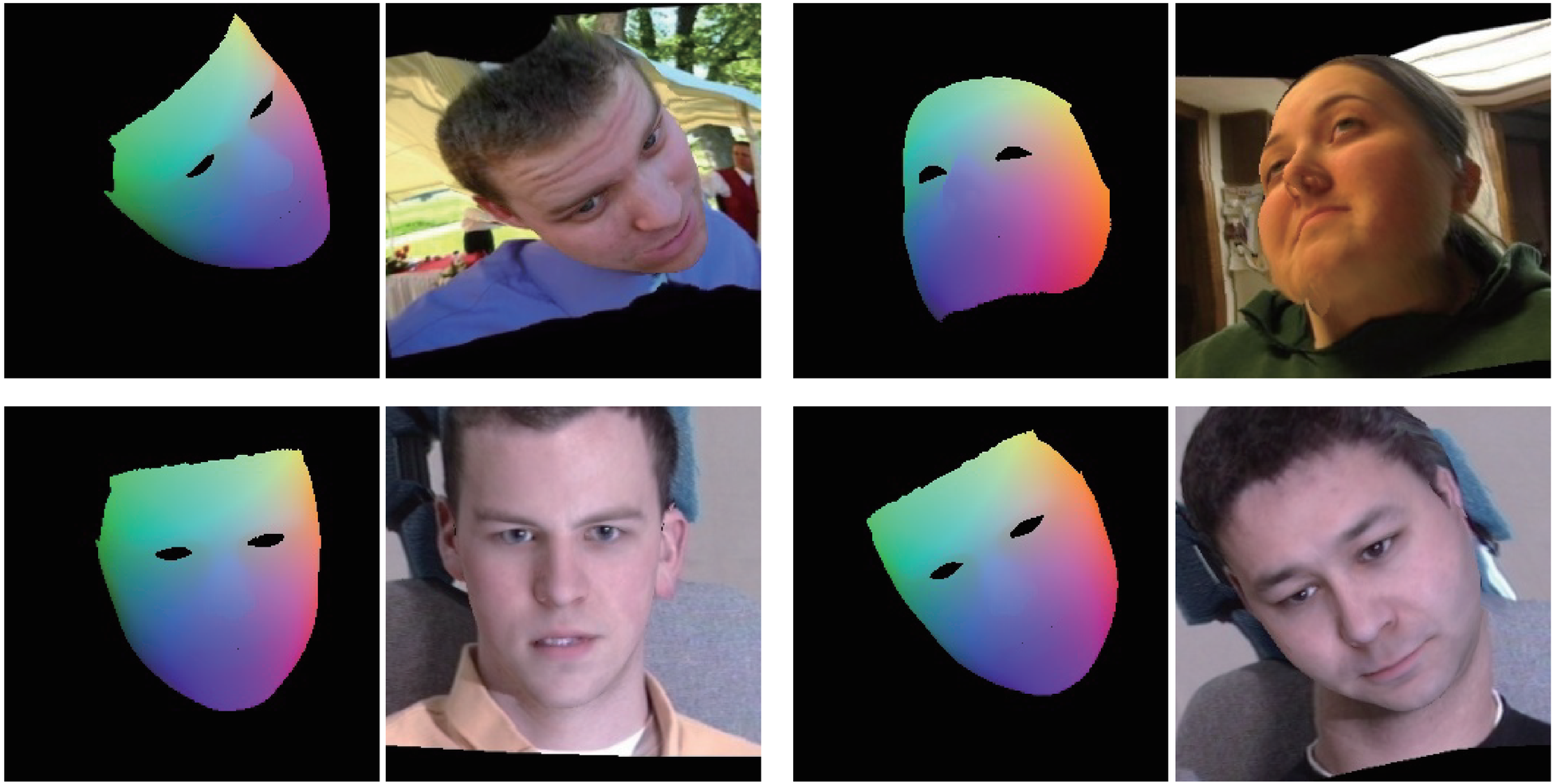
Data Augmentation - Fine



Data Augmentation - Fine



Data Augmentation - Video

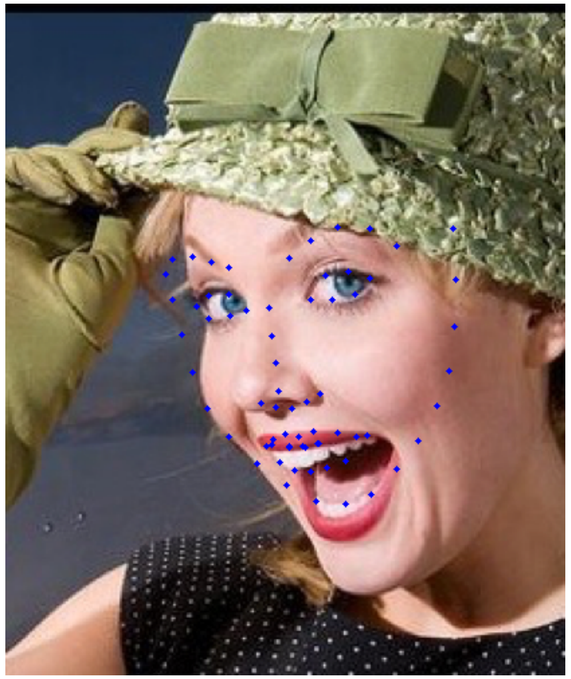


Comparison: Optimization vs CNN

Optimization

CNN

Comparison: Optimization vs CNN



Landmarks

Optimization

CNN

Result - Comparison

- [Garrido et al.2016] costs 175.5s for each frame.
- Ours costs 20ms for each frame.

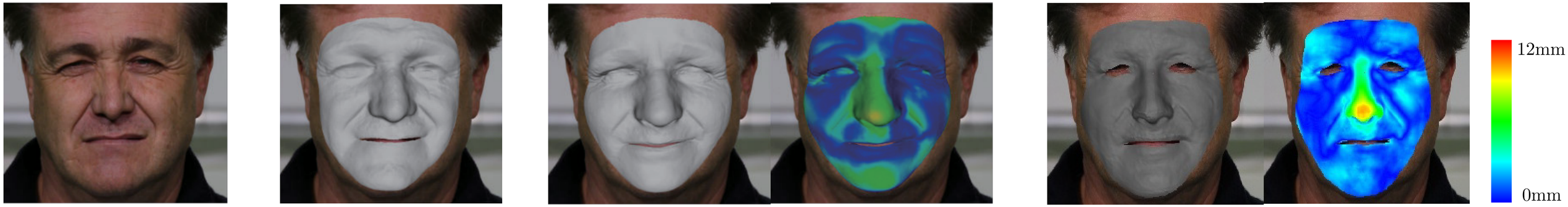
Input

[Garrido et al. 2016]

Ours

Pablo Garrido, et.al. Reconstruction of personalized 3d face rigs from monocular video. TOG, 2016.

