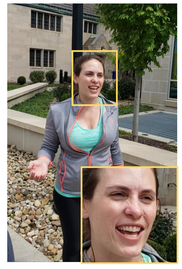
Self-Supervised Adaptation of High-Fidelity Face Models for Monocular Face Performance Tracking

Jae Shin Yoon[1], Takaaki Shiratori[2], Shoou-I Yu[2] and Hyun Soo Park[1]
[1] University of Minnesota [2] Facebook Reality Labs, Pittsburgh

Motivation



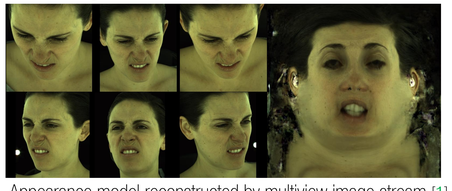
Input: monocular video

Output: rendered high fidelity 3D face model

Challenge

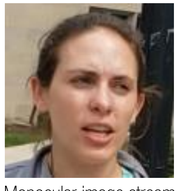
Existing approaches [1] to learn high fidelity appearance model require dense multiview image streams under a controlled environment, which cannot be applied to a monocular video.

Challenge 1: modality mismatch



Appearance model reconstructed by multiview image stream [1]

VS.



Monocular image stream

Challenge 2: domain mismatch

Approach

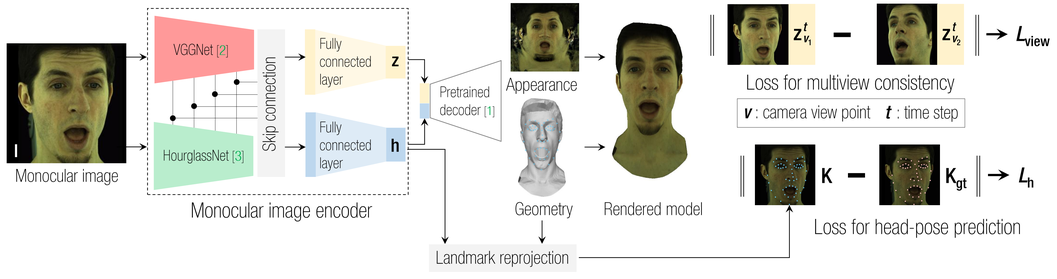
To address two mismatches, we
(1) design a new encoder that can take an monocular image to regress the latent code (z and h)
(2) propose a new self-supervision approach to refine the encoder by adapting to the testing monocular video.

Monocular image encoder for modality mismatch

$$L_E = \lambda_z L_z + \lambda_h L_h + \lambda_{view} L_{view}$$

Overall loss for training monocular image encoder

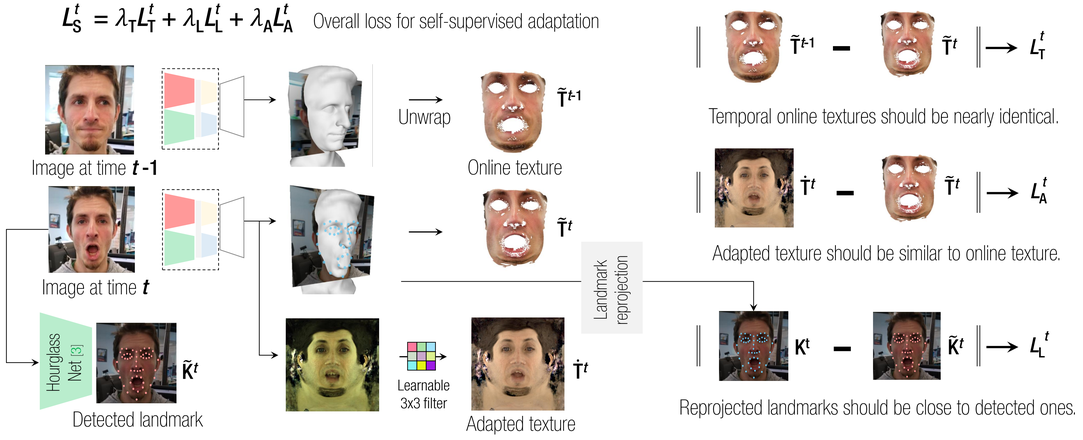
z : output latent face code
h : output head-pose parameter
K : reprojected face landmarks



Self-supervised adaptation by tracking on texture/landmark/appearance for domain mismatch

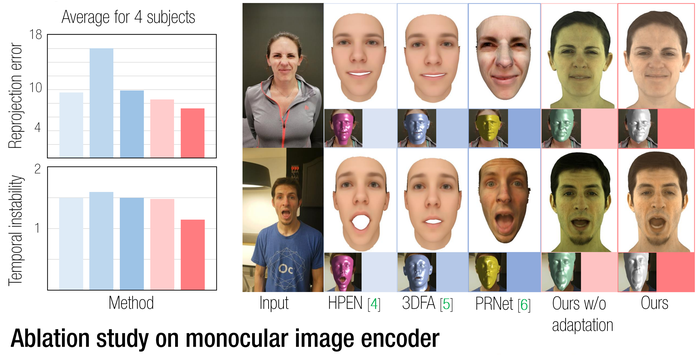
$$L_S^t = \lambda_T L_T^t + \lambda_L L_L^t + \lambda_A L_A^t$$

Overall loss for self-supervised adaptation

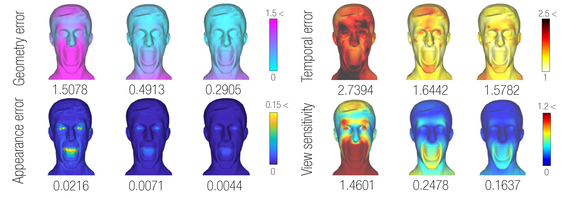


Evaluation

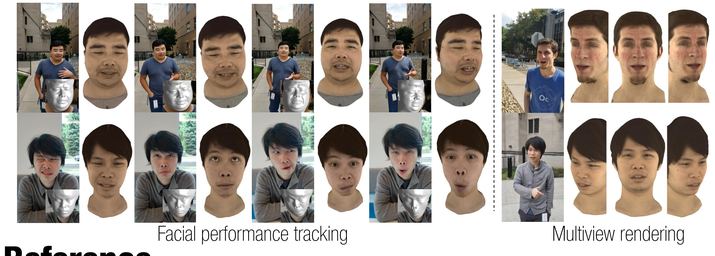
Comparison with state-of-the-art



Ablation study on monocular image encoder



Results



Reference

[1] Lombardi et al. "Deep Appearance Models for Face Rendering." SIGGRAPH 2018
[2] Simonyan et al. "Very Deep Convolutional Networks for Large-scale Image Recognition." ICLR 2015.
[3] Newell et al. "Stacked hourglass networks for human pose estimation." ECCV 2016
[4] Zhu et al. "High-fidelity Pose and Expression Normalization for Face Recognition In The Wild." CVPR. 2015.
[5] Zhu et al. "Face Alignment Across Large Poses: A 3D Solution." CVPR 2016.
[6] Feng et al. "Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network." ECCV 2018.