Joint Face Detection and Facial Motion Retargeting for Multiple Faces

Bindita Chaudhuri[1], Noranart Vesdapunt[2], Baoyuan Wang[2]

[1]University of Washington, [2]Microsoft Corporation

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Microsoft

webpage

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Objective

End-to-End Network
with real-time performance on mobile devices

Face Detection   Facial Retargeting
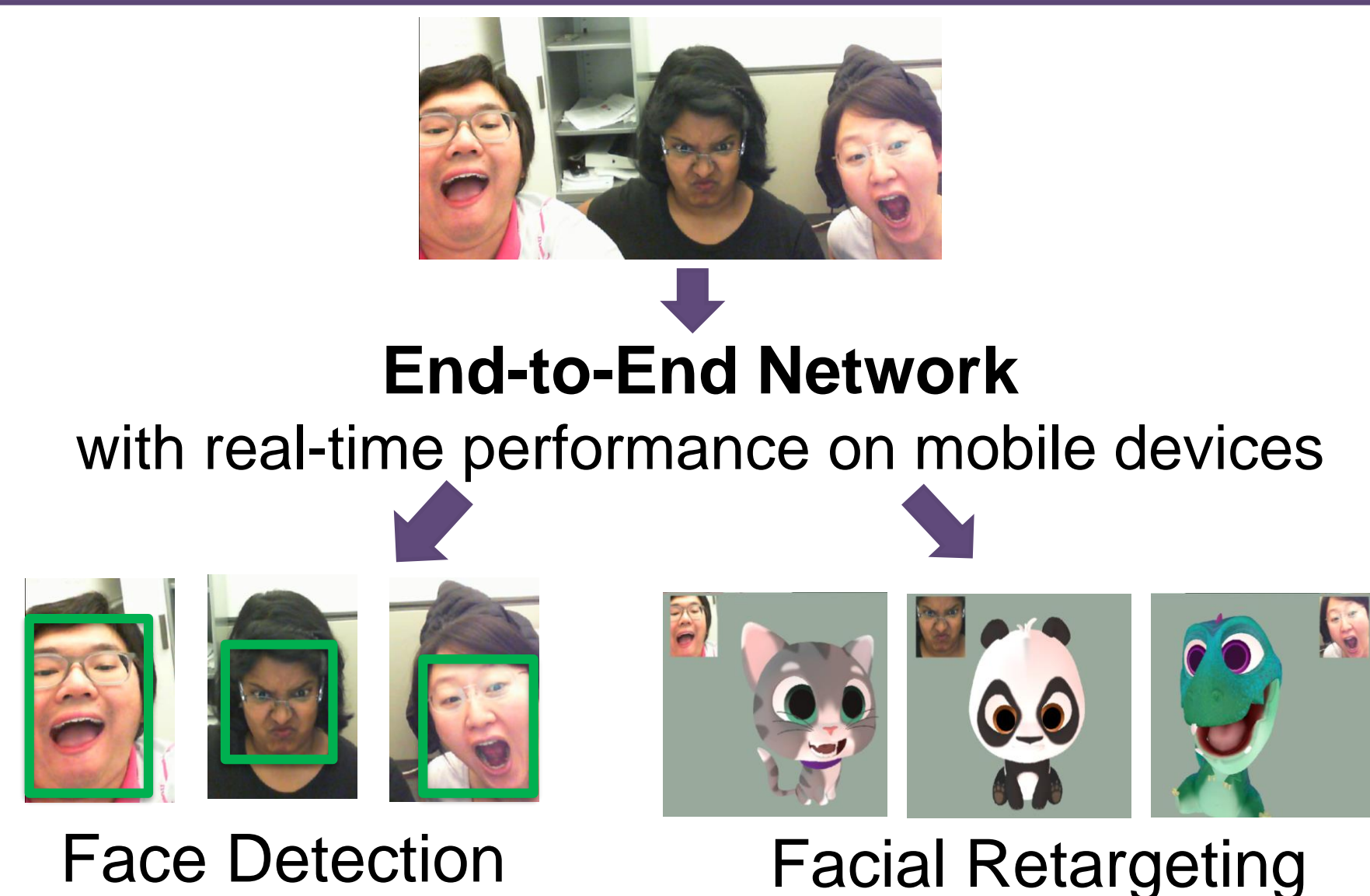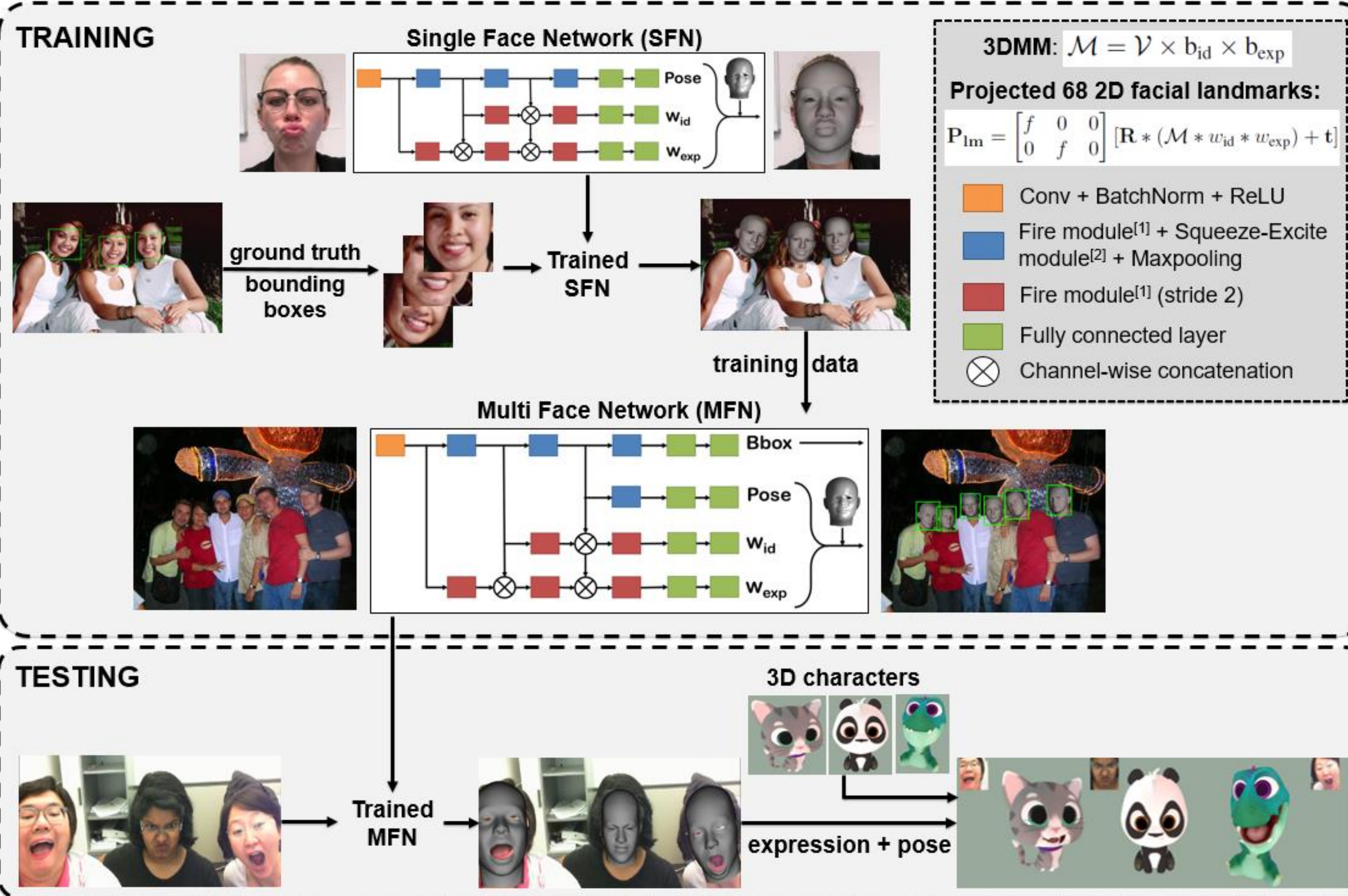
Contributions