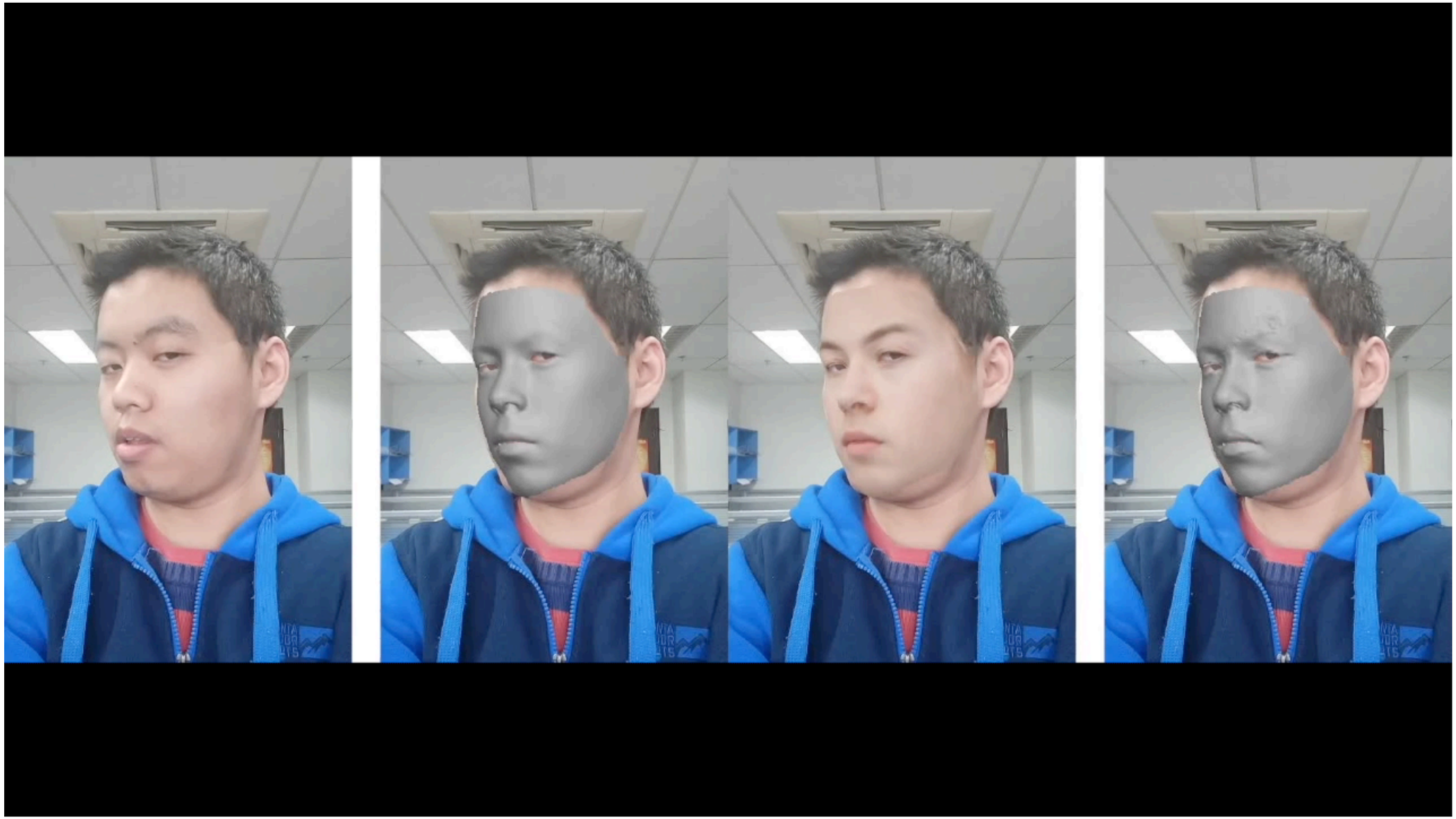
Comparison with GroundTruth



Input Stereo [Garrido et al. 2016] Ours

- The mean reconstruction error is 1.96mm compared to the binocular facial performance capture.
- Comparable with optimization based approach (1.96mm vs. 1.8mm) while with much less time.

Real-time facial performance capture



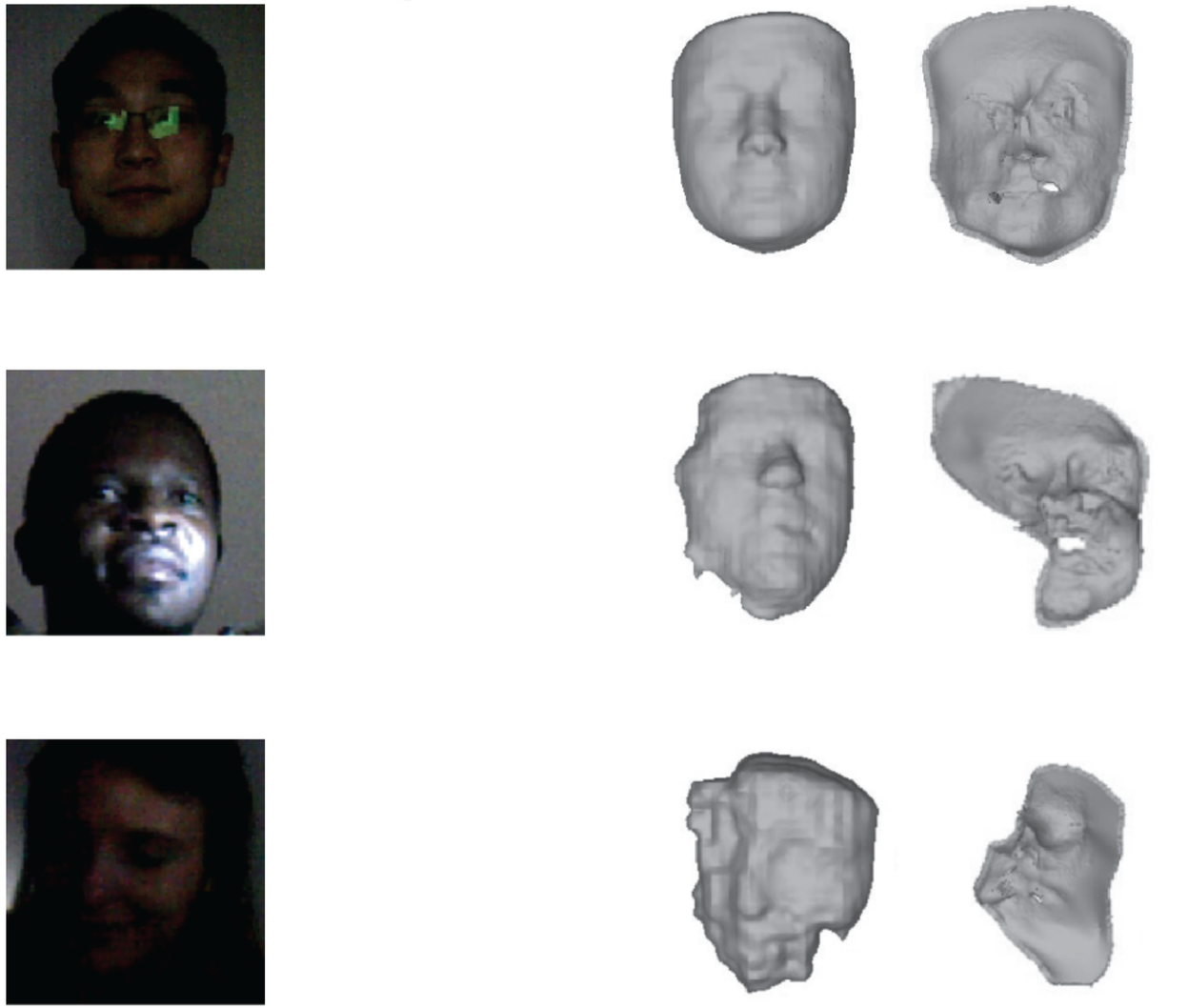
Application in VR

Real-time 3D Face-Eye Performance Capture of a PersonWearing VR Headset
ACM Multimedia (Full research paper), 2018

3D Caricature Reconstruction

Alive Caricature from 2D to 3D
CVPR 2018

Limitation of RGB-based Methods



Input [Jackson et al. 2017] [Sela et al. 2017]

