Self-supervised Multi-level Face Model Learning for Monocular Reconstruction at over 250 Hz
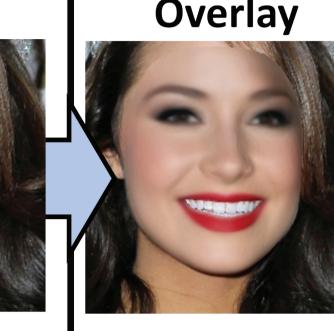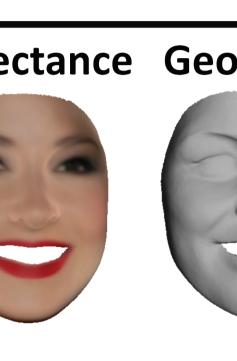
Ayush Tewari[1,2]   Michael Zollhöfer[1,2,3]   Pablo Garrido[1,2]   Florian Bernard[1,2]   Hyeongwoo Kim[1,2]   Patrick Pérez[4]   Christian Theobalt[1,2]
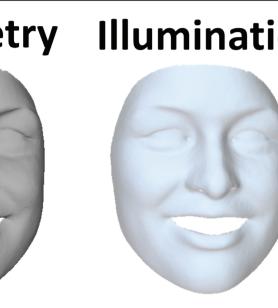
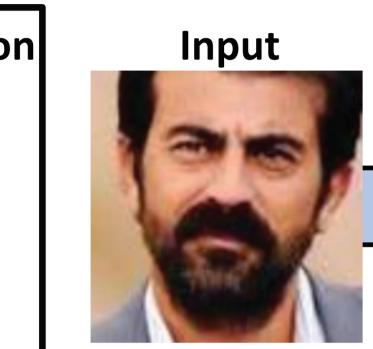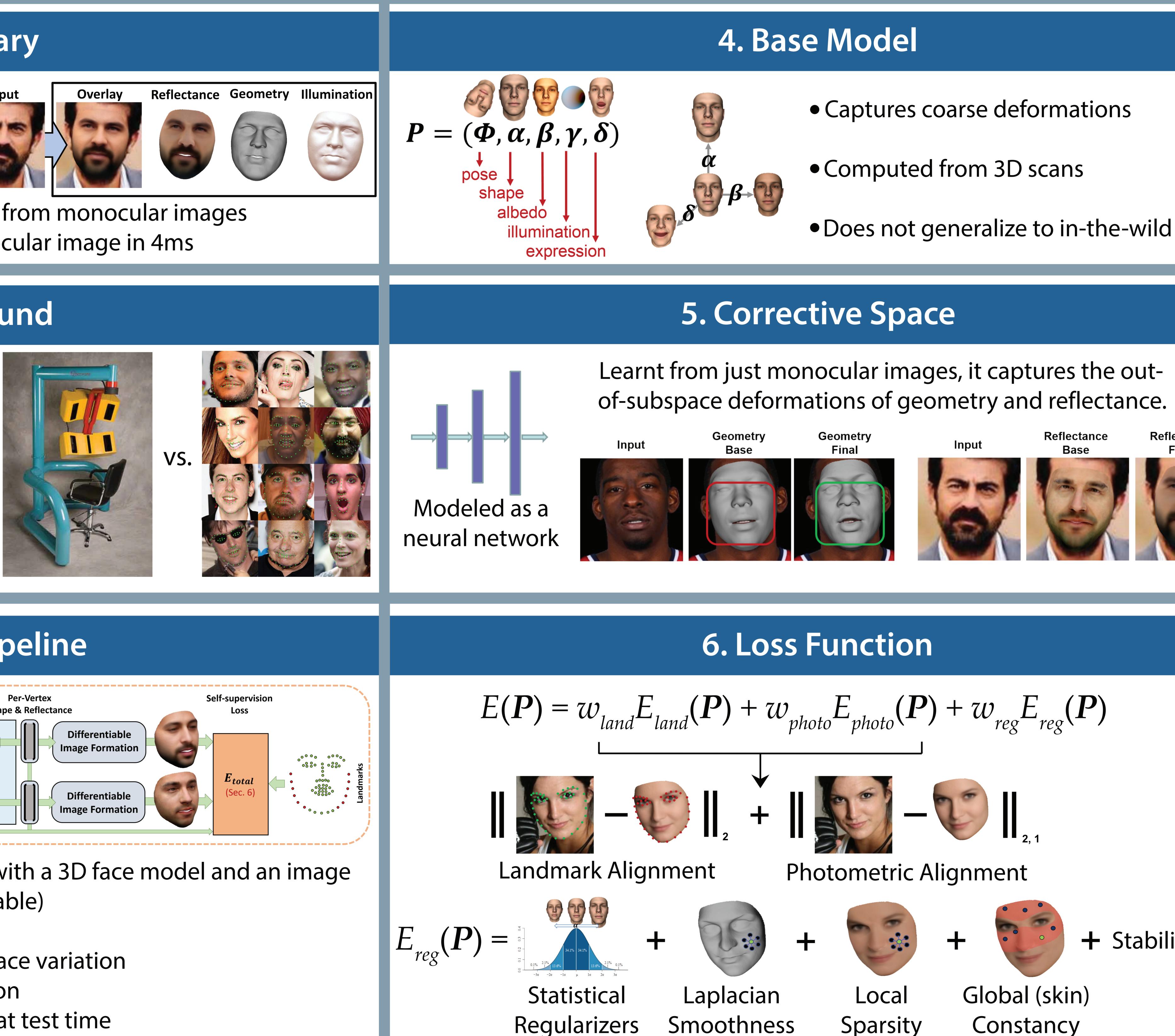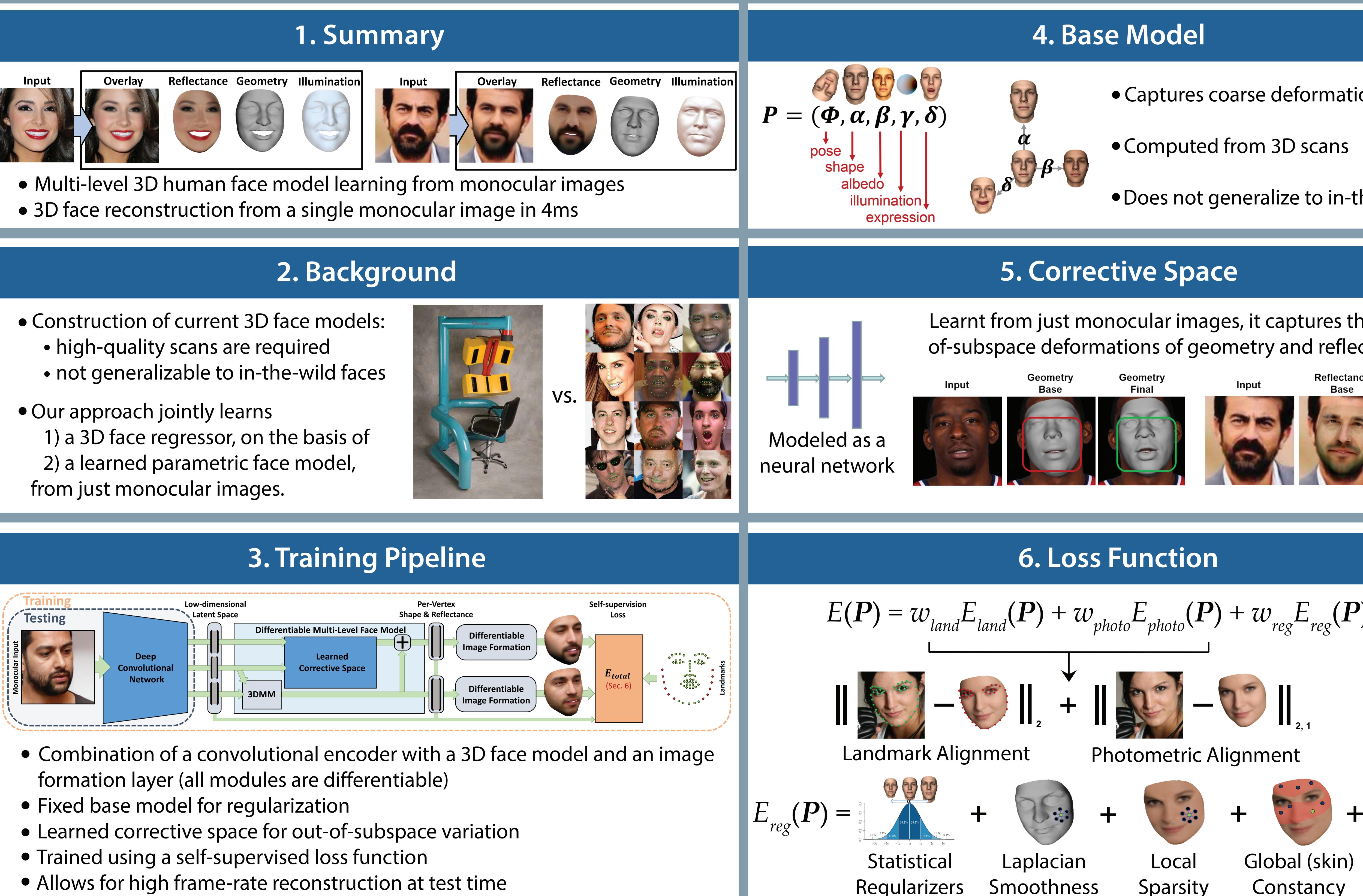[1]Max-Planck-Institute for Informatics   [2]Saarland Informatics Campus   [3]Stanford University   [4]Technicolor
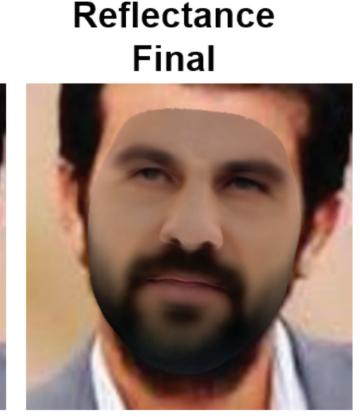
1. Summary

