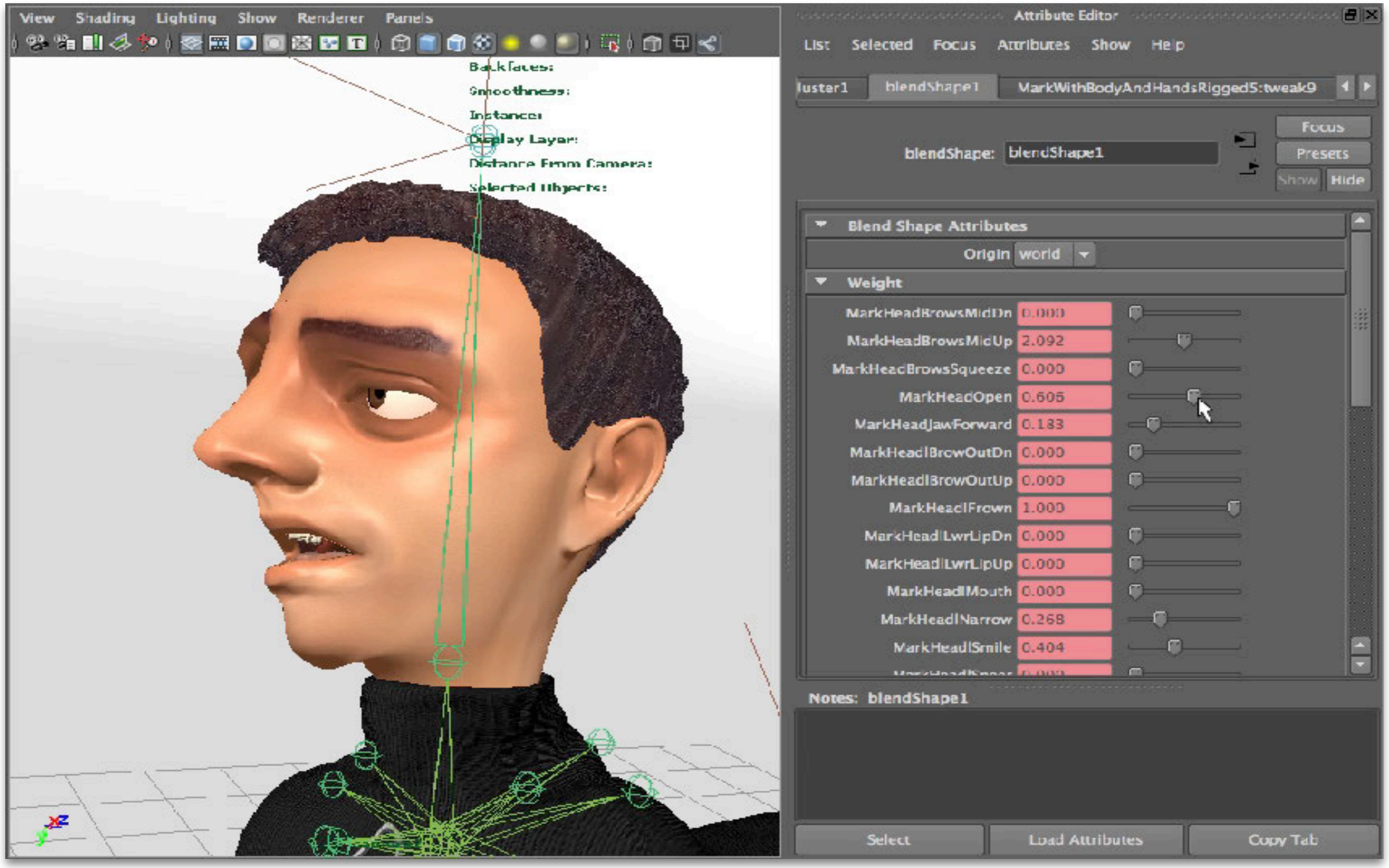
Real-time 3D Face Reconstruction with Geometry Details

张举勇
中国科学技术大学

Jun 7, 2018

3D Face Modeling - Manual



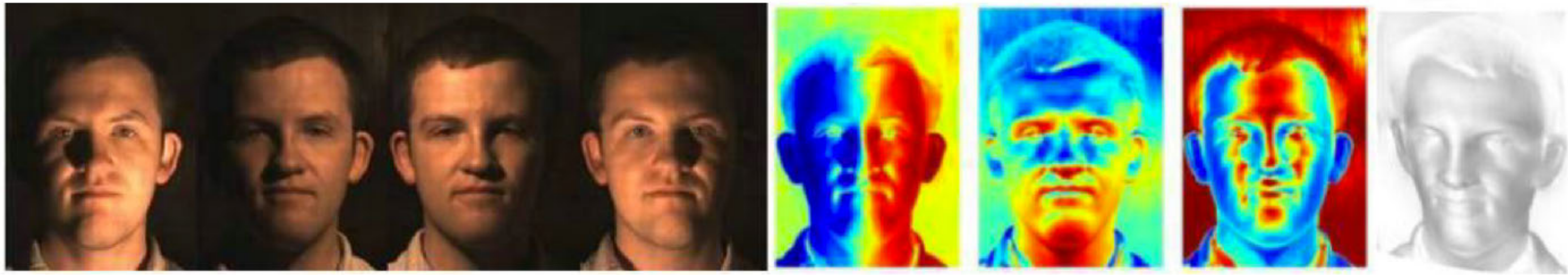
Manual

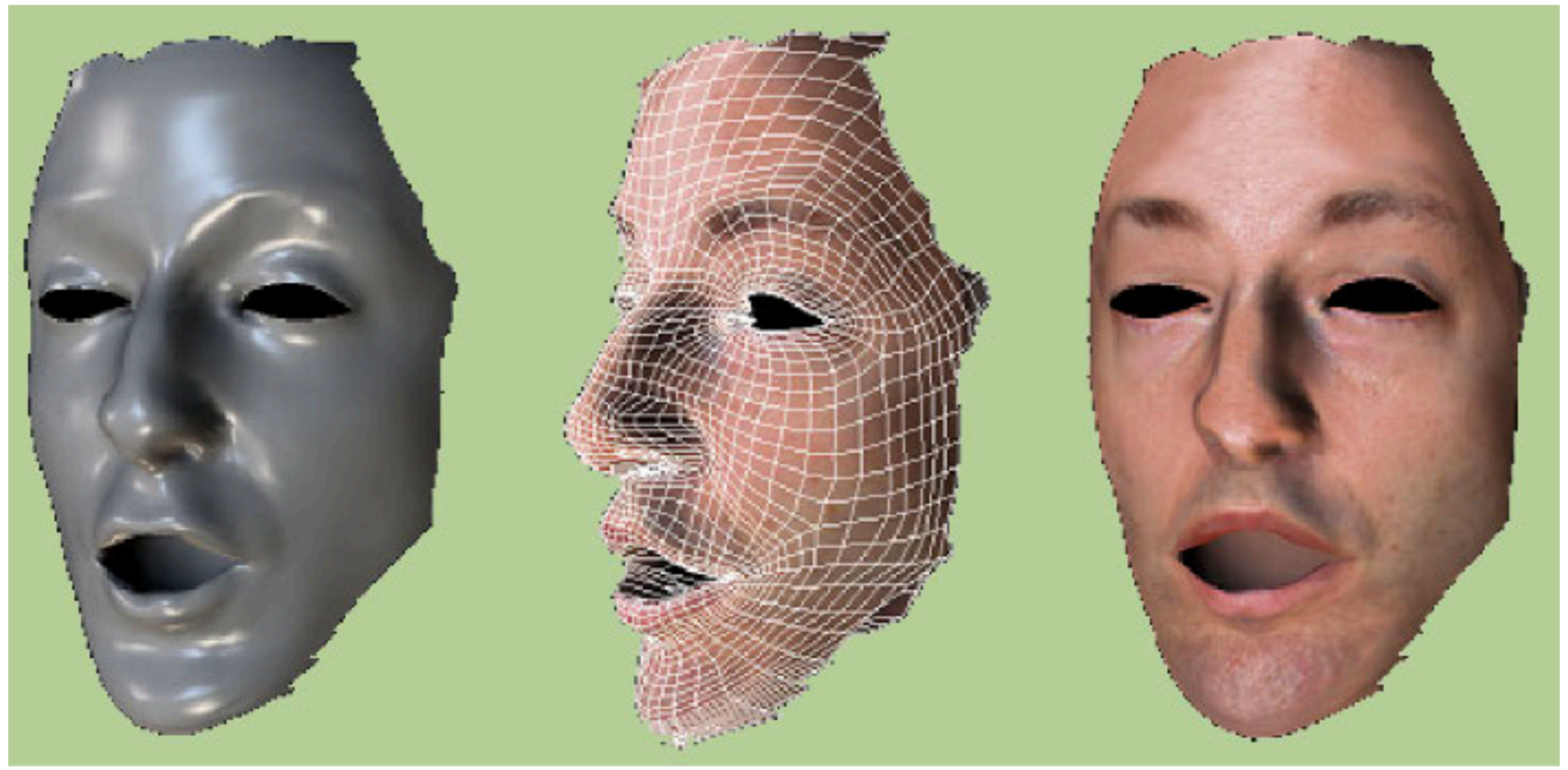
3D Face Modeling - Motion Capture

Motion Capture

3D Scanners

- Structured light, multi-view reconstruction, Laser Scanning, etc
- Most 3D sensors is quite large and expensive, thus hard to be widely used





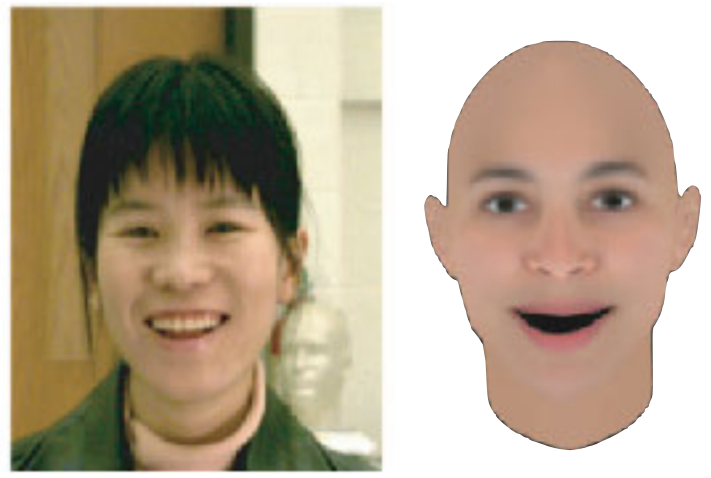
Related Work

Markers

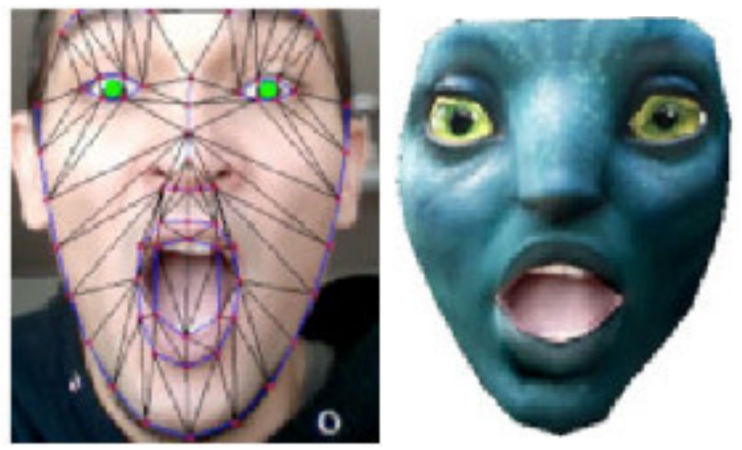
[Ma et al, 2008] [Bickel et al, 2008] [Huang et al, 2011]

Webcam

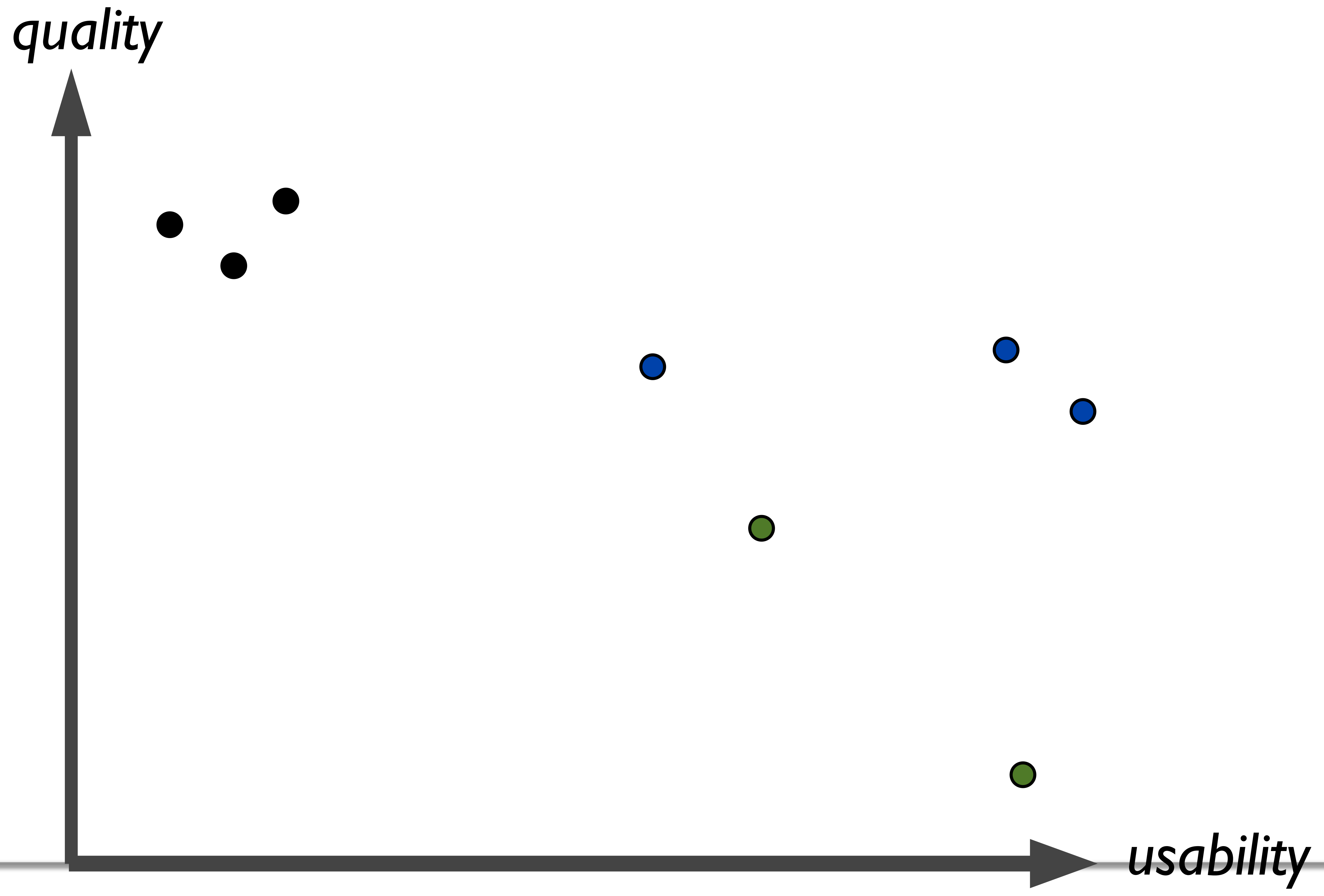
[Chen et al, 2015]



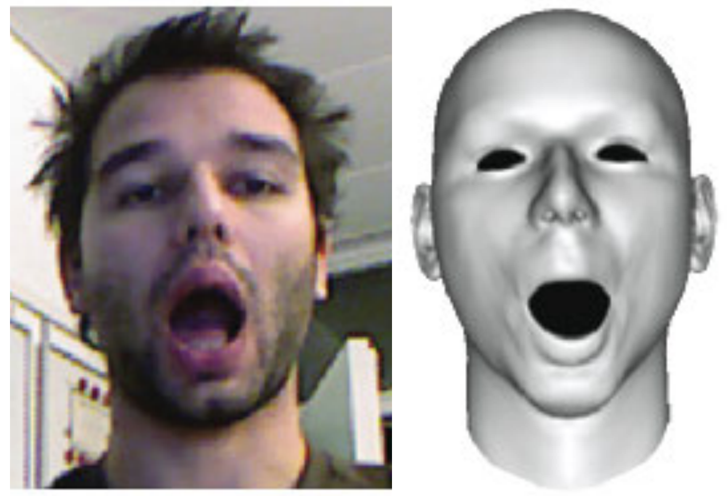
[Chai et al, 2003]