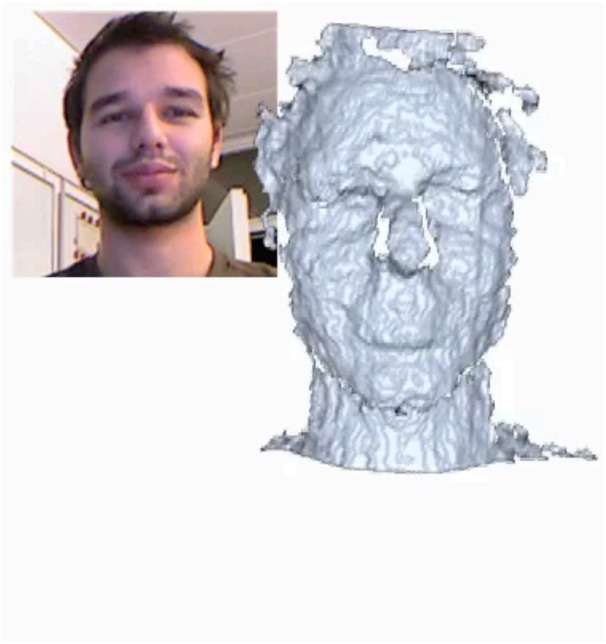
RGB ⟶ RGB-D

Supervised ⟶ Self-Supervised

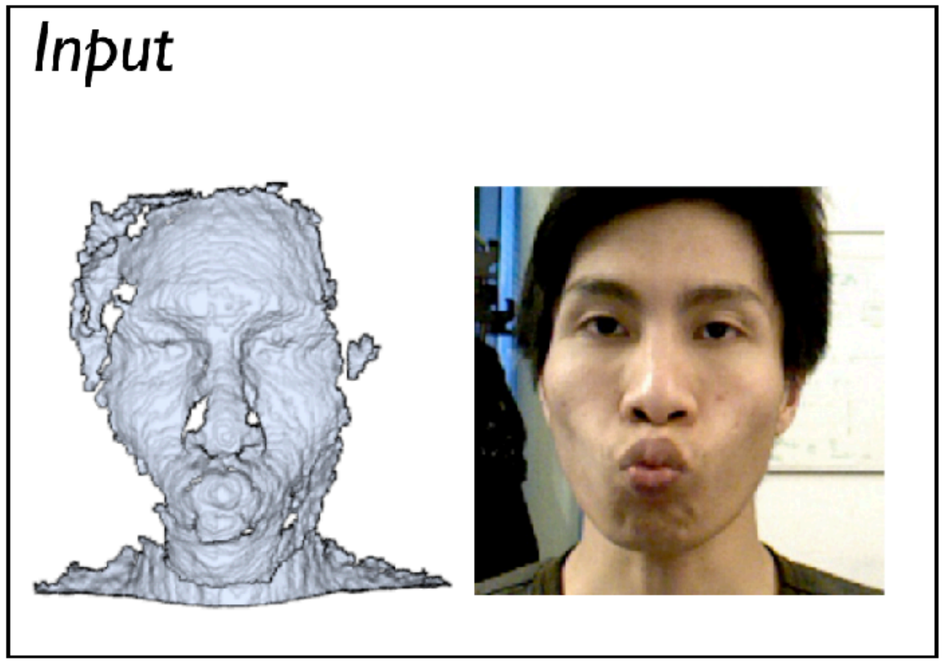
Self-supervised CNN for Unconstrained 3D Facial Performance Capture from an RGB-D Camera

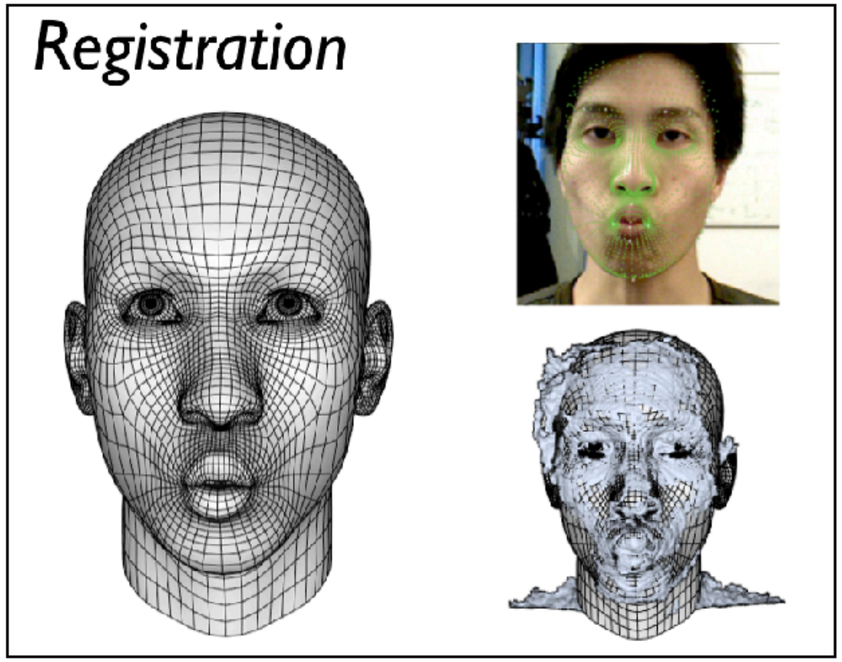
https://arxiv.org/abs/1808.05323

Depth Sensor



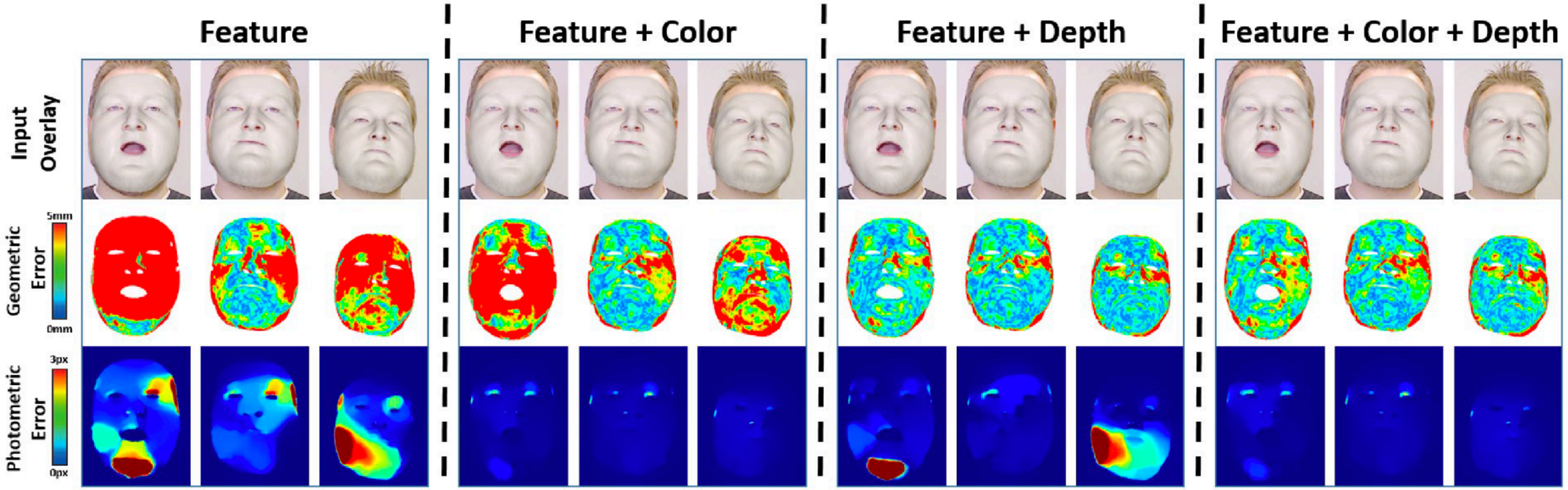
Sofien. B, et.al. Online Modeling For Realtime Facial Animation, Siggraph 2013