- A novel top-down approach to **jointly learn**
  - bounding box locations
  - 3D Morphable Model (3DMM) parameters

- Multi-scale representation learning to **disentangle** the 3DMM parameters

- An **end-to-end real-time** memory-efficient system for practical applications with multi-face images (**26 fps** on Google Pixel 2)
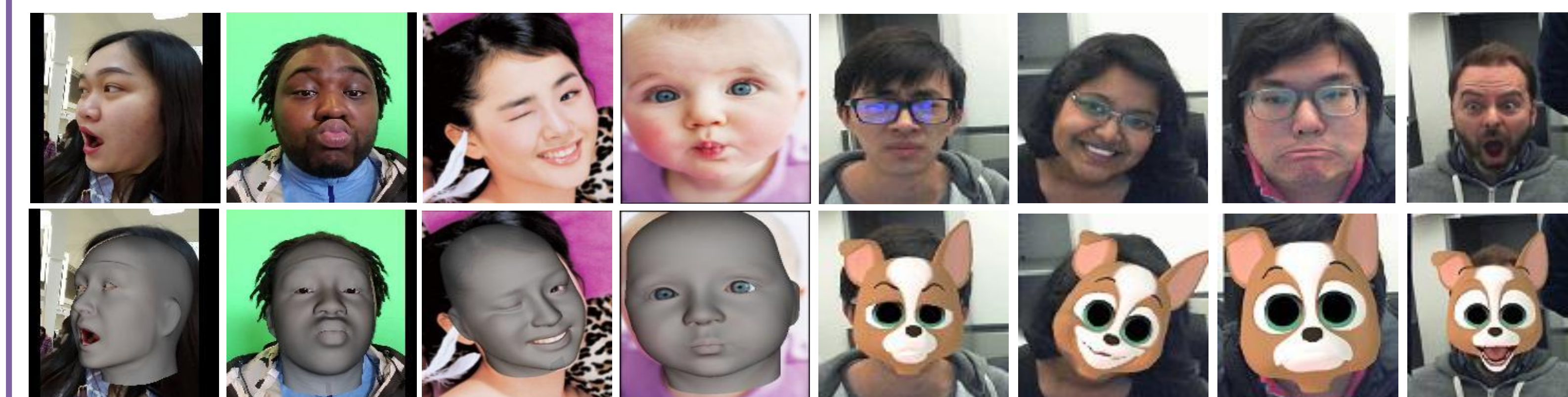
Multi-scale Representation

Correct fitting   Incorrect fitting   Same 2D landmarks

- Disentangling of 3DMM parameters is important for retargeting purpose

- Pose ($\mathbf{R}$, $\mathbf{t}$, f), identity (w_{id}) and expression (w_{exp}) parameters are learnt from global, regional and local features respectively

Methodology

TRAINING

Single Face Network (SFN)

Pose, W_{id}, W_{exp}

ground truth bounding boxes → Trained SFN → training data

Multi Face Network (MFN)

Bbox, Pose, W_{id}, W_{exp}

3DMM: $\mathcal{M} = \mathcal{V} \times b_{id} \times b_{exp}$

Projected 68 2D facial landmarks:

$$\mathbf{P}_{lm} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \end{bmatrix} [\mathbf{R} * (\mathcal{M} * w_{id} * w_{exp}) + \mathbf{t}]$$

- ■ Conv + BatchNorm + ReLU
- ■ Fire module[1] + Squeeze-Excite module[2] + Maxpooling
- ■ Fire module[1] (stride 2)
- ■ Fully connected layer
- ⊗ Channel-wise concatenation

TESTING

3D characters

Trained MFN → expression + pose → 3D characters

Quantitative Evaluation

Advantages of multi-scale design (ablation)

Normalized Mean Error (NME): 2.89, 2.65, 2.23

Single scale MFN with GT from Single scale SFN; Single scale MFN with GT from Multi-scale SFN; Multi-scale MFN with GT from Multi-scale SFN

- Lower 2D landmark error
- Better ground truth for MFN

Performance on single face videos

Area Under the Curve (AUC); Scenario 1, Scenario 2, Scenario 3
0.748, 0.76, 0.726; 0.847, 0.83, 0.769; 0.901, 0.884, 0.842
MTCNN [6]   MHM [7]   Multi-scale SFN

- Higher alignment accuracy than state-of-the-art

Advantages of joint training for face detection

Average Precision (AP): 92.1, 97.2, 97.9, 98.2, 98.8
MFN (only detection), Hyperface [4], Faceness Net [5], Single scale MFN, Multi-scale MFN
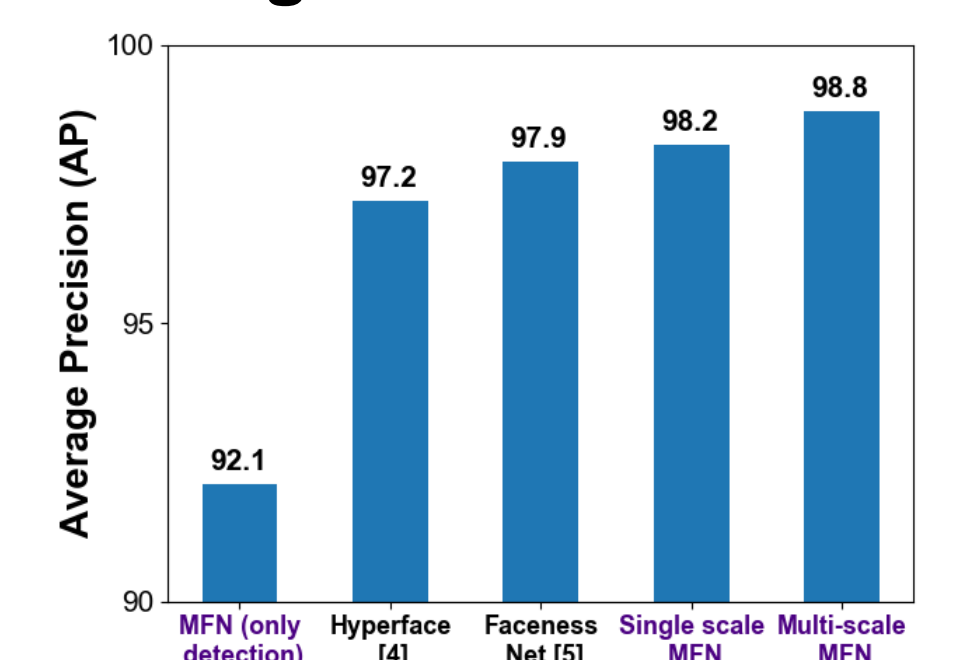
- Improved face detection due to 3DMM constraints