- Multi-level 3D human face model learning from monocular images
- 3D face reconstruction from a single monocular image in 4ms

2. Background

- Construction of current 3D face models:
  - high-quality scans are required
  - not generalizable to in-the-wild faces
- Our approach jointly learns
  1) a 3D face regressor, on the basis of
  2) a learned parametric face model,
  from just monocular images.

3. Training Pipeline

- Combination of a convolutional encoder with a 3D face model and an image formation layer (all modules are differentiable)
- Fixed base model for regularization
- Learned corrective space for out-of-subspace variation
- Trained using a self-supervised loss function
- Allows for high frame-rate reconstruction at test time

4. Base Model

$$P = (\Phi, \alpha, \beta, \gamma, \delta)$$

pose, shape, albedo, illumination, expression

- Captures coarse deformations
- Computed from 3D scans
- Does not generalize to in-the-wild faces

5. Corrective Space

Learnt from just monocular images, it captures the out-of-subspace deformations of geometry and reflectance.

Modeled as a neural network

6. Loss Function

$$E(P) = w_{land}E_{land}(P) + w_{photo}E_{photo}(P) + w_{reg}E_{reg}(P)$$

$$\| \cdot \|_2 + \| \cdot \|_{2,1}$$

Landmark Alignment   Photometric Alignment

$$E_{reg}(P) = \text{Statistical Regularizers} + \text{Laplacian Smoothness} + \text{Local Sparsity} + \text{Global (skin) Constancy} + \text{Stabilization}$$

7. Results and Evaluations

Evaluations of the Multi-level Face Model

Evaluation of different sizes of the corrective space

Comparisons to other methods