J. Thies, et.al. Real-time Expression Transfer for Facial Reenactment, Siggraph Asia 2015

$$E(\boldsymbol{\mathcal{P}}) = E_{\mathrm{emb}}(\boldsymbol{\mathcal{P}}) + w_{\mathrm{col}} E_{\mathrm{col}}(\boldsymbol{\mathcal{P}}) + w_{\mathrm{lan}} E_{\mathrm{lan}}(\boldsymbol{\mathcal{P}}) + w_{\mathrm{reg}} E_{\mathrm{reg}}(\boldsymbol{\mathcal{P}})$$

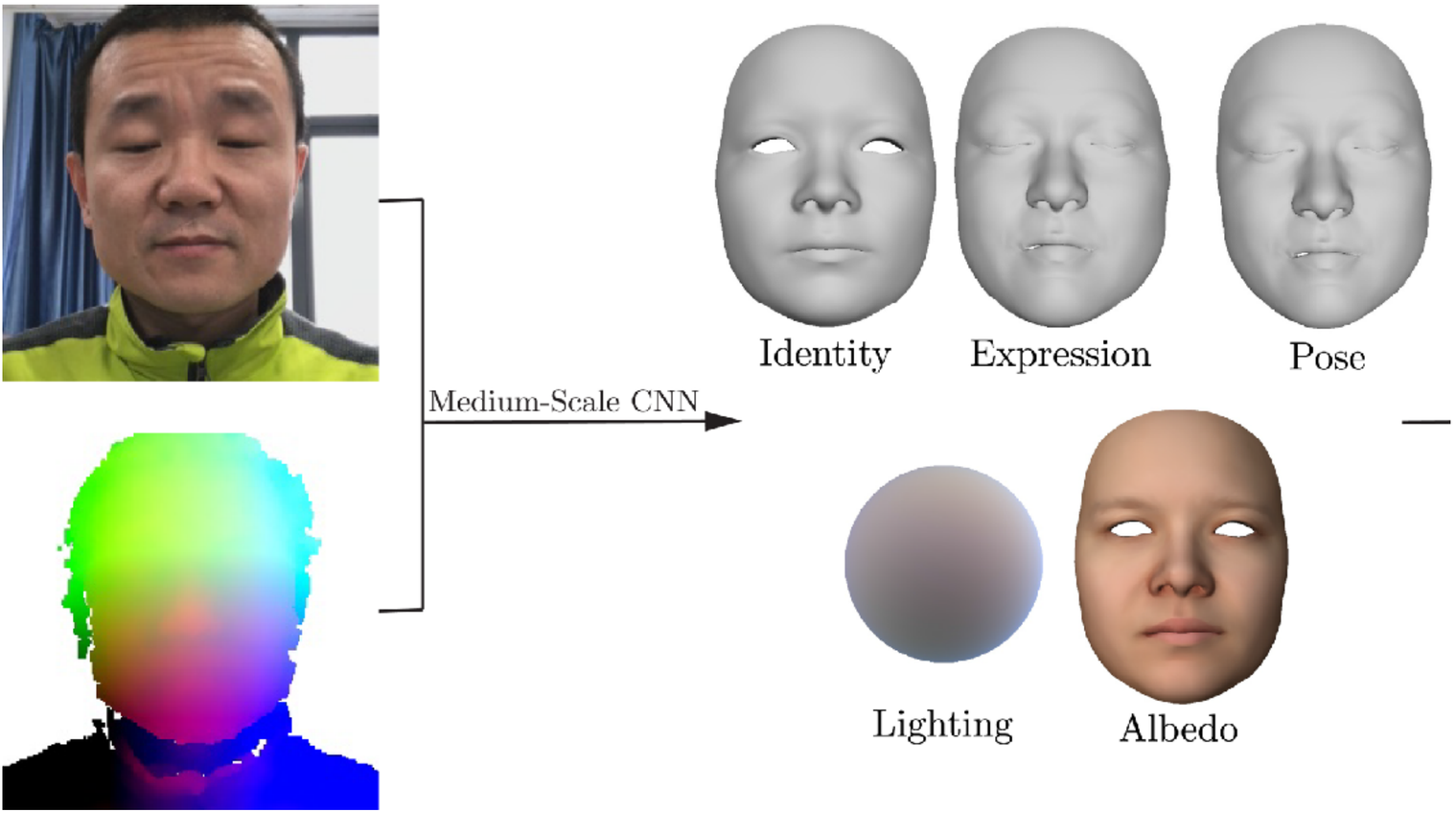


Limitations of Existing Methods

- The 3D face modeling is formulated as an optimization problem, which includes the following steps. **Hard to code!**
  - depth to point cloud
  - rigid registration: ICP problem
  - non-rigid registration: Sparse optimization
  - blendshape refinement
- **High computation cost**. Not easy to port it to mobile platform.