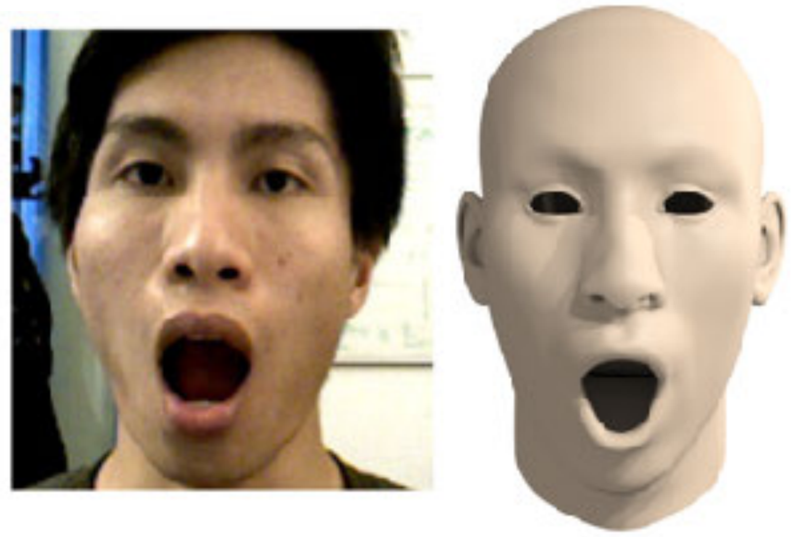
[Saragih et al, 2011]



RGB-D



[Weise et al, 2011]



[Bouaziz et al, 2013]

3D Face Applications

- Expression t
- Face recog

Reconstruction Accuracy

- Sparse landmark
- Color consistency
- Geometry consistency
- Prior knowledge/statistical models: 3DMM, FaceWareHouse, etc…

Computation Speed

- Fast numerical optimization
- Multi-threaded optimization
- GPU computing
- Learning based: offline training, testing in real-time

Usability

- Equipment
  - Laser scanner, motion capture
  - RGB-D camera
  - RGB camera
- User-specific calibration or Manual assistance

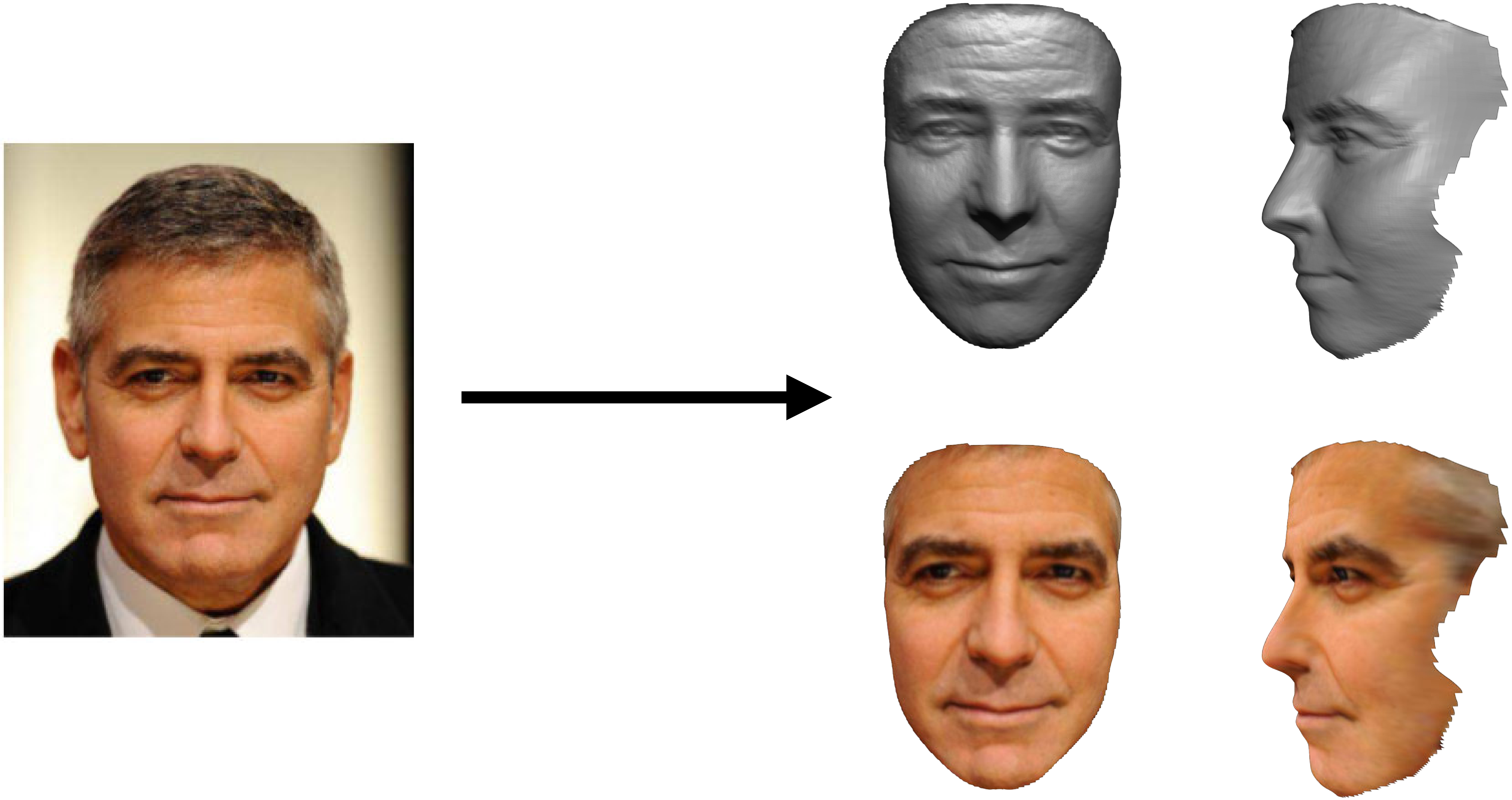
Outline of This Talk

- Optimization based 3D Face Reconstruction from a Single Image
- CNN based 3D Dense Face Tracking from Monocular Camera
- Monocular RGB Camera to monocular RGB-D Camera
- Normal 3D face to Caricature face

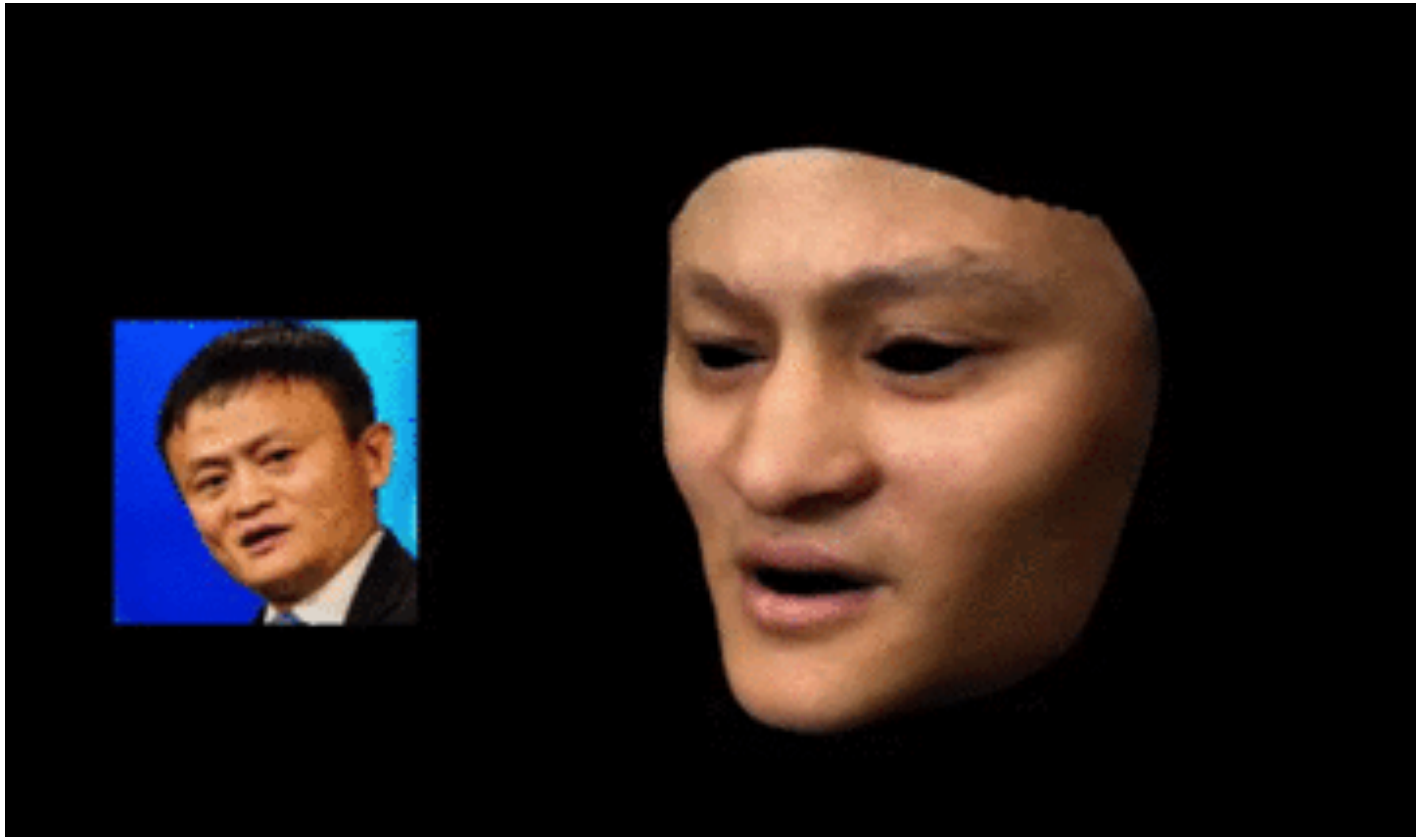
Single Image based Face Reconstruction

3D Face Reconstruction with Geometry Details from a Single Image
IEEE Transactions on Image Processing, 2018.

Reconstruction From Image



Inverse Process



Preliminaries - Rendering Equation

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

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

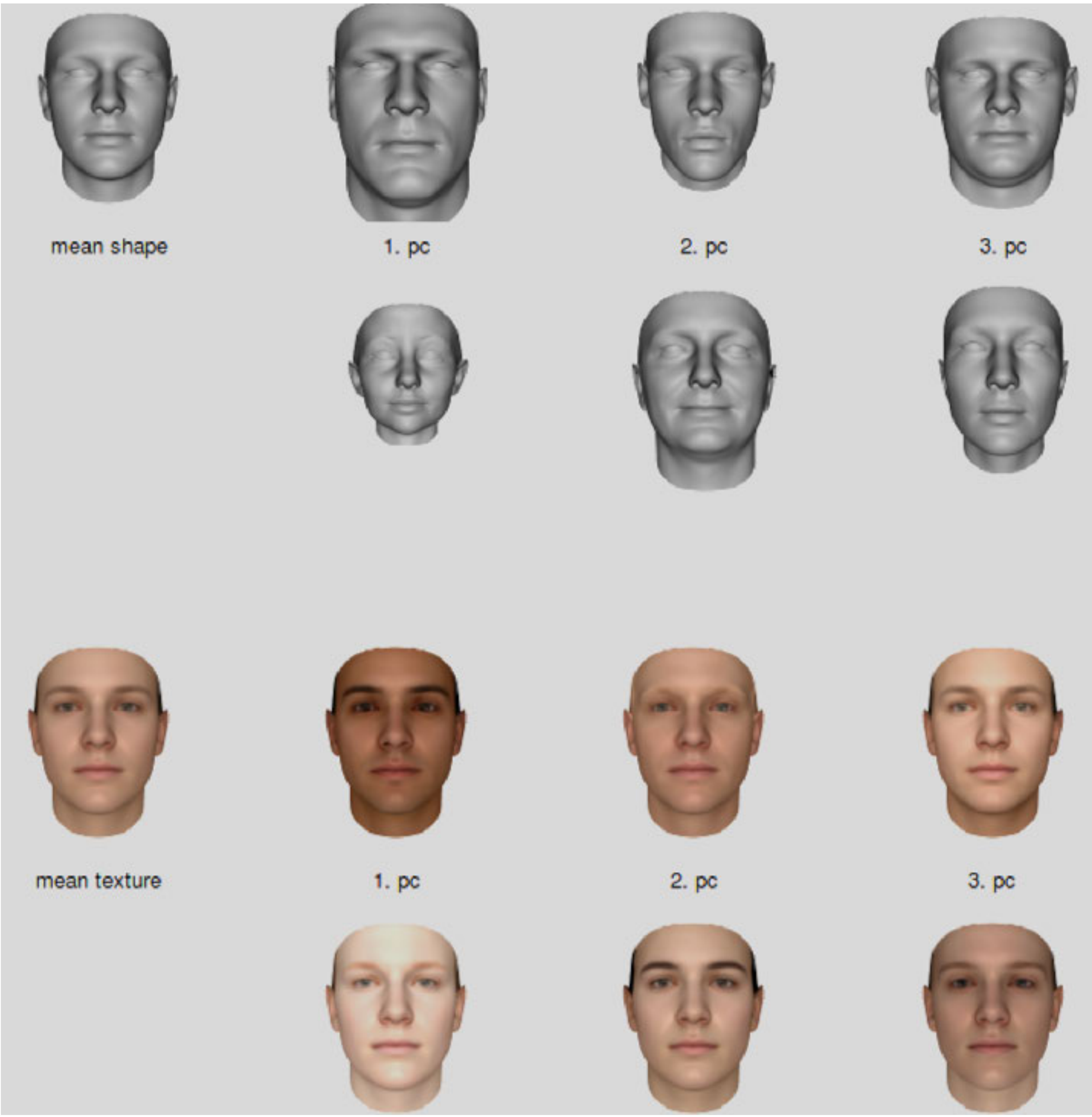
Image Lighting parameter Geometry Albedo