Time complexity analysis

Runtime (ms) vs #faces; Joint detection and retargeting; Separate detection and retargeting
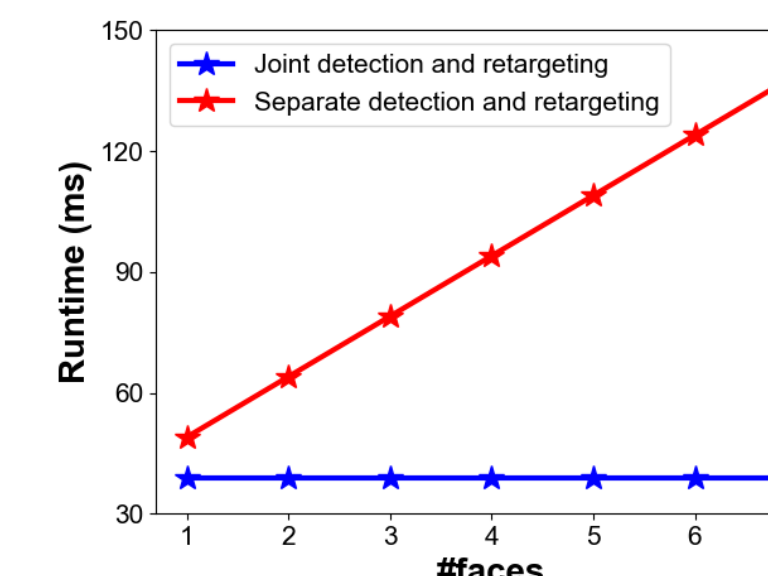
- Constant runtime
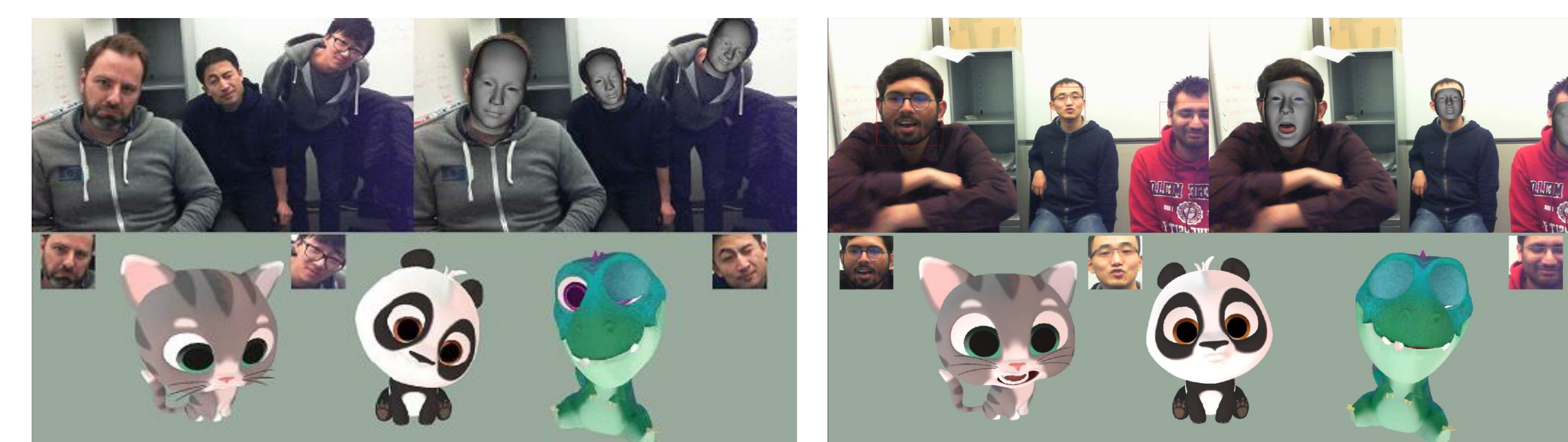- Faster compared to separate networks