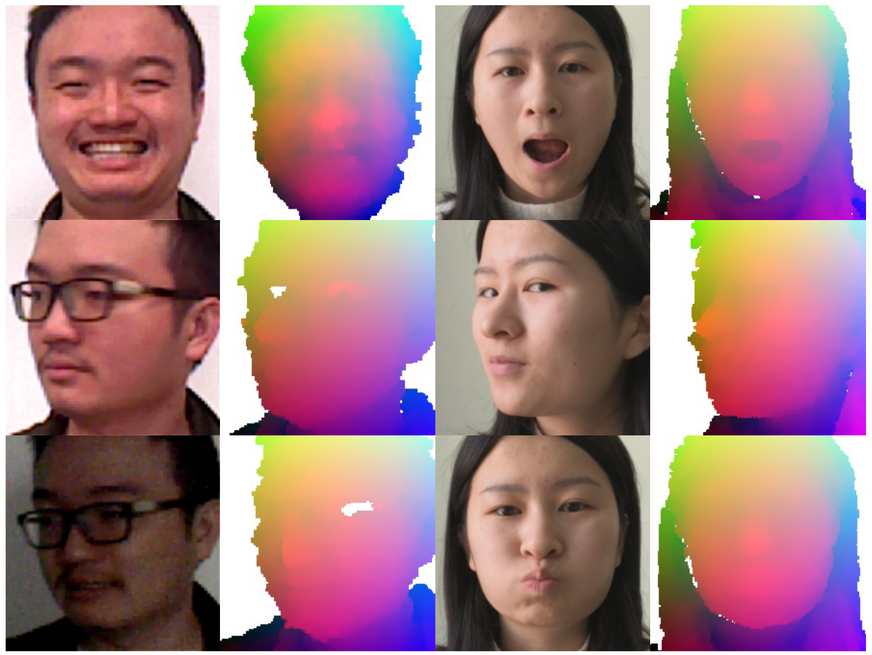
Our Proposed Pipeline

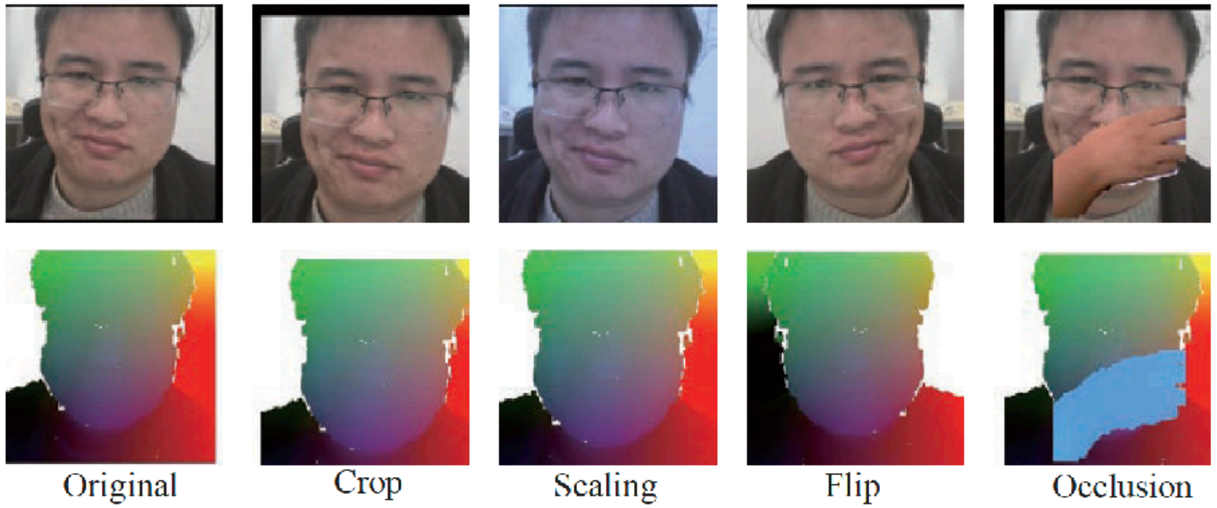
Our Proposed Pipeline



Training data

- 600 people, and around 300k RGB-D frames
- Data augmentation, including flip, scaling, occlusion





Self-supervised Learning

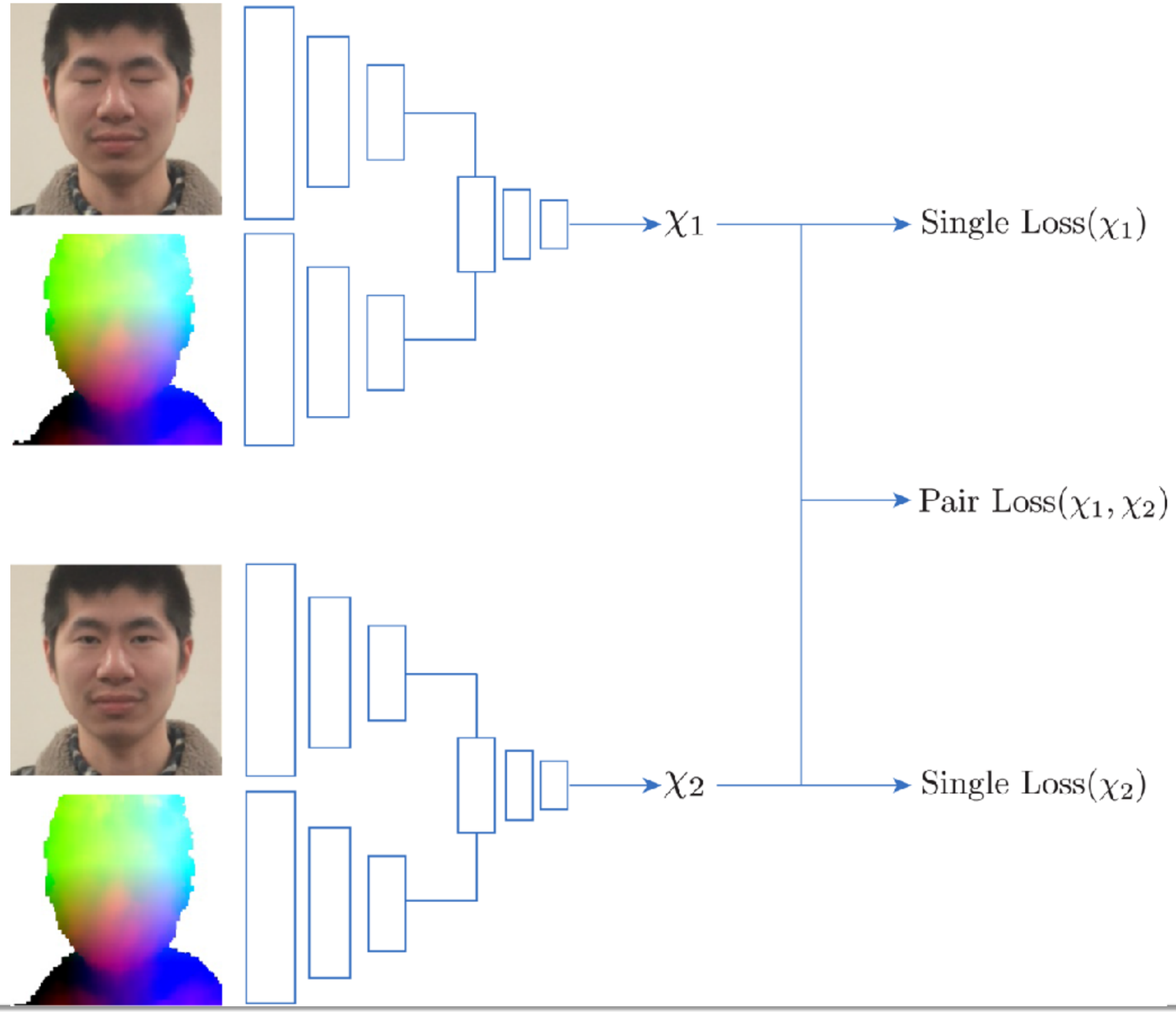
- Offline training the CNN model, real-time tracking during testing
- A novel strategy to optimize the rigid registration energy in CNN training

$$E_{\text{loss}} = \underbrace{E_{\text{geo}} + w_{\text{col}}E_{\text{col}} + w_{\text{lan}}E_{\text{lan}} + w_{\text{reg}}E_{\text{reg}}}_{\text{Single Loss}} + \underbrace{w_{\text{flow}}E_{\text{flow}} + w_{\text{same}}E_{\text{same}}}_{\text{Pair Loss}}$$

$$E_{\text{flow}}(\chi_n, \chi_{n-1}) = \frac{1}{|\mathcal{F}|}\sum_{m\in\mathcal{F}} \|(Pr_n(\mathbf{p}) - Pr_{n-1}(\mathbf{p})) - f(m)\|_2^2,$$

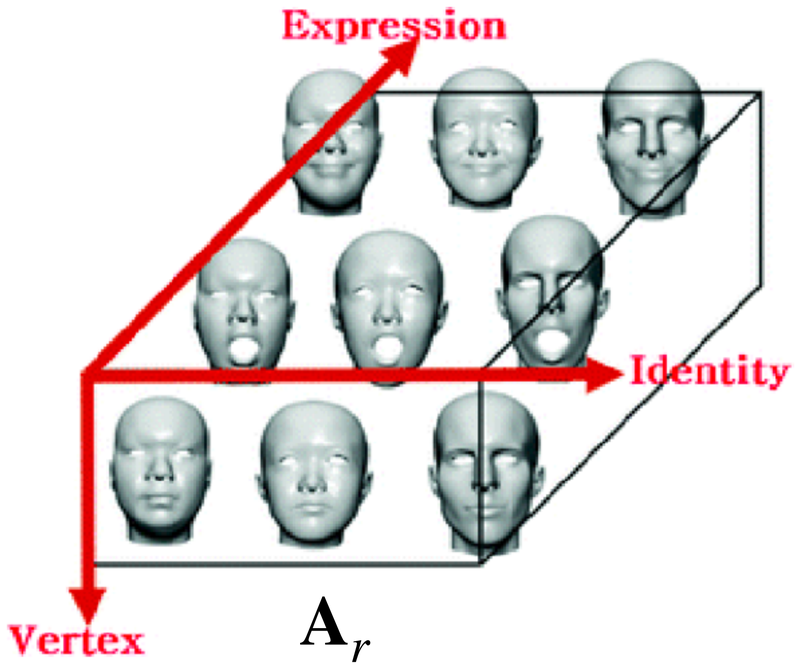
$$E_{\text{same}}(\chi_{n_1}, \chi_{n_2}) = \left(\|\boldsymbol{\alpha}_{\text{id},n_1} - \boldsymbol{\alpha}_{\text{id},n_2}\|_2^2 + \|\boldsymbol{\alpha}_{\text{alb},n_1} - \boldsymbol{\alpha}_{\text{alb},n_2}\|_2^2\right)$$