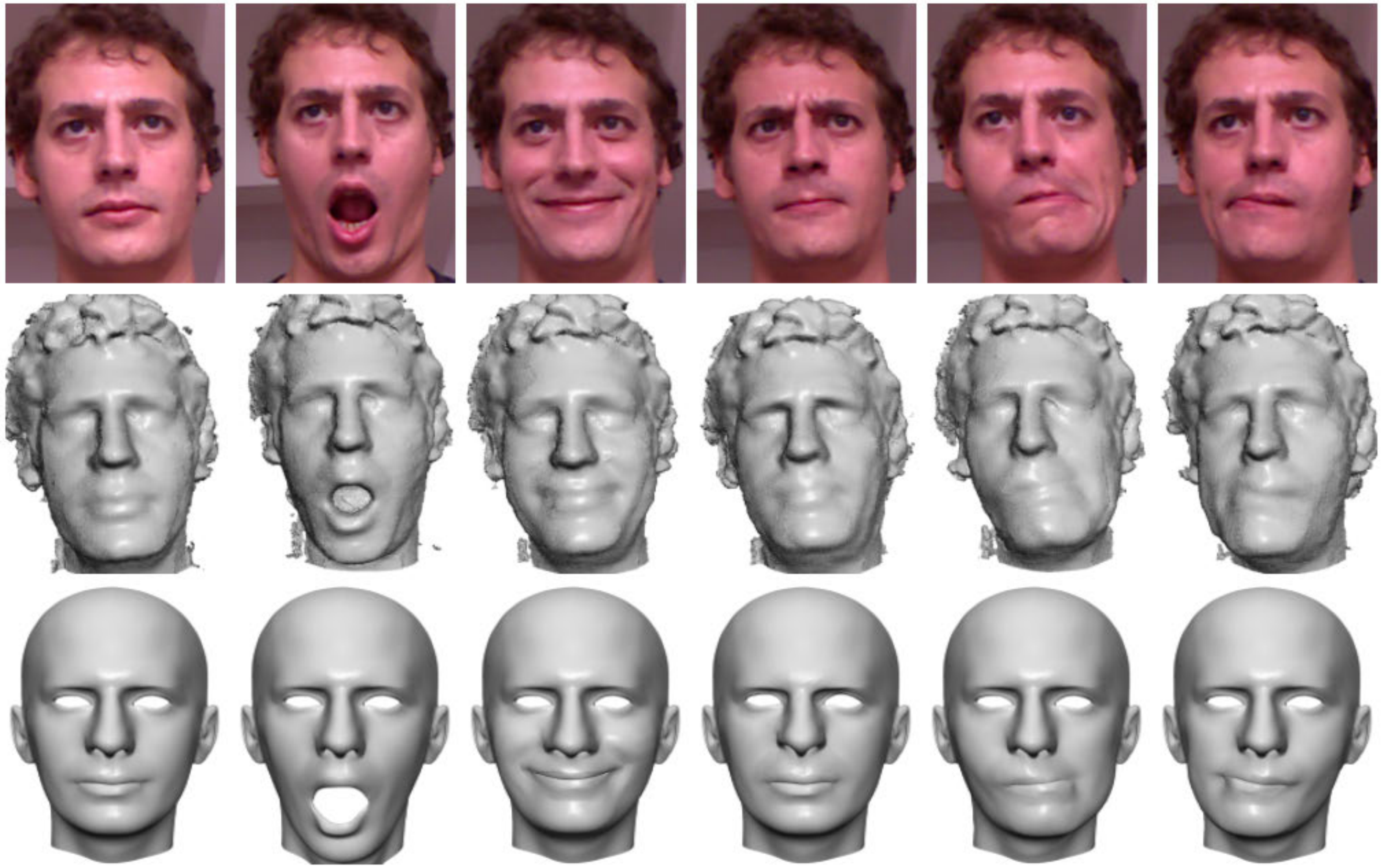
Preliminaries - 3D Face Representation

Mean Face Identity Expression Displacement

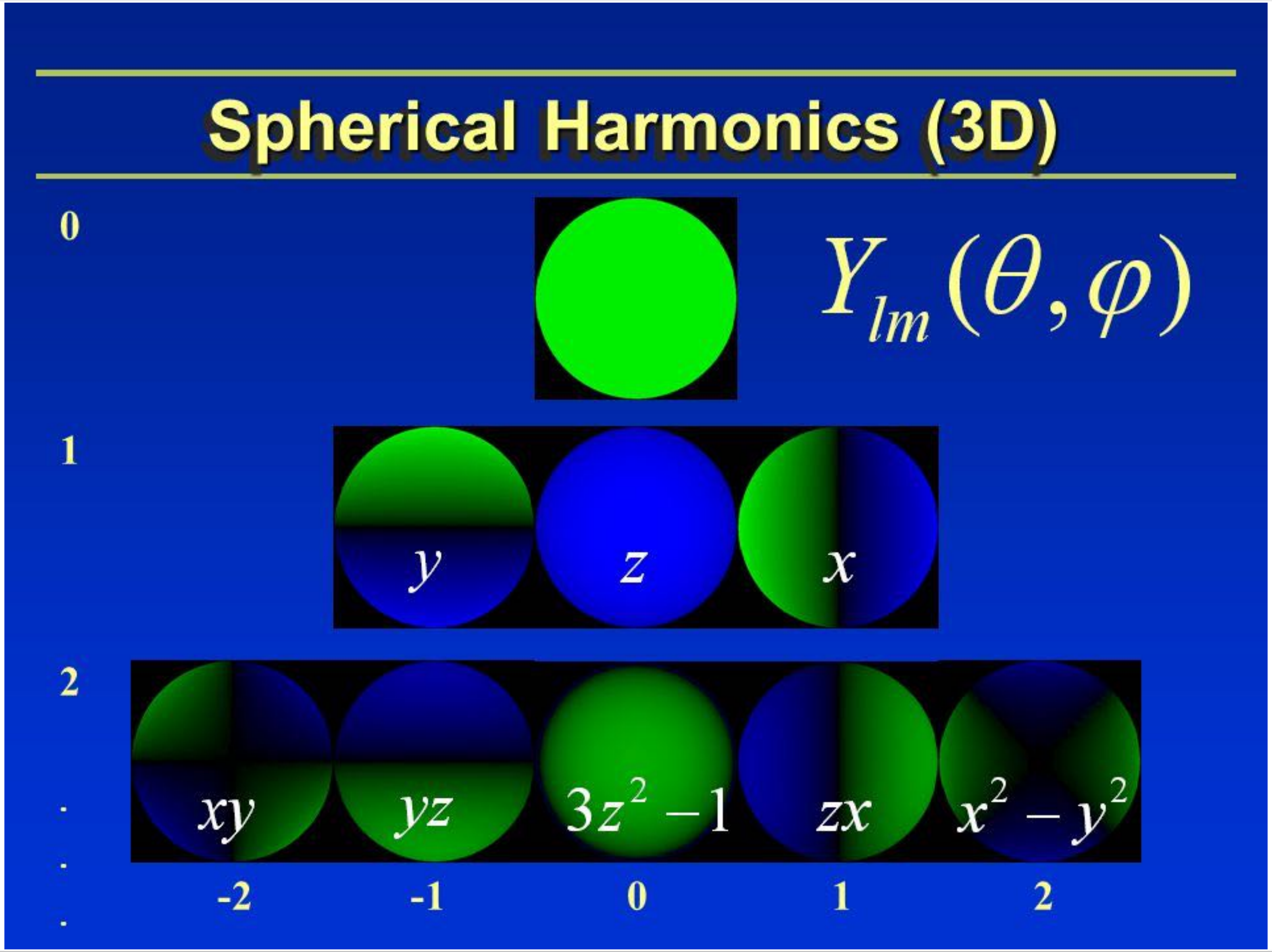
$$F = (\bar{F} + A_{\mathrm{id}}\alpha_{\mathrm{id}} + A_{\mathrm{exp}}\alpha_{\mathrm{exp}}) + F_{\mathrm{disp}}$$

Preliminaries - 3DMM & FaceWarehouse



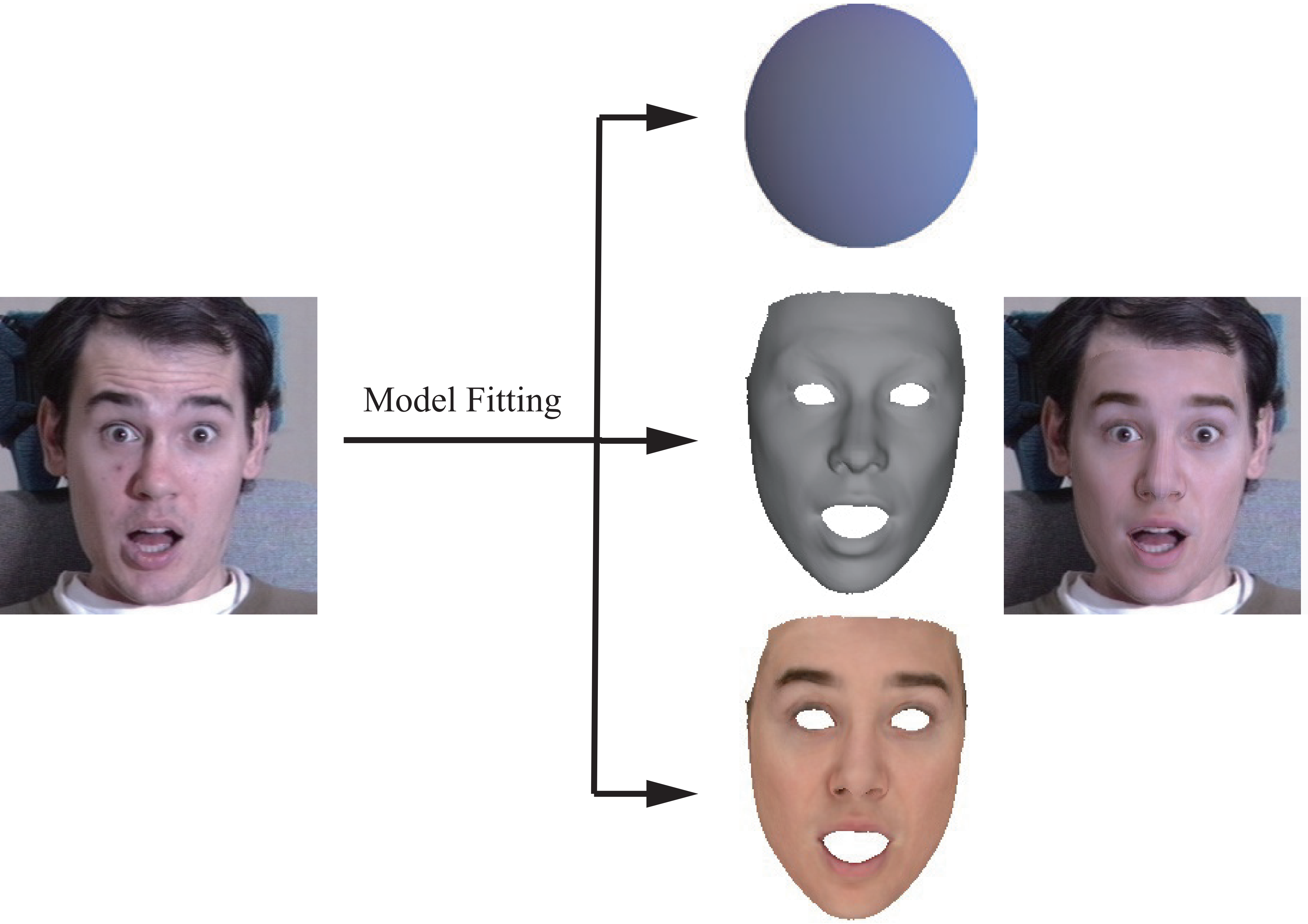


Preliminaries - Lighting



Inverse Rendering - Coarse

Inverse Rendering - Coarse



Inverse Rendering - Coarse

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

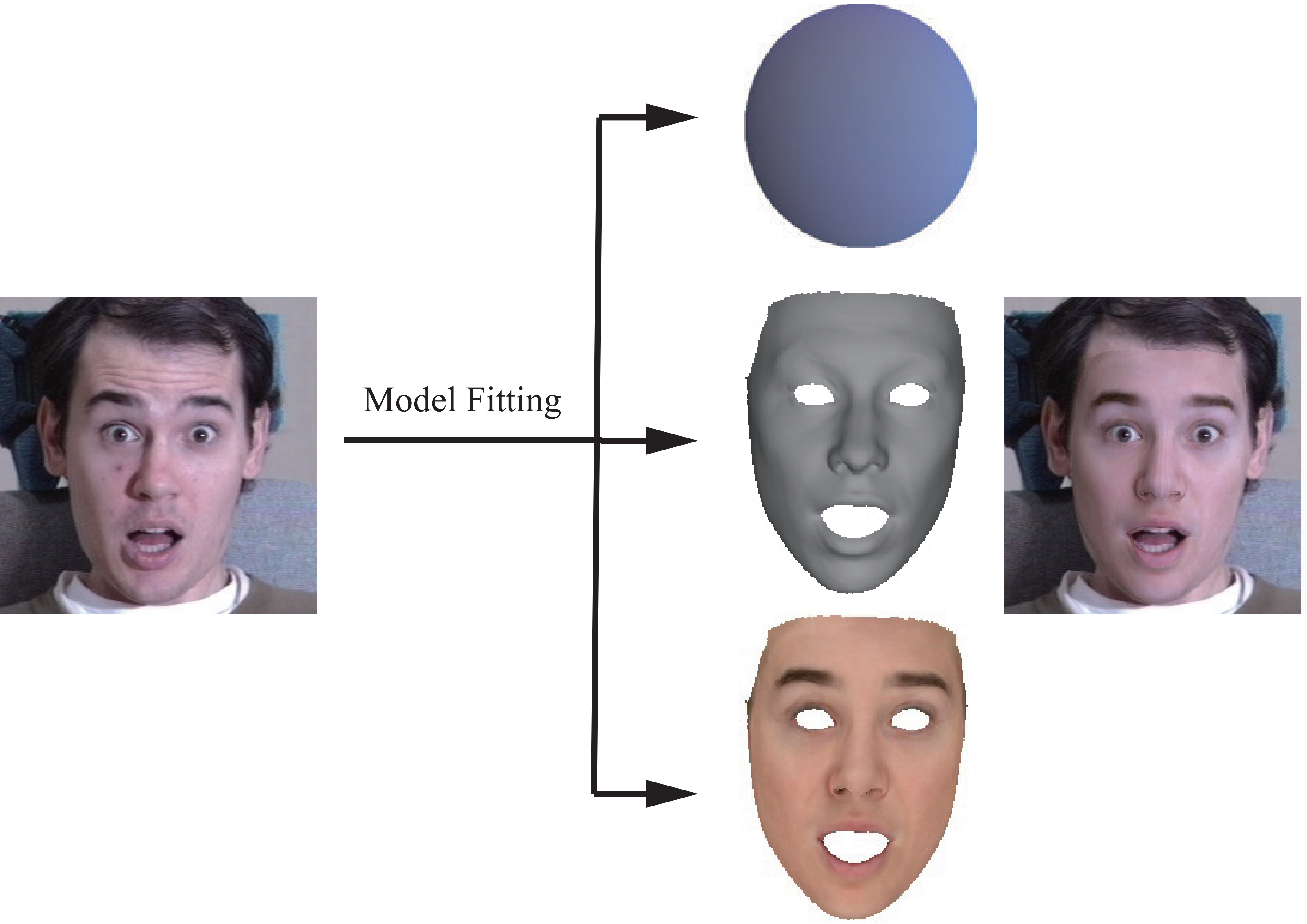
$$E(\chi) = E_{\text{con}} + w_{\text{lan}} E_{\text{lan}} + w_{\text{reg}} E_{\text{reg}}$$

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

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

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

Inverse Rendering - Geometry



Inverse Rendering - Geometry

Inverse Rendering - Geometry

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

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

Inverse Rendering - Albedo

Inverse Rendering - Albedo

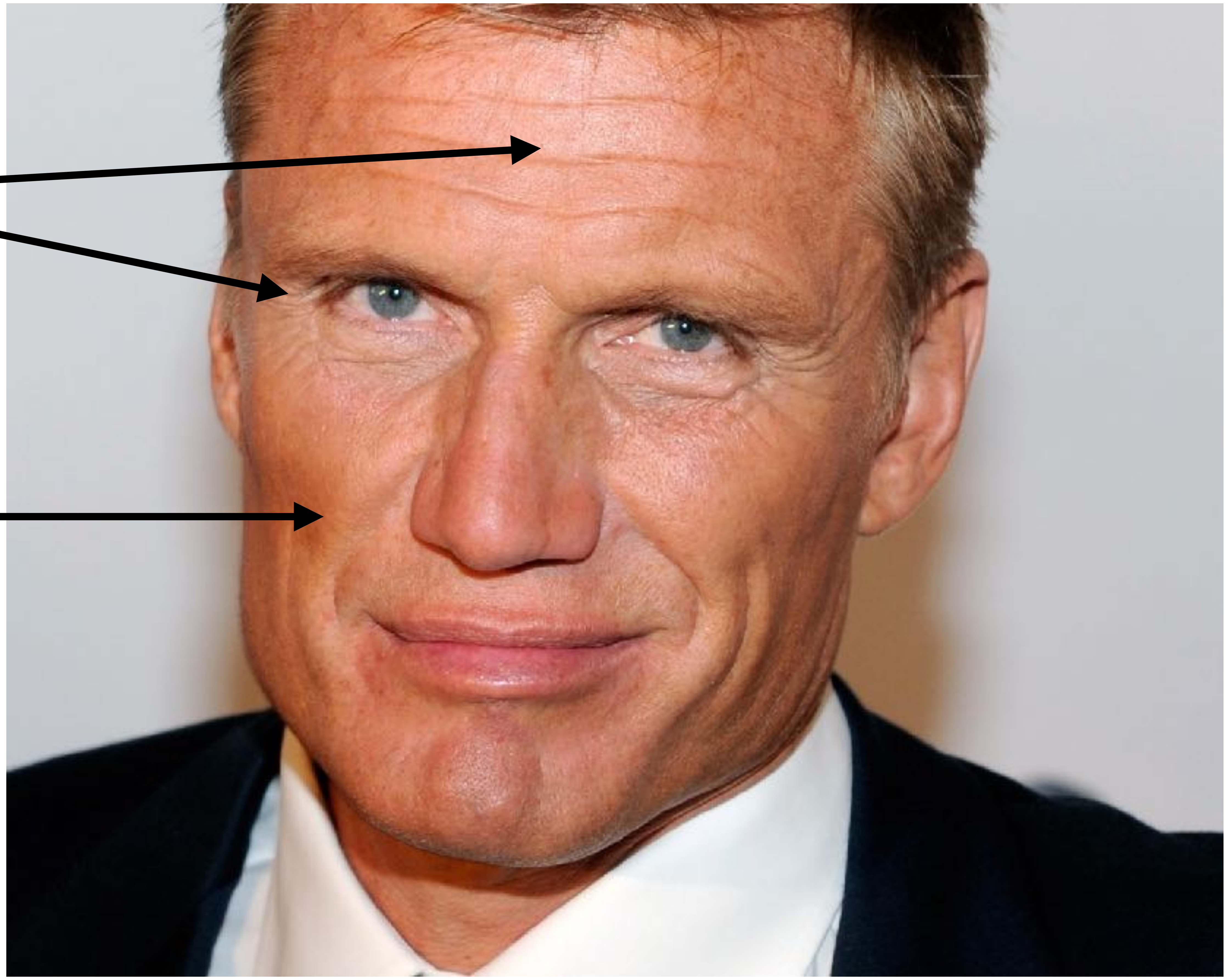
Inverse Rendering - Albedo

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

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

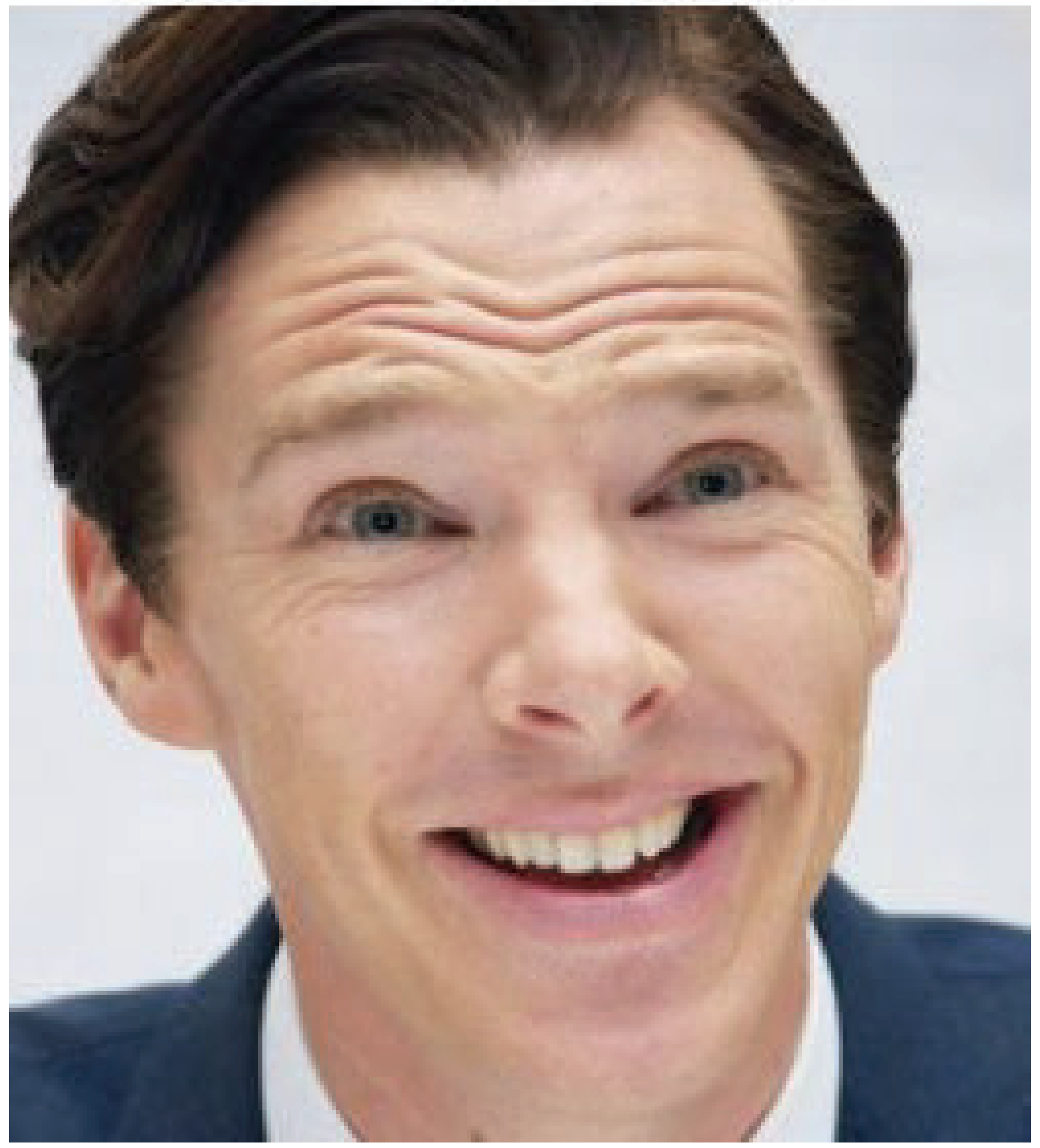
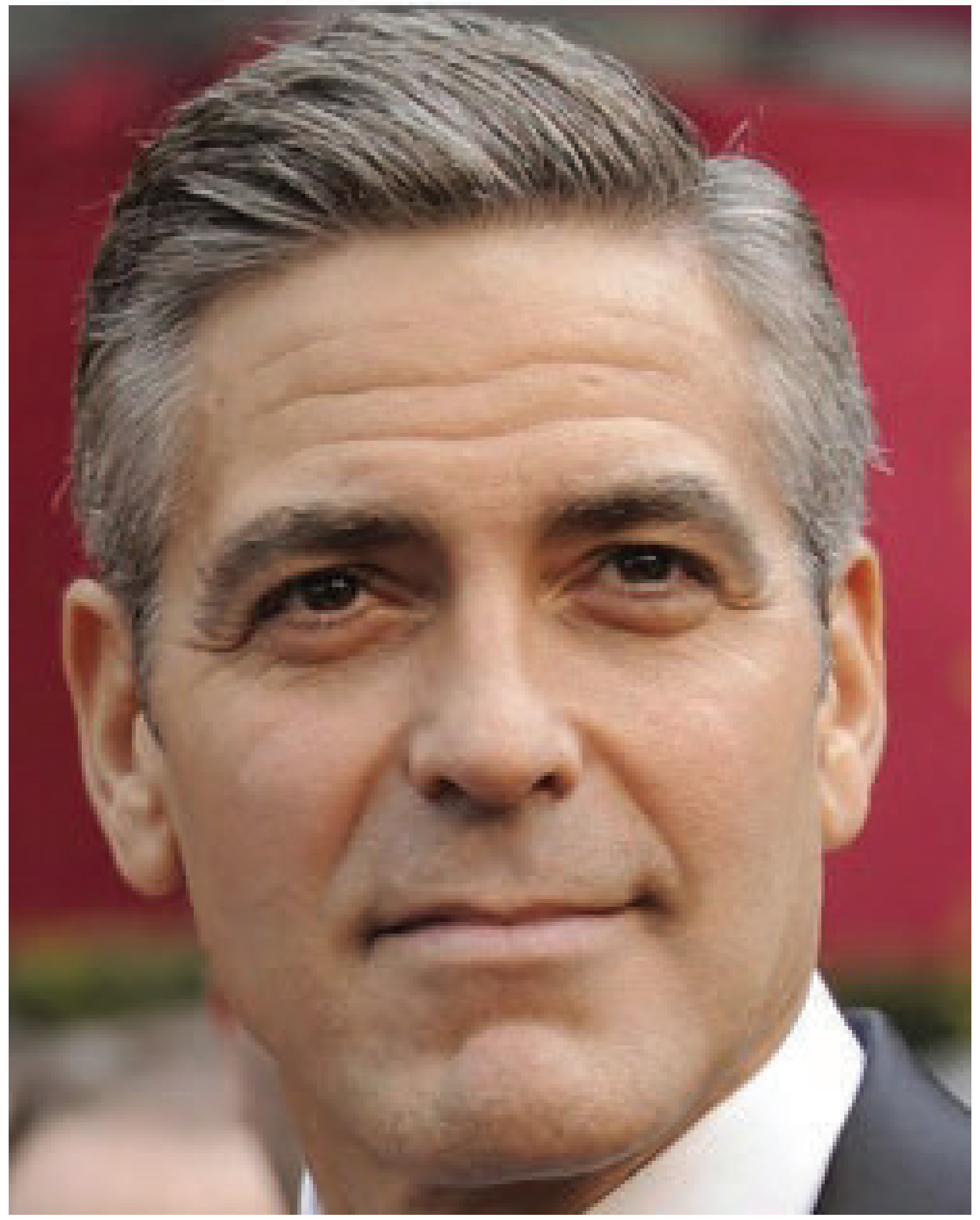
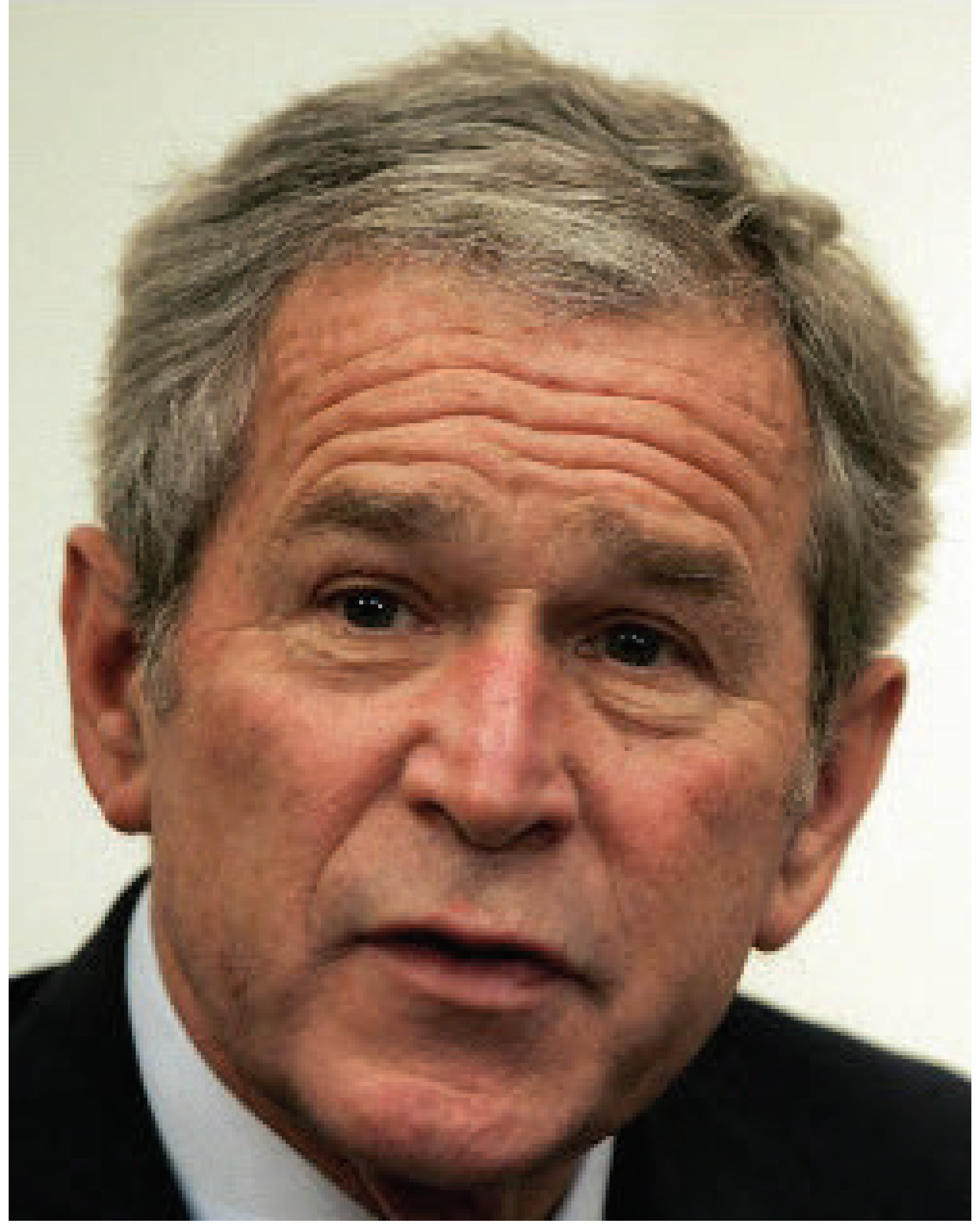
$\beta = 0.65$

$\beta = 0.35$



Recap - Inverse Rendering Process

Input Images

Coarse Results

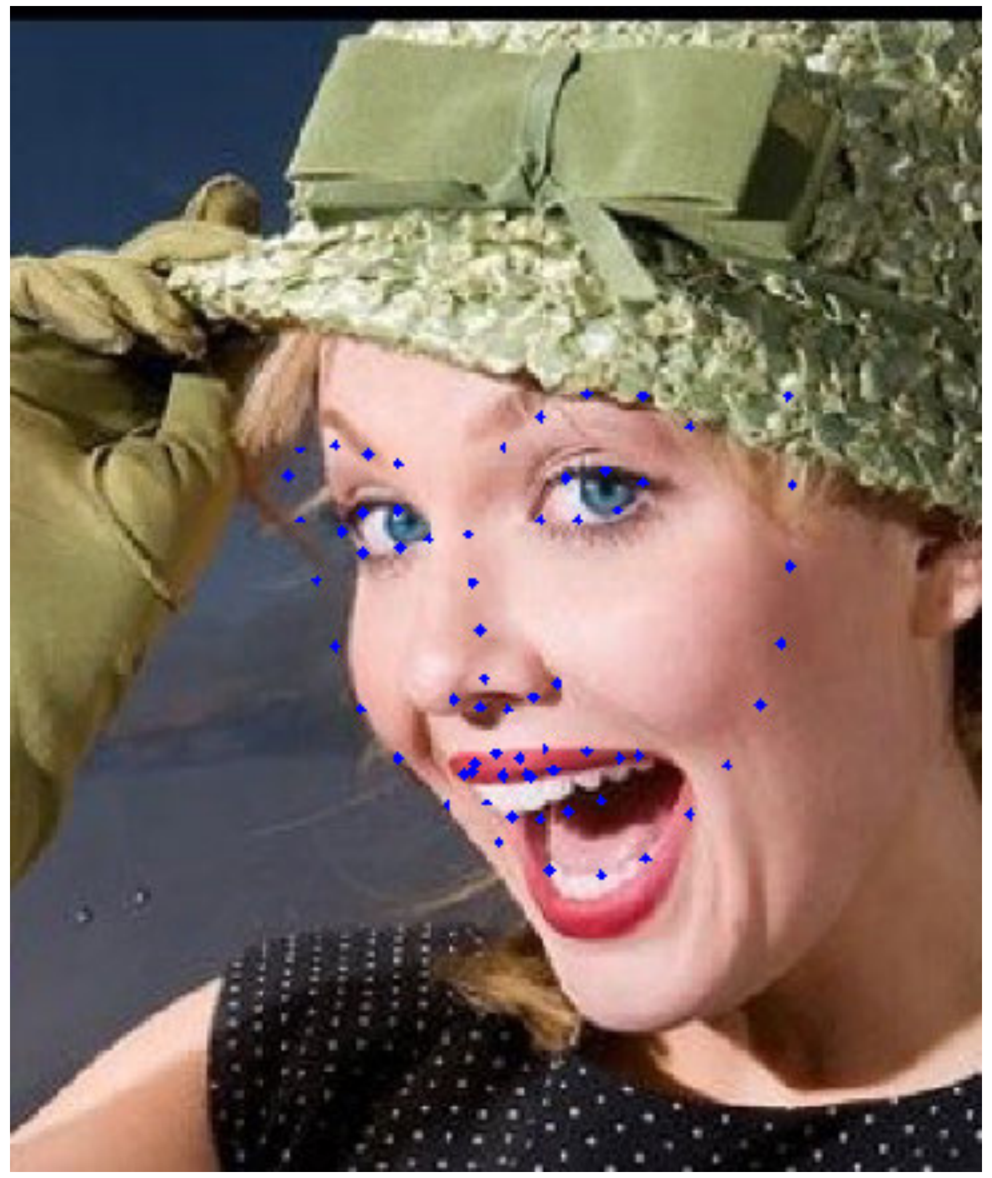



Fine Results

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

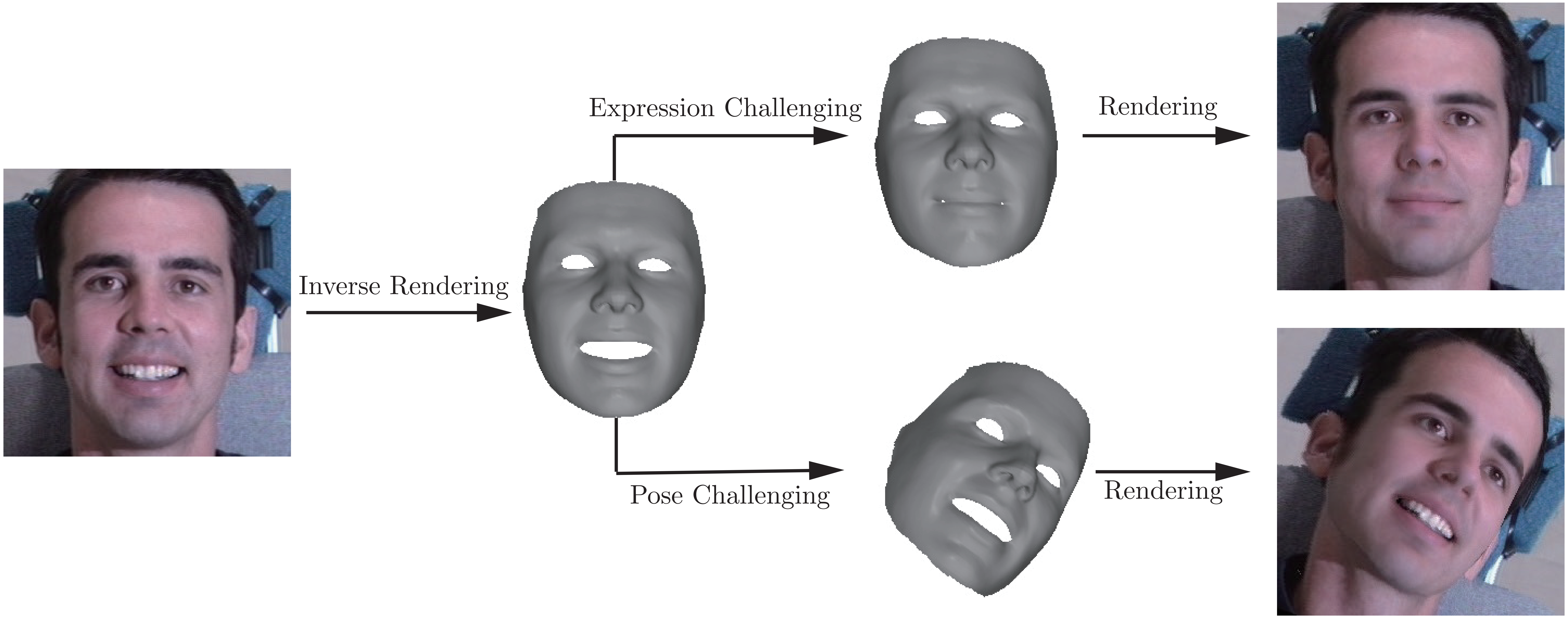
CNN-based Real-time Dense Face Reconstruction with
Inverse-rendered Photo-realistic Face Images
IEEE Trans on PAMI, 2018

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

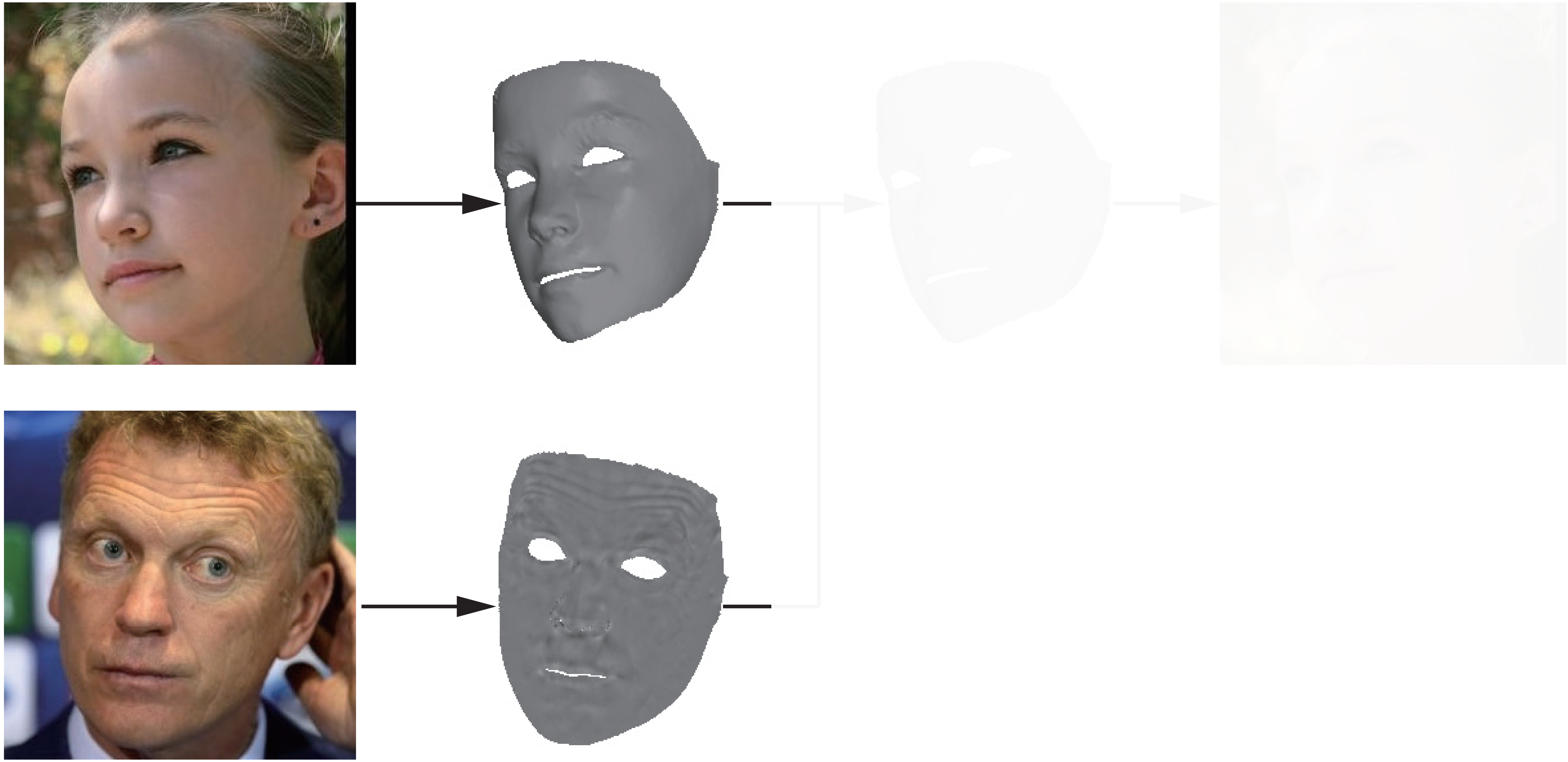
Data Augmentation - Coarse



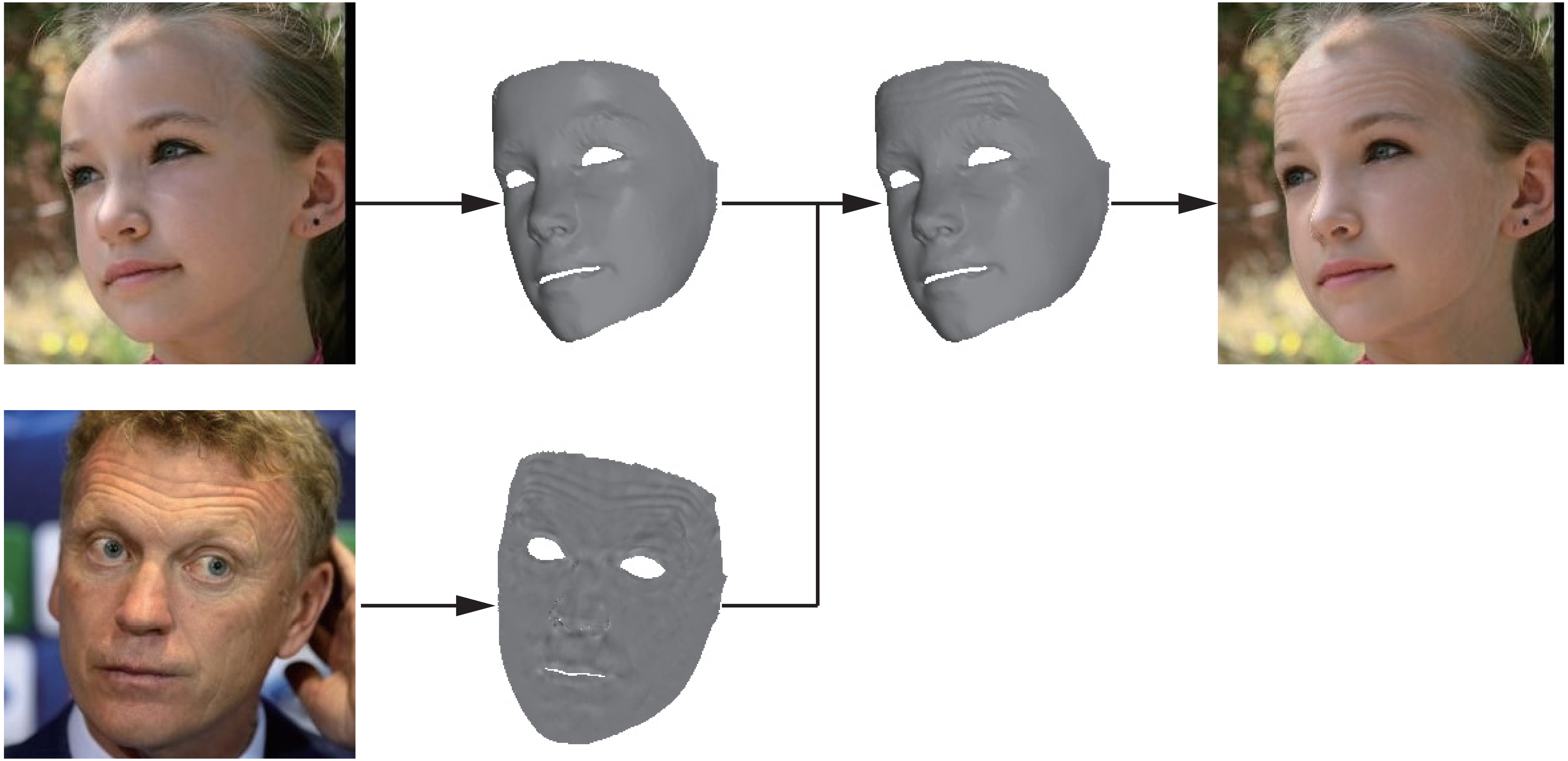
More Augmentation Examples

Data Augmentation - Fine

Data Augmentation - Fine



Data Augmentation - Fine



CoarseNet

- ResNet 18

layer name	output size	18-layer
conv1	112×112	
conv2_x	56×56	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$

- Variable: 3DMM parameter, pose parameter
- Loss: pixels' distance

$$Proj(\mathcal{P}) = \Pi R(\bar{S}_p + A_{p,id} \cdot \alpha_{\mathrm{id}} + A_{p,\mathrm{exp}} \cdot \alpha_{\mathrm{exp}}) + t$$

$$\mathcal{L}_{\mathrm{pose}} = \|Proj(\mathcal{P}_g) - Proj(\mathcal{P}_{n,\mathrm{pose}}, \mathcal{P}_{g,\mathrm{geo}})\|_2^2$$

$$\mathcal{L}_{\mathrm{geo}} = \|Proj(\mathcal{P}_g) - Proj(\mathcal{P}_{n,\mathrm{geo}}, \mathcal{P}_{g,\mathrm{pose}})\|_2^2$$

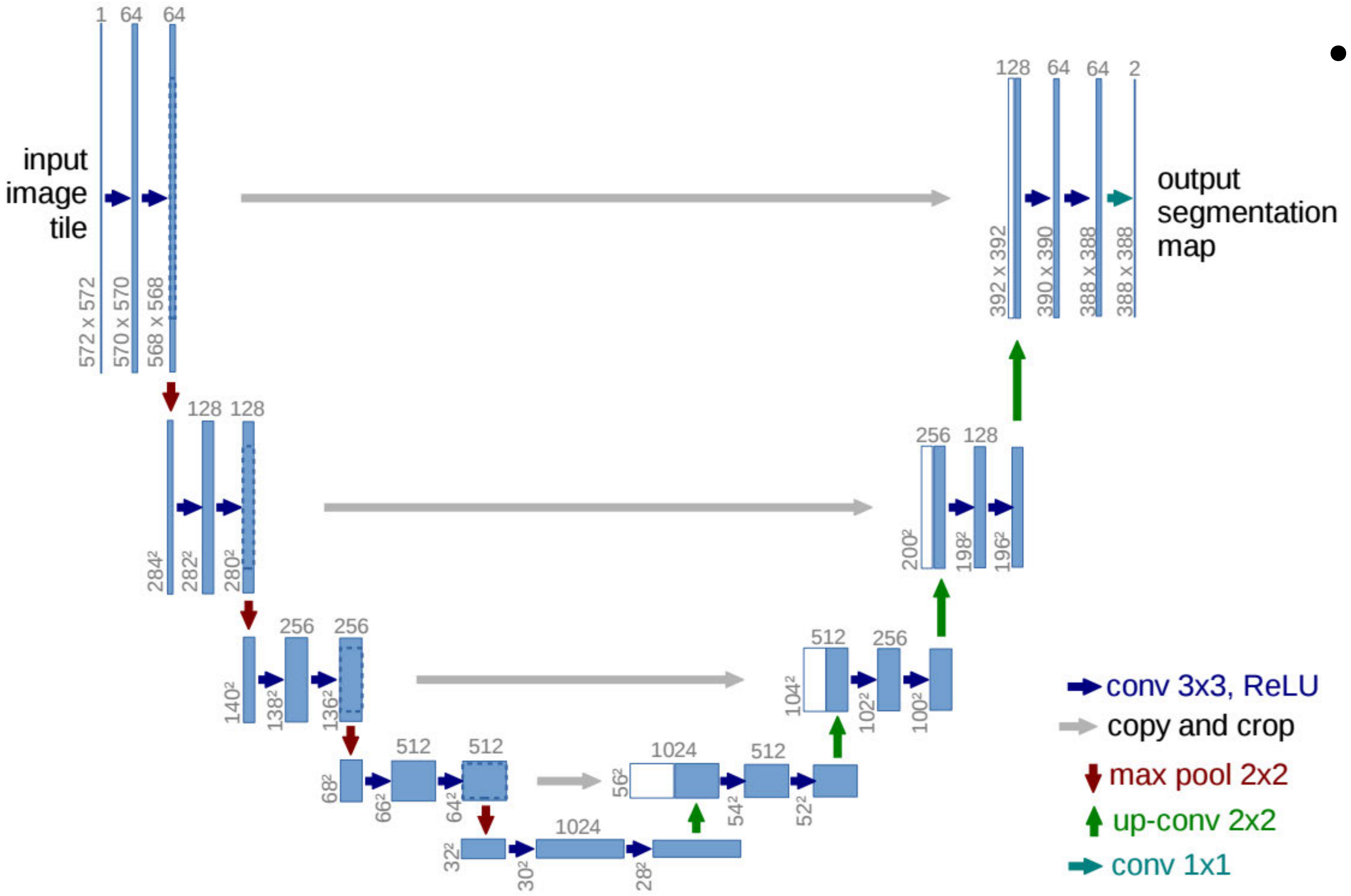
$$\mathcal{L} = w \cdot \mathcal{L}_{\mathrm{pose}} + (1 - w) \cdot \mathcal{L}_{\mathrm{geo}}$$

Comparison between Euclidean Loss

	Pixel Distance(Pose)	Pixel Distance(Geometry)
Parameter L2 loss (only learn pose)	29.35	
Parameter L2 loss (only geometry)		5.43
Proposed Loss	**7.69**	**4.07**

FineNet

- U-Net



- Variable: depth displacement
- Loss: euclidean distance

Comparison: Optimization vs CNN

Optimization

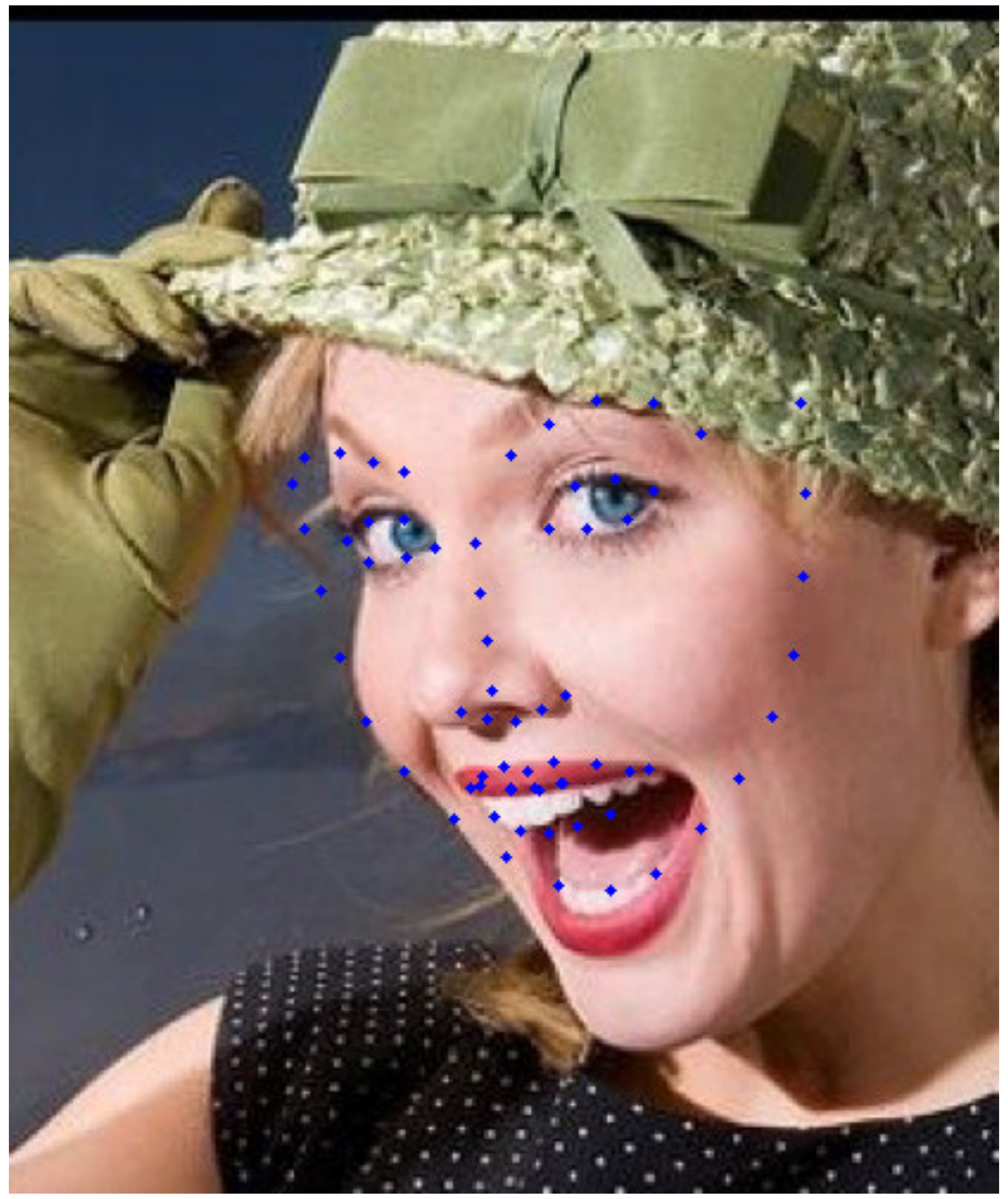
CNN

Comparison: Optimization vs CNN

Optimization

CNN

Comparison: Optimization vs CNN



Landmarks

Optimization

CNN

Reconstruction Results

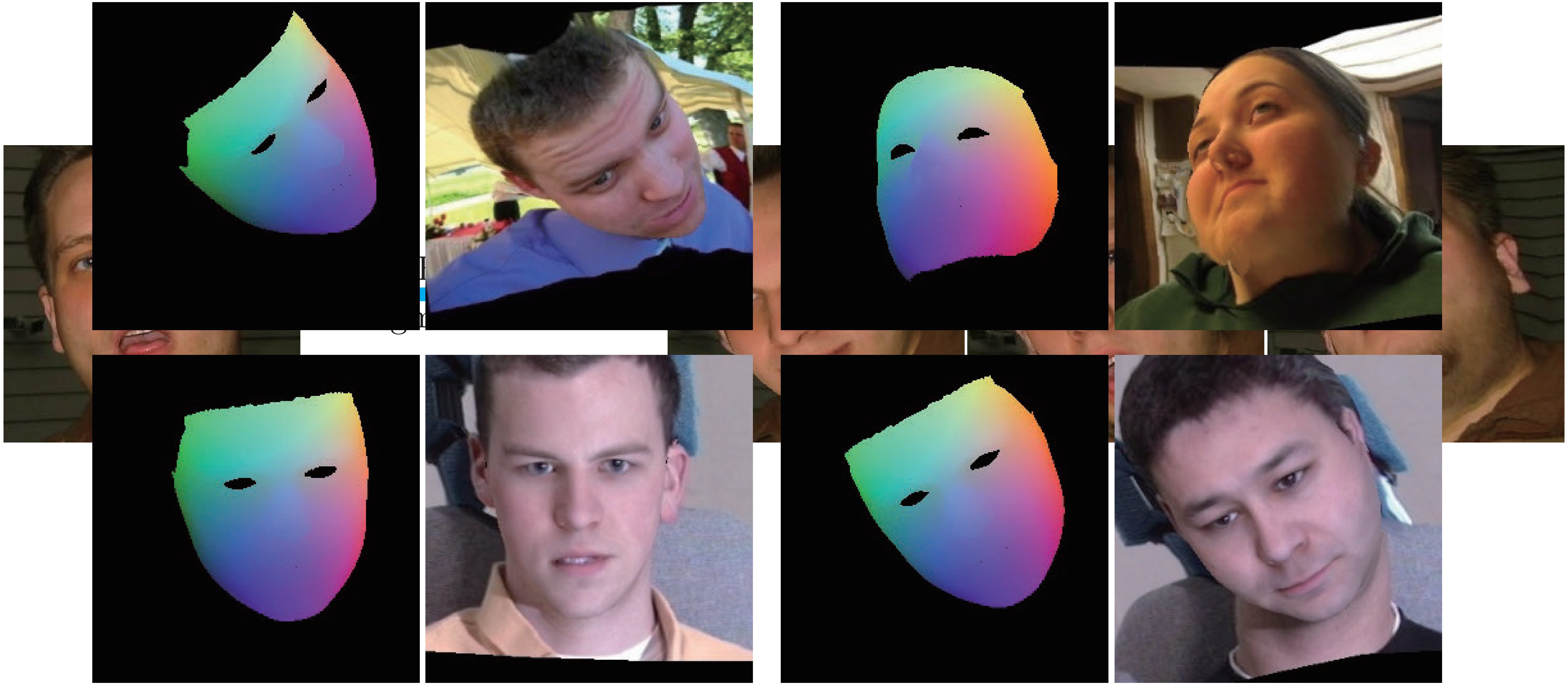
Image $\longrightarrow$ Video

Dense Face Tracking From RGB Video

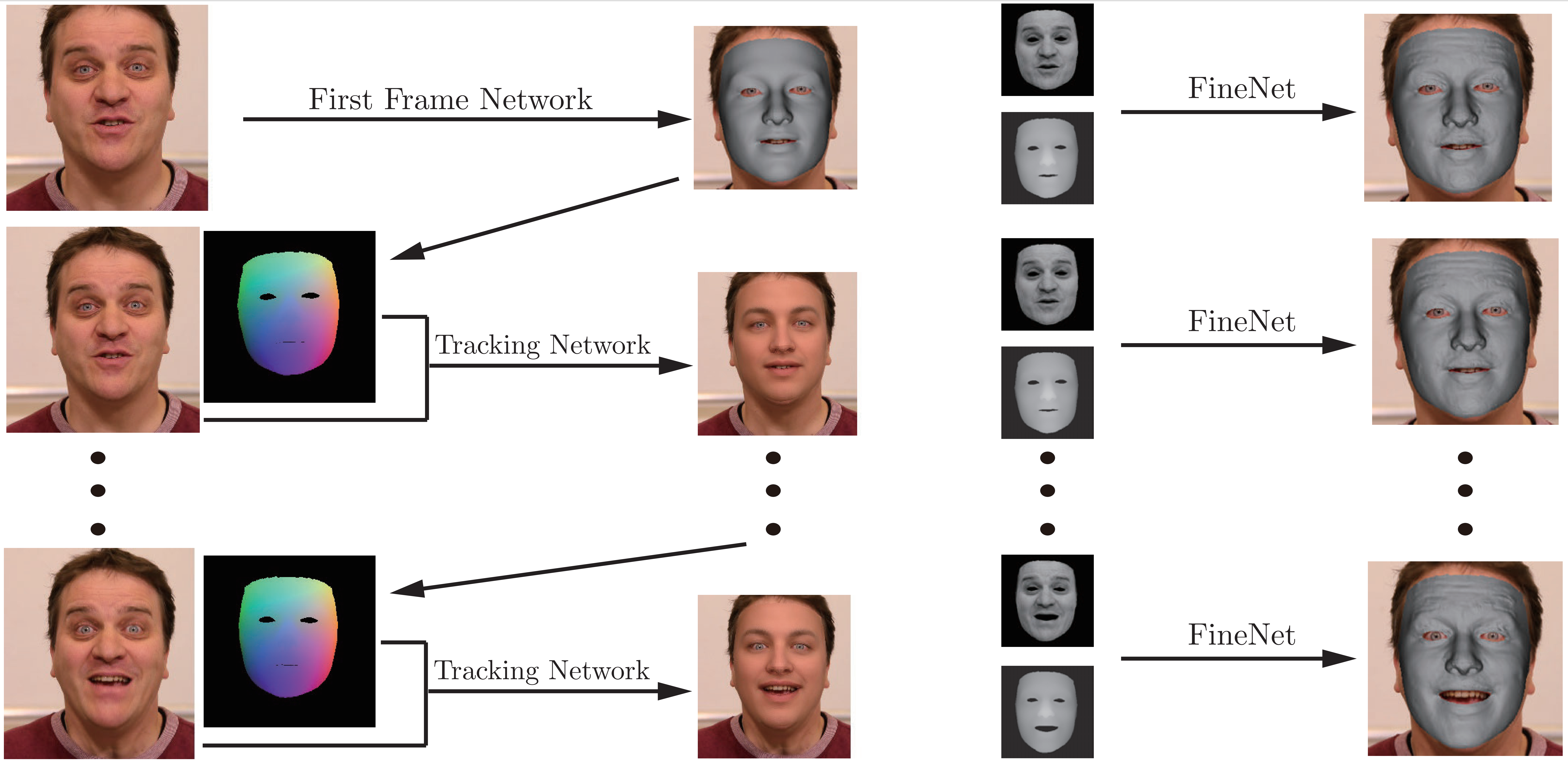
Problem Formulation

- Input: a face video sequences
- Output: detailed 3D face geometry, albedo, lighting
- Main challenges: there doesn’t exist public dataset
- Solution: Construct video type datas from images.

Training Data Construction



Algorithm Pipeline



Network Design

- First Frame Network
  - Output: pose, geometry
  - Structure: Reset-18
- Tracking Network
  - Output: pose difference with last frame, geometry, albedo, lighting
  - Structure: Reset-18
- Fine-Level Network
  - Output: depth displacement for each pixel
  - Structure: U-Net [Ronneberger et al. 2015]

Results - coarse&fine

Input CoarseNet Output FineNet Output

Result - Comparison

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

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

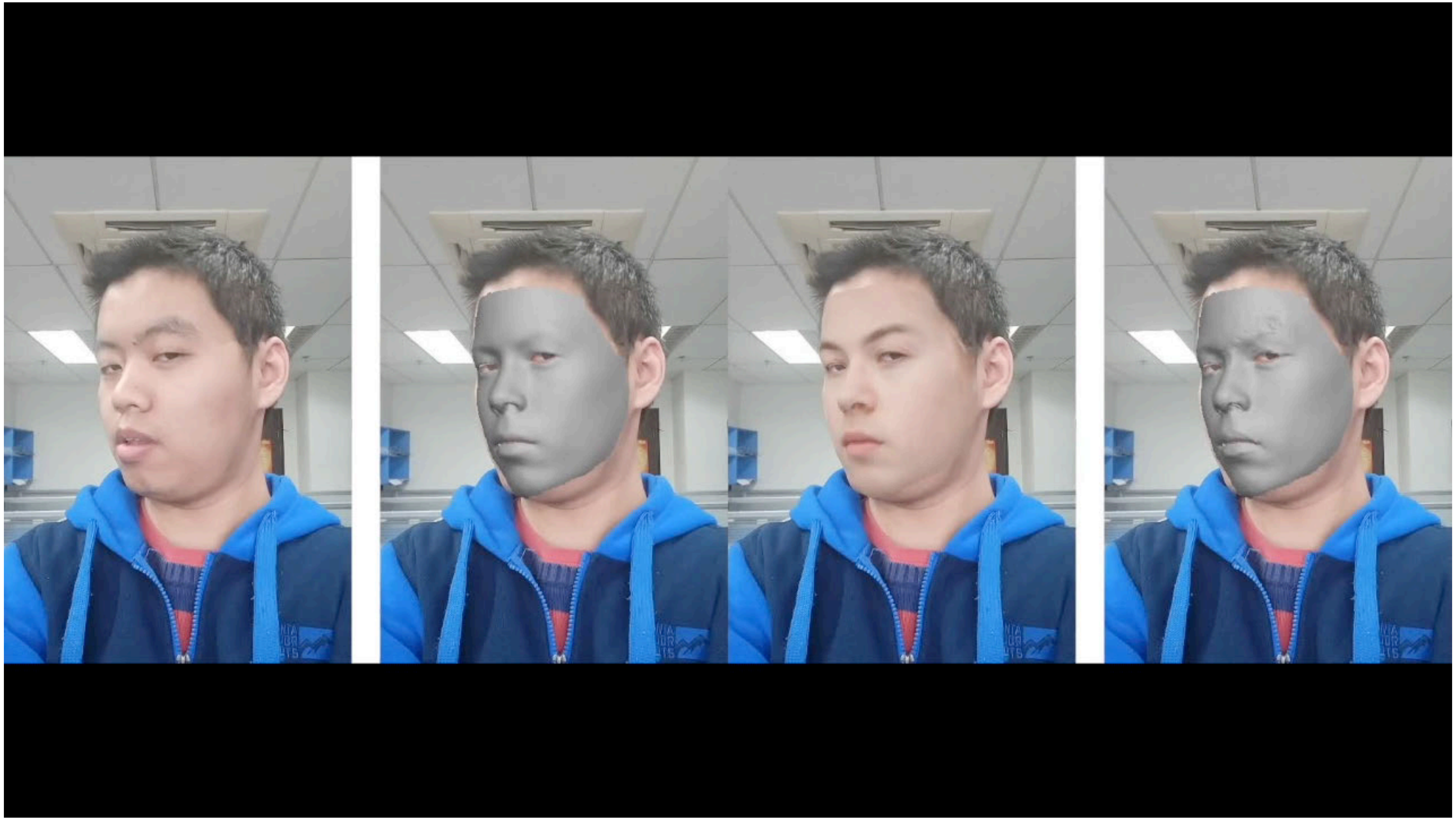
Comparison with GroundTruth

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.

Result - Video Comparison

Real-time facial performance capture



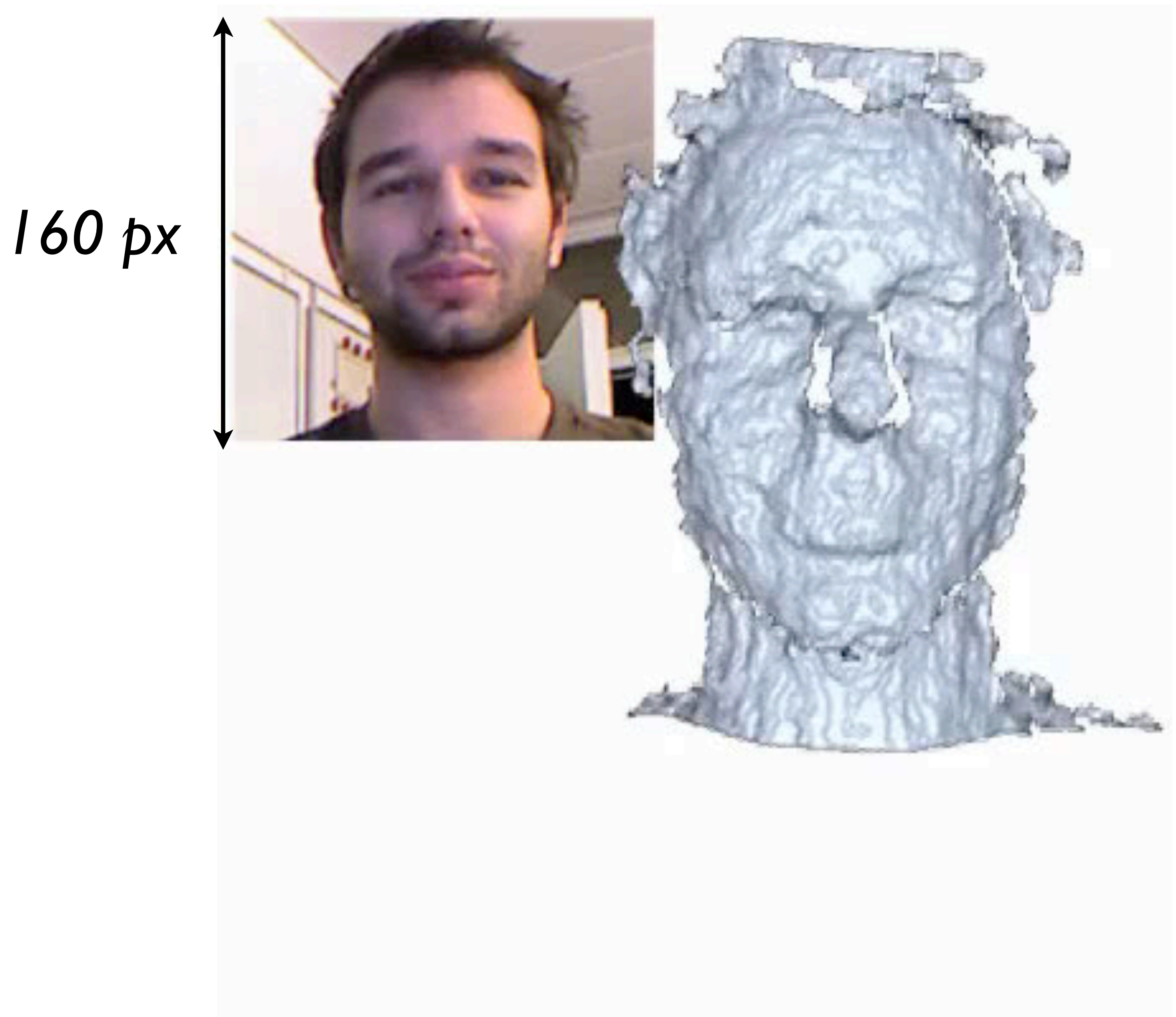
RGB $\longrightarrow$ RGB-D

Depth Sensors

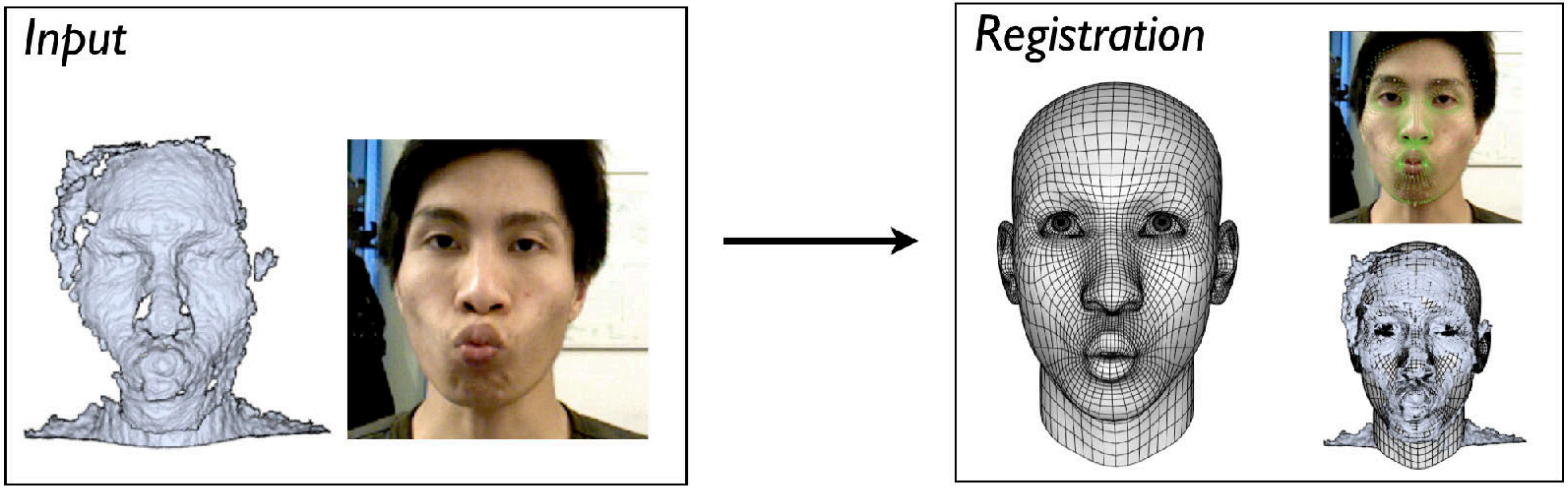
…

\+ Price/Size -

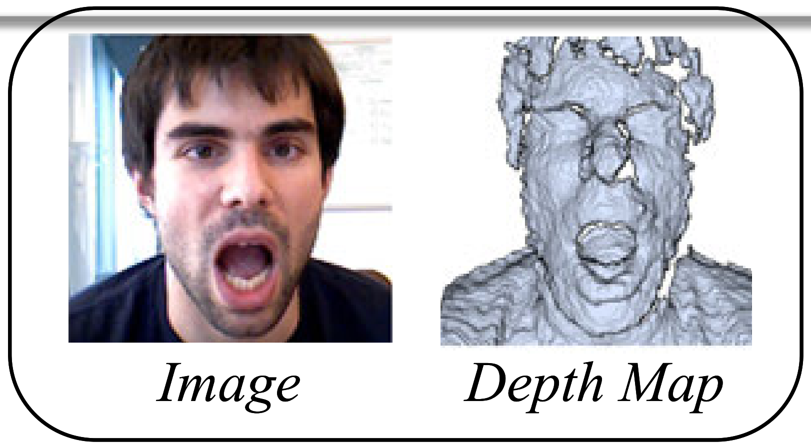
Kinect-Xbox



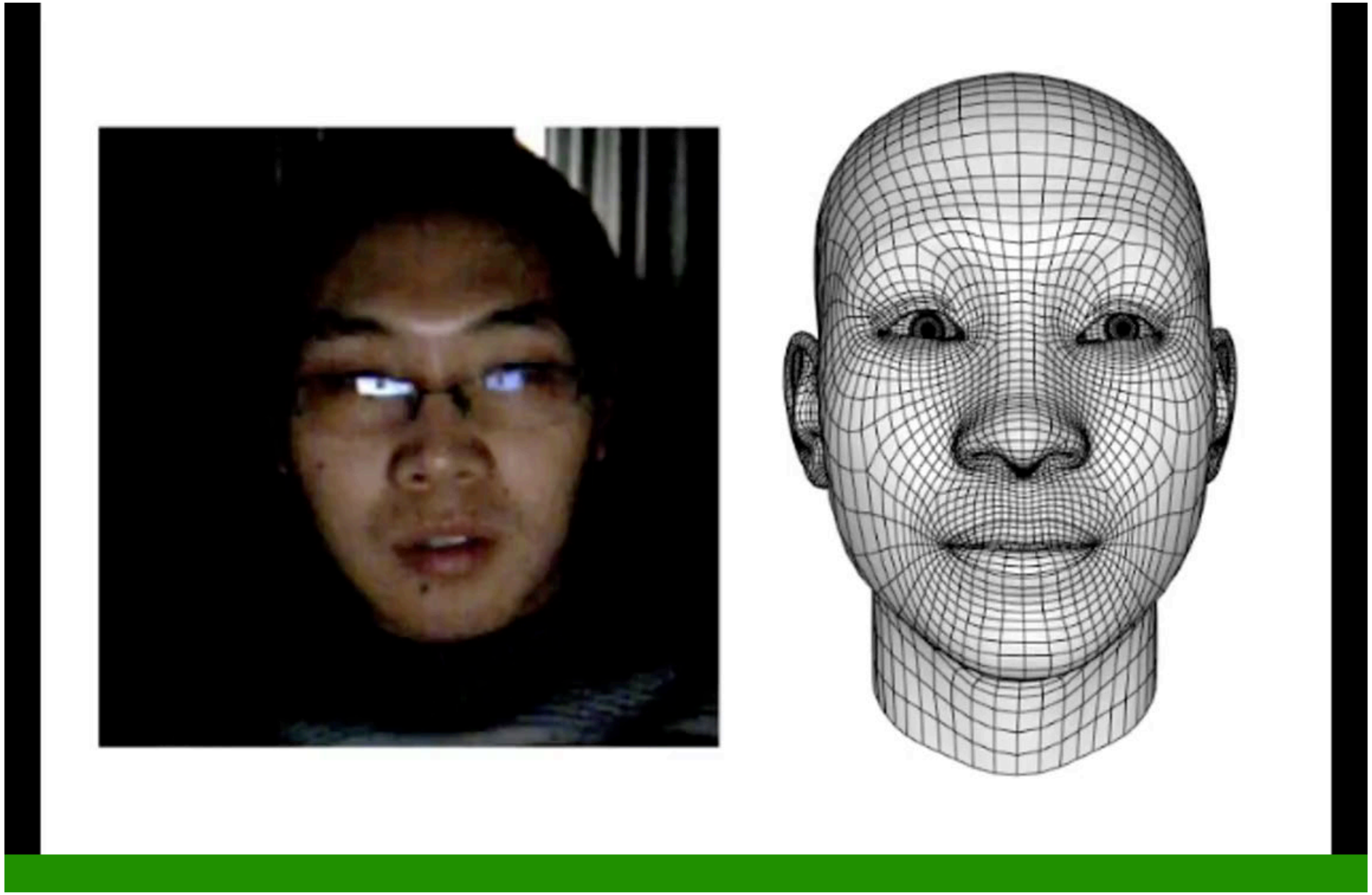
Sofien. B, et.al. Online Modeling For Realtime Facial Animation, Siggraph 2013



Pipeline

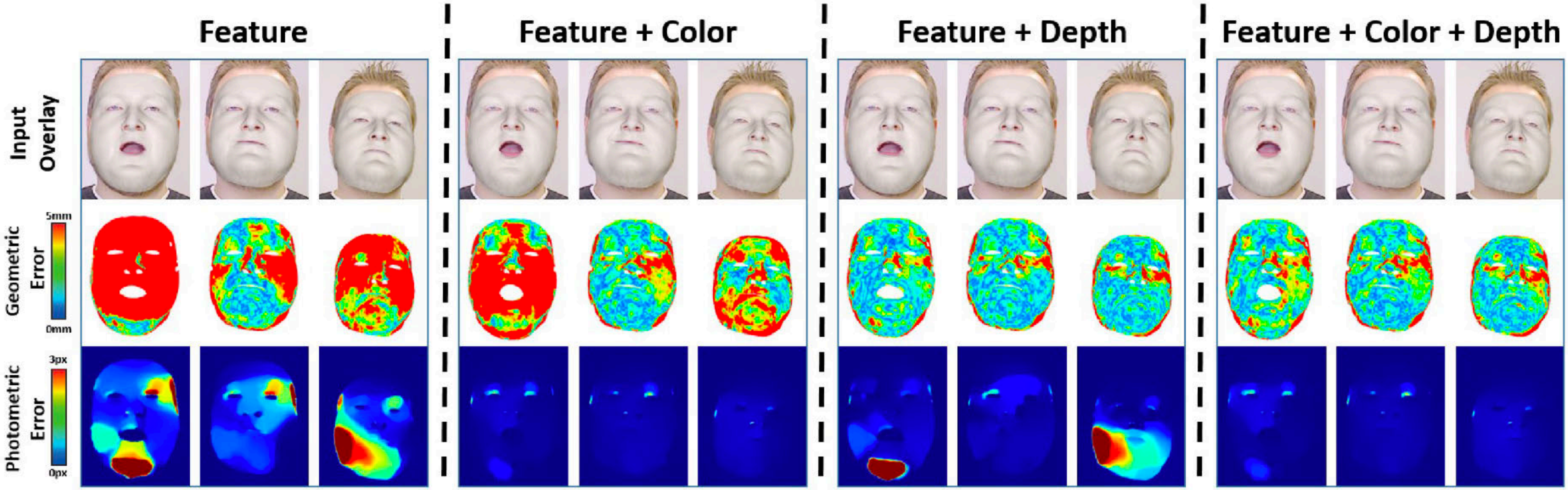


Results



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

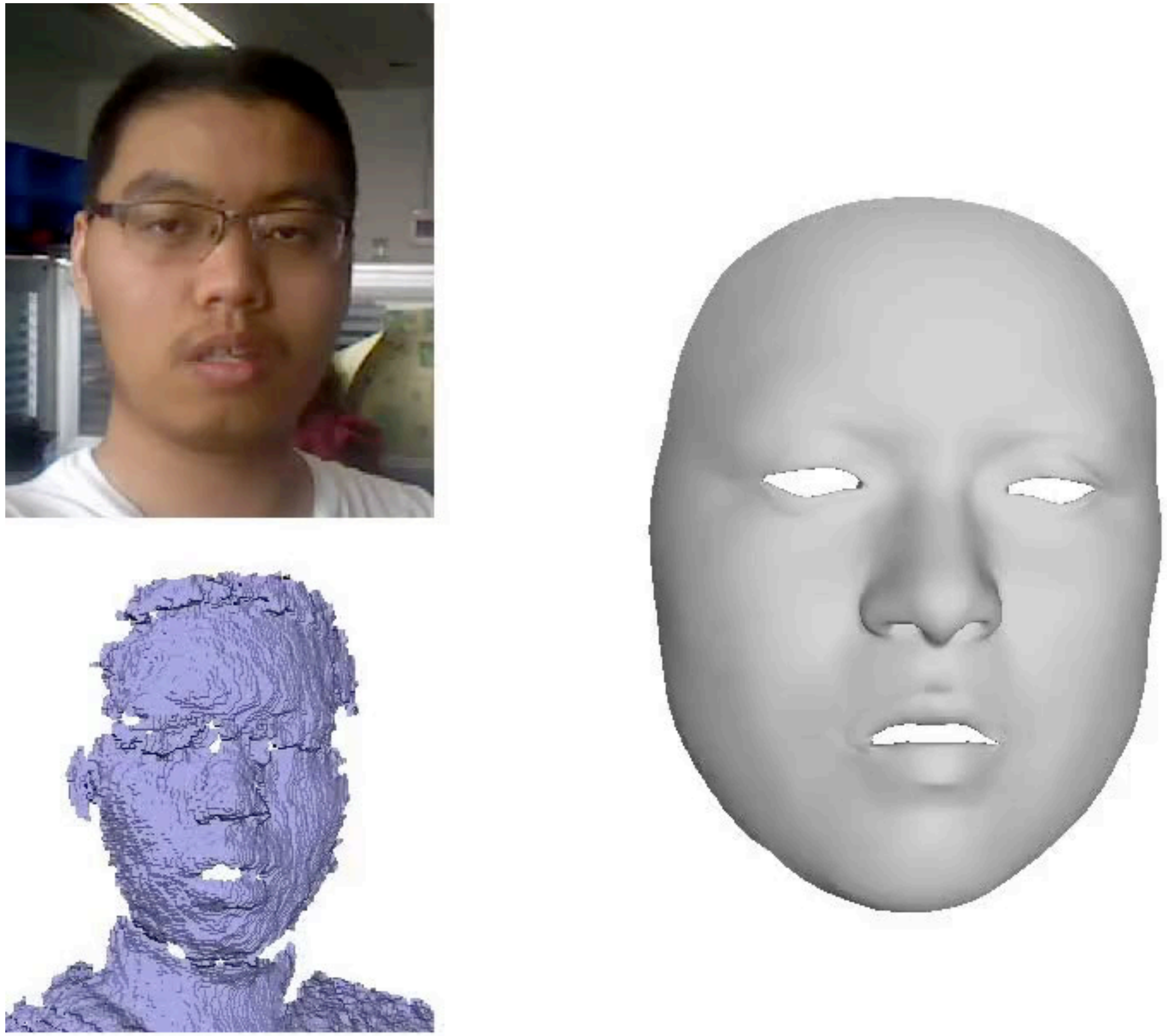
$$E(\mathcal{P}) = E_{\text{emb}}(\mathcal{P}) + w_{\text{col}} E_{\text{col}}(\mathcal{P}) + w_{\text{lan}} E_{\text{lan}}(\mathcal{P}) + w_{\text{reg}} E_{\text{reg}}(\mathcal{P})$$



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
  - non-rigid registration
  - blendshape refinement
- **High computation cost**. Not easy to port it to mobile platform.

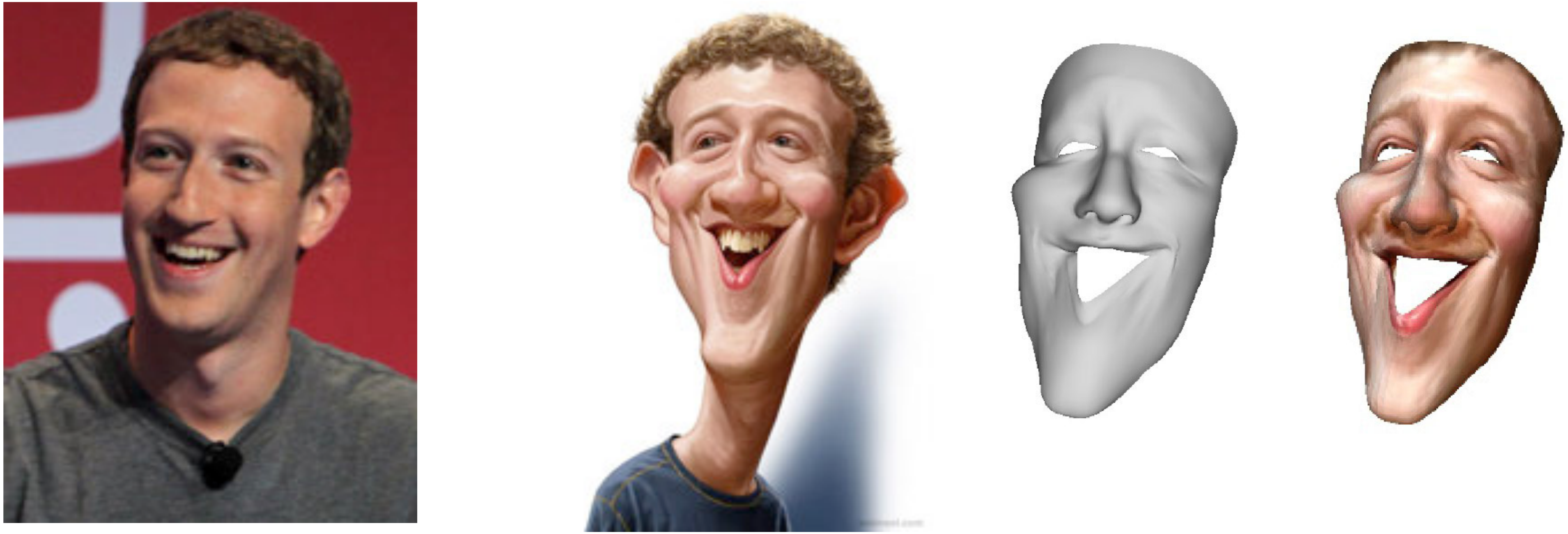
Our demo



Normal Face → Caricature

Alive Caricature from 2D to 3D
CVPR 2018, Spotlight Presentation

Problem



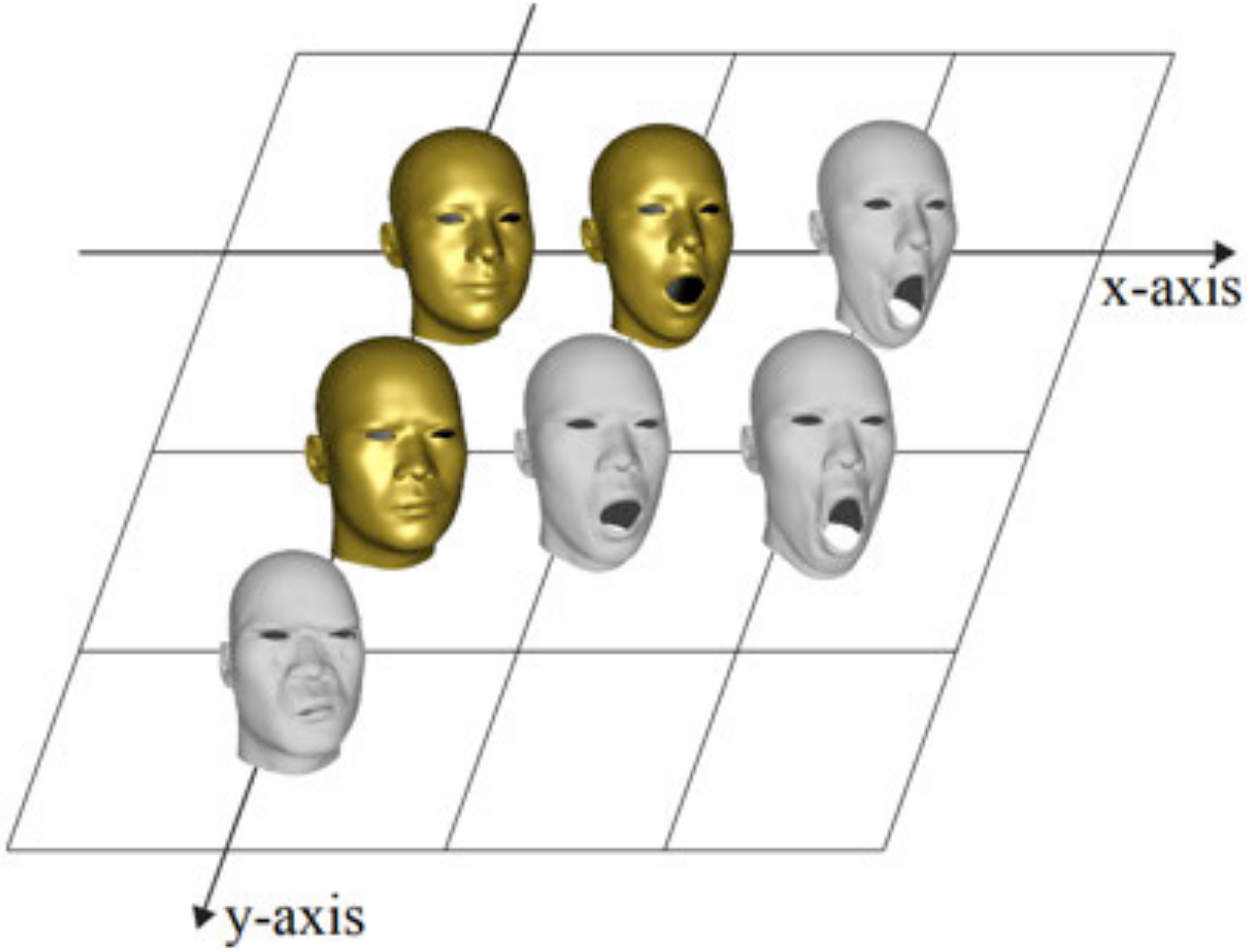
Blendshapes

novel expression neutral weights

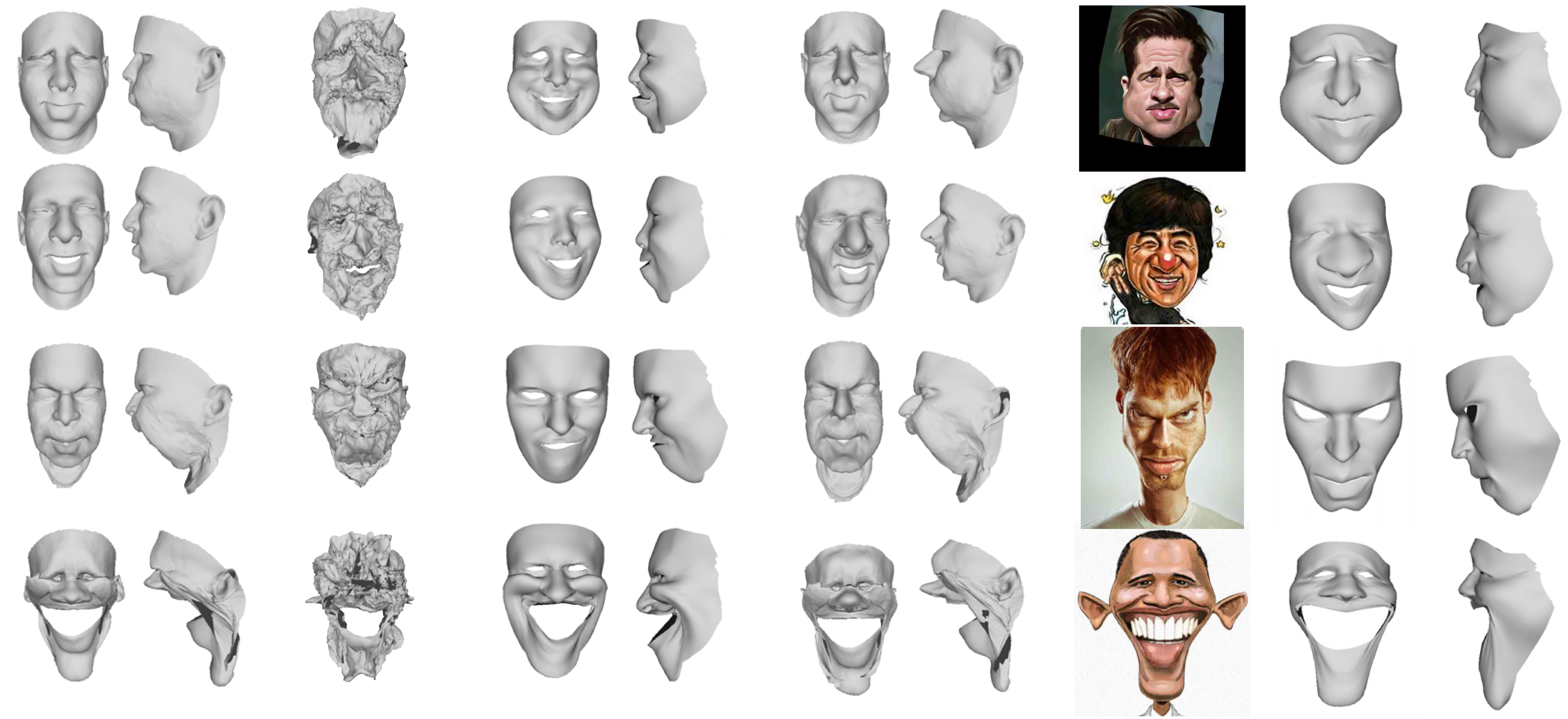
$$\mathbf{F}(\mathbf{x}) = \mathbf{b}_0 + \Delta\mathbf{B}\mathbf{x}$$

$$\Delta\mathbf{B} = [\mathbf{b}_1 - \mathbf{b}_0, \ldots, \mathbf{b}_n - \mathbf{b}_0]$$

3D Face Representation for extrapolation



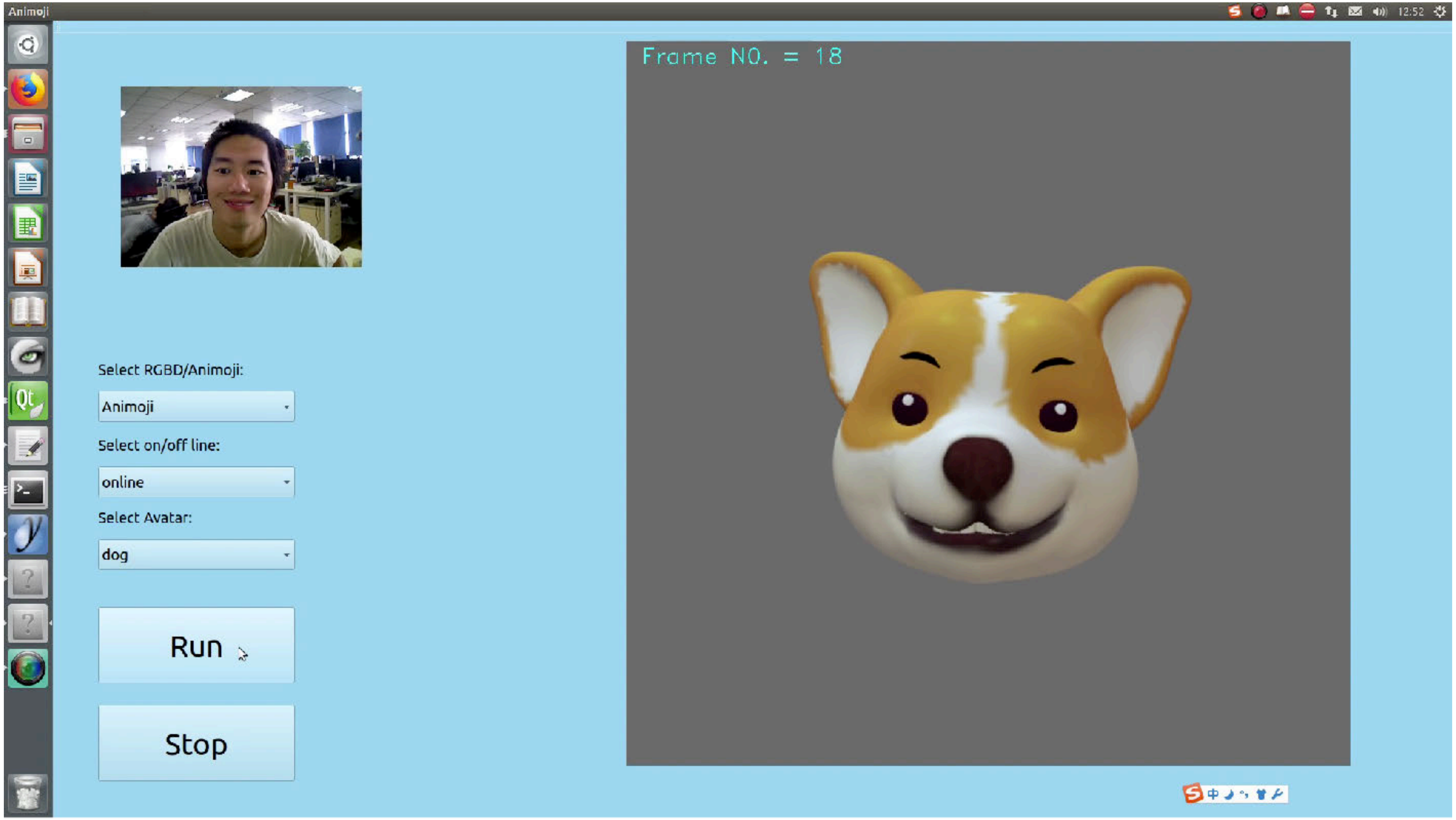
Results



(a)3DMM (b)3DMM(-) (c)FaceWareHouse (d)Caricatured 3DMM (e) Our Method

Animoji Demo

Animoji

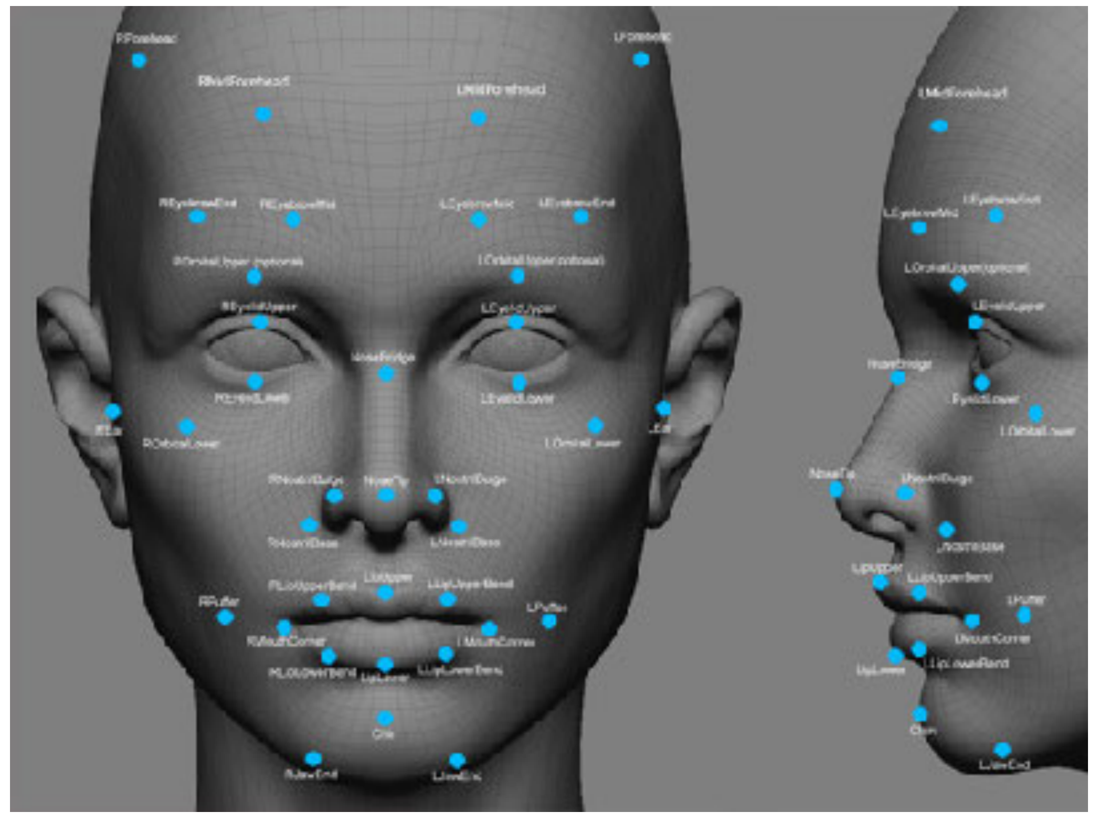


Animoji
Frame NO. = 18
Select RGBD/Animoji:
Animoji
Select on/off line:
online
Select Avatar:
dog
Run
Stop

Other Applications

Video Games

Communication



Security

Acknowledgments

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

NANYANG TECHNOLOGICAL UNIVERSITY

CARDIFF UNIVERSITY PRIFYSGOL CAERDYDD

NSFC