Medium-scale network architecture



3D Face Representation (fixed Basis)

$$\mathbf{p} = \mathbf{A}_r \times_2 \alpha_{id} \times_3 \alpha_{exp}$$

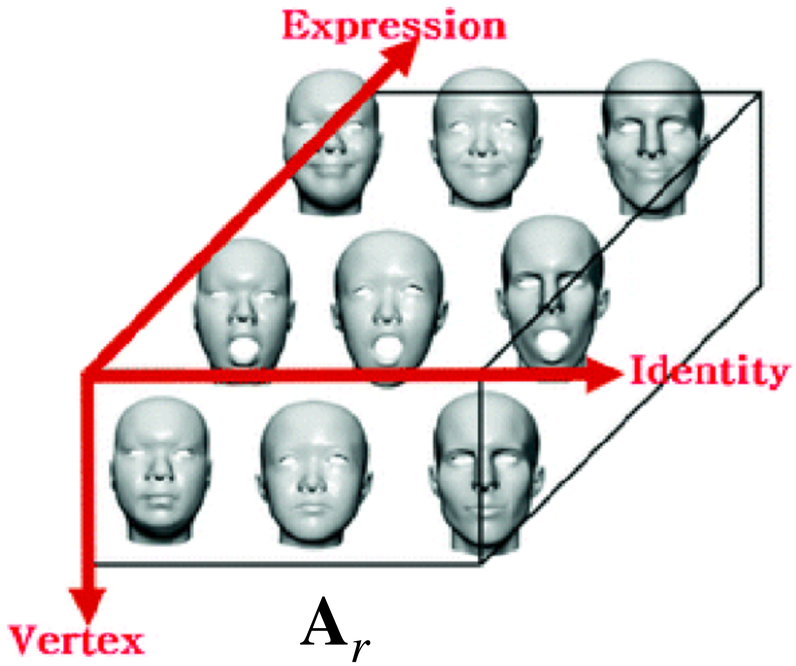


$$\chi = \{\alpha_{id}, \alpha_{exp}, \alpha_{alb}, pitch, yaw, roll, \mathbf{t}, \mathbf{r}\}$$

$$E_{\text{loss}}(\chi) = \underbrace{E_{\text{geo}} + w_{\text{col}}E_{\text{col}} + w_{\text{lan}}E_{\text{lan}} + w_{\text{reg}}E_{\text{reg}}}_{\text{Single Loss}} + \underbrace{w_{\text{flow}}E_{\text{flow}} + w_{\text{same}}E_{\text{same}}}_{\text{Pair Loss}}$$

3D Face Representation (refined Basis)

$$\mathbf{p} = \mathbf{A}_r \times_2 \alpha_{id} \times_3 \alpha_{exp}$$



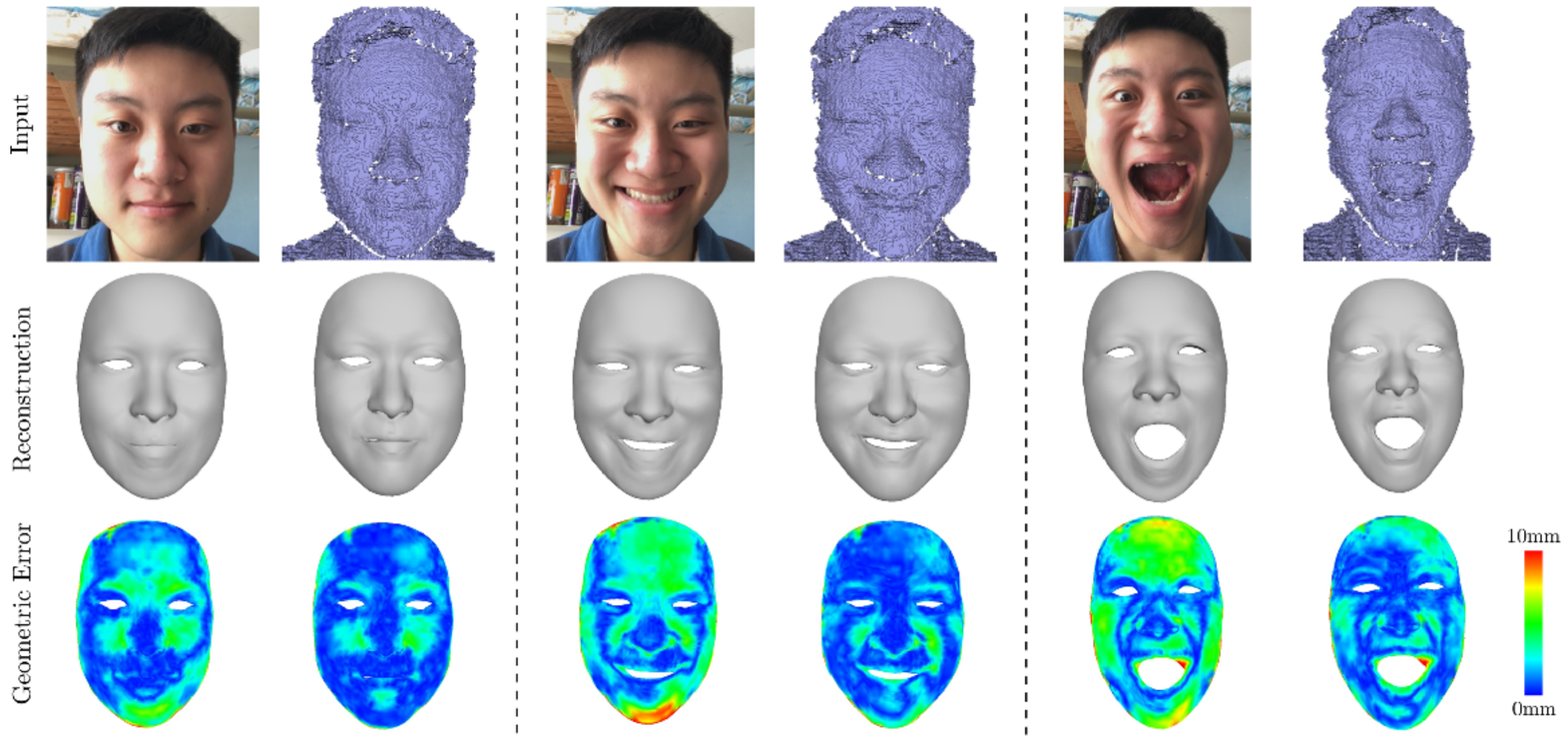
$$\chi = \{\alpha_{id}, \alpha_{exp}, \alpha_{alb}, pitch, yaw, roll, \mathbf{t}, \mathbf{r}\}$$

$$E_{\text{loss}}(\mathbf{A}_r, \chi) = \underbrace{E_{\text{geo}} + w_{\text{col}}E_{\text{col}} + w_{\text{lan}}E_{\text{lan}} + w_{\text{reg}}E_{\text{reg}}}_{\text{Single Loss}} + \underbrace{w_{\text{flow}}E_{\text{flow}} + w_{\text{same}}E_{\text{same}}}_{\text{Pair Loss}}$$

Fixed vs Refined

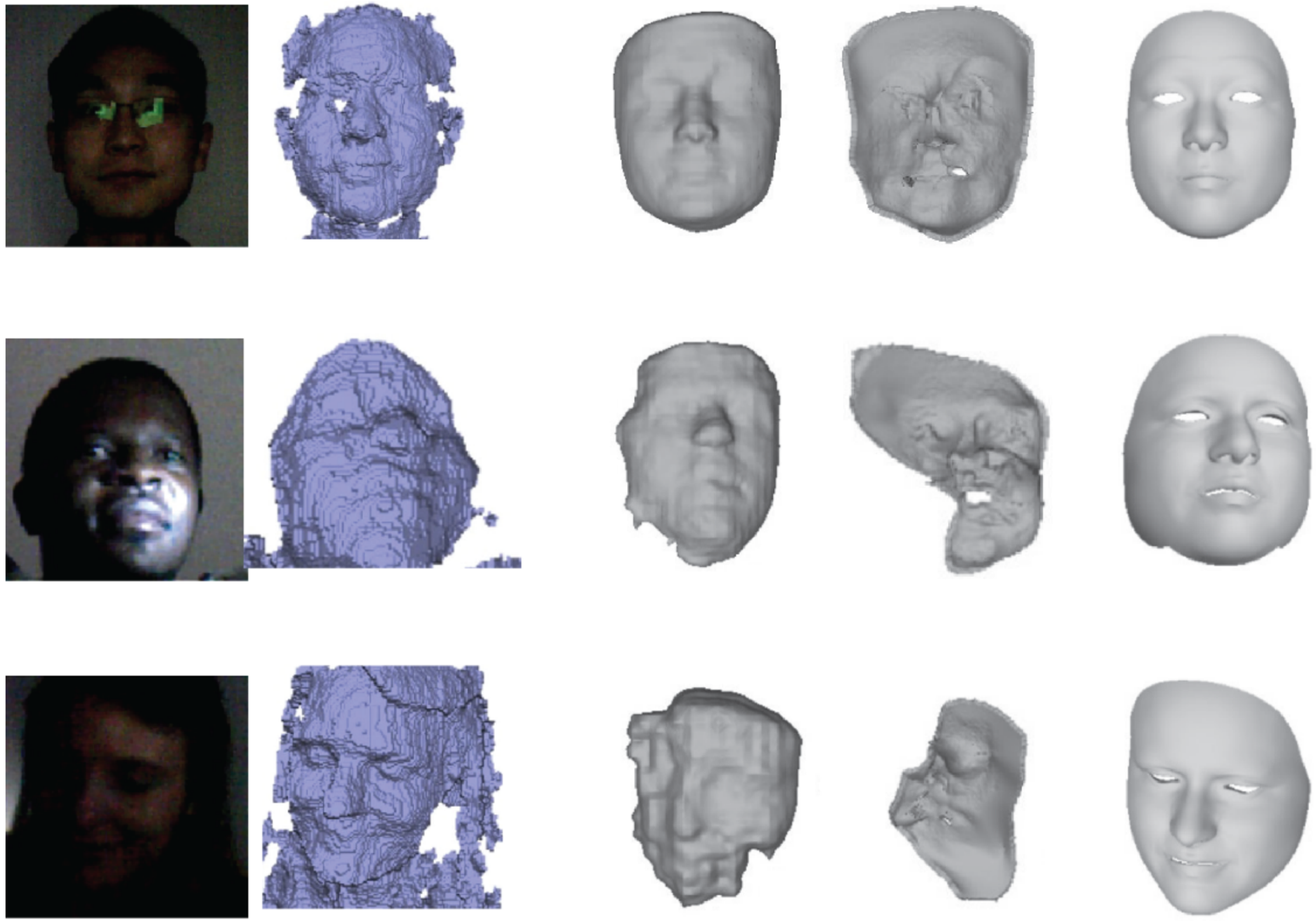
- The parametric shape model is not fixed, which is refined during training
- Regularization to the refined model:

$$E_{\text{regA}}(A'_r) = w_{\text{Areg}} \|A'_r - A_r\|^2 + w_{\text{Asmo}} \| \triangle(A'_r - A_r)\|^2$$



Fixed: 2.33mm / 0.45mm
Refined: 1.91mm / 0.23mm

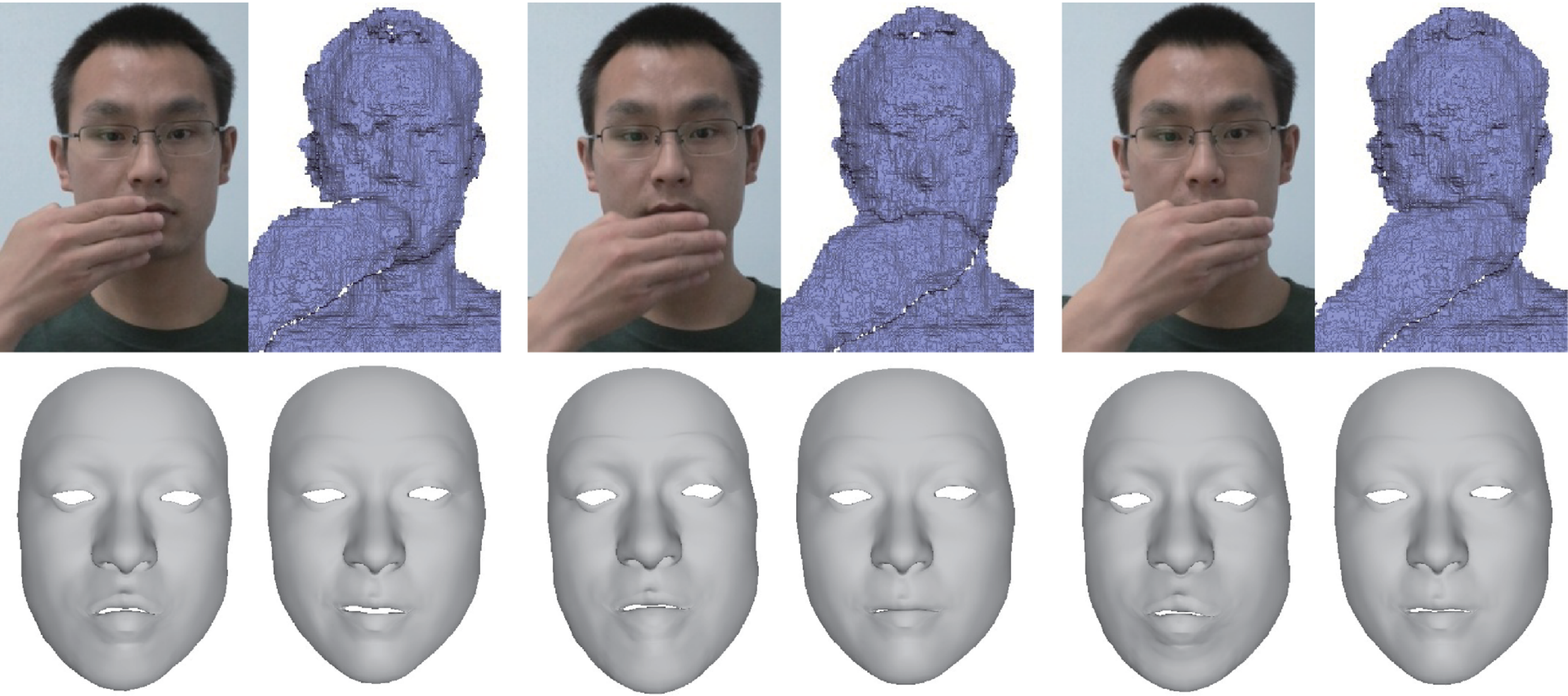
Results of RGB-D based method



Input [Jackson et al. 2017] [Sela et al. 2017] Ours

Robustness

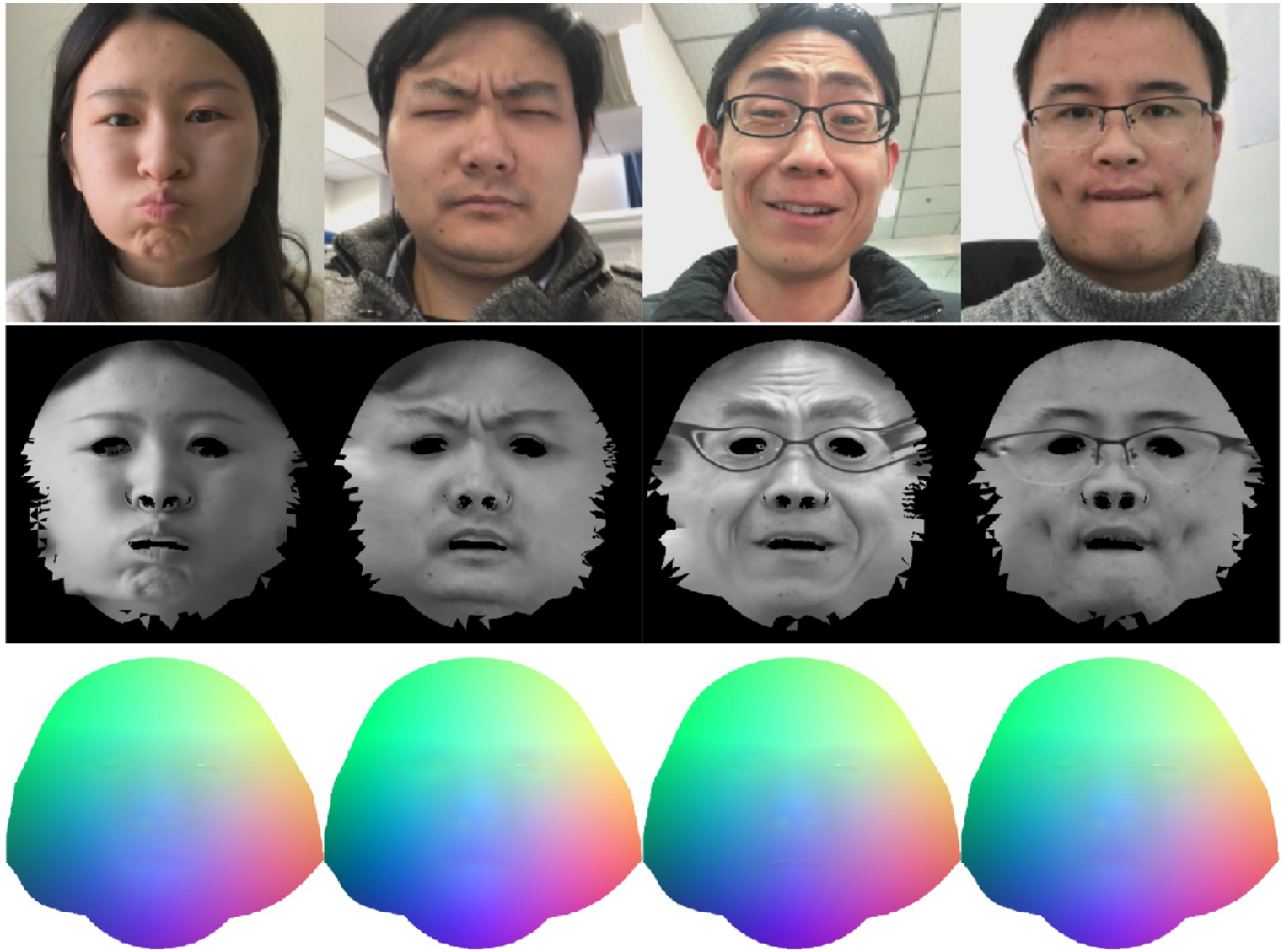
Occlusion



Geometry Details

$$\hat{\mathbf{p}} = \mathbf{p} + \mathbf{n} \cdot d$$

Geometry Details - representation



Geometry Details - loss

Shading Term: $$w_{\text{face}} \sum_{i \in \mathcal{T}} \|\mathbf{I}(\mathbf{n}_i \,|\, b_i, \gamma) - c_i\|^2 + w_{\text{edge}} \sum_{i \bigcap j \in \mathcal{E}} \|(\mathbf{I}(\mathbf{n}_i \,|\, b_i, \gamma) - \mathbf{I}(\mathbf{n}_j \,|\, b_j, \gamma)) - (c_i - c_j)\|^2$$

Coherence Term: $$w_{cl} \sum_{i \in \mathcal{F}} \|\mathbf{n}_{n,i} - \mathbf{n}_{n-1,i}\|_2^2$$

Regularization Term: $$w_{\text{sm}} \sum_{v_i \in V} \|\Delta \mathbf{p}_i\|_2 + w_{\text{mi}} \|\mathbf{d}\|_2^2$$

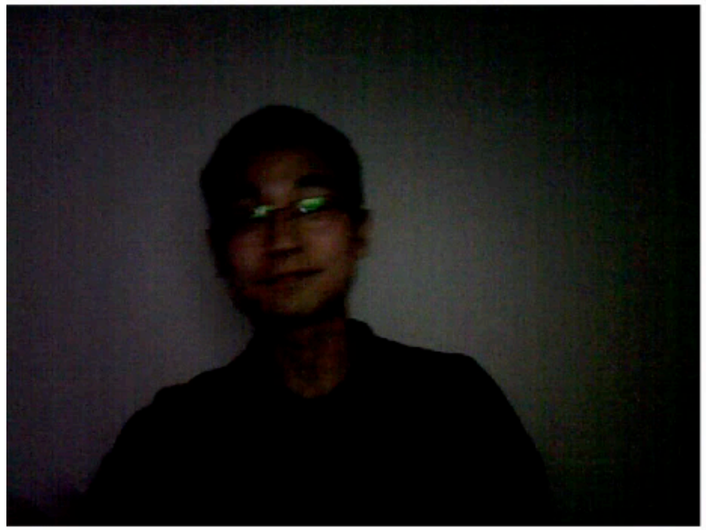
Geometry Details - results

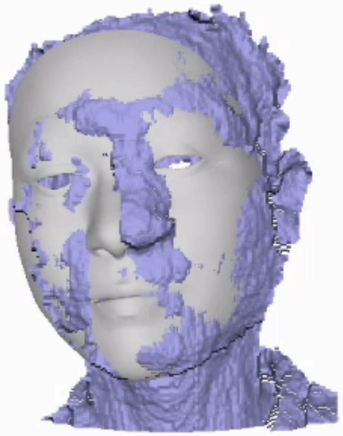
Input

Coarse

Fine

Demo





Acknowledgments

Yudong Guo, Luo Jiang, Boyi Jiang, Lin Cai, Hao Li
Jianfei Cai, Jianmin Zheng, Bailin Deng, Yu-kun Lai, Ligang Liu

NANYANG TECHNOLOGICAL UNIVERSITY

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

NSFC

Thank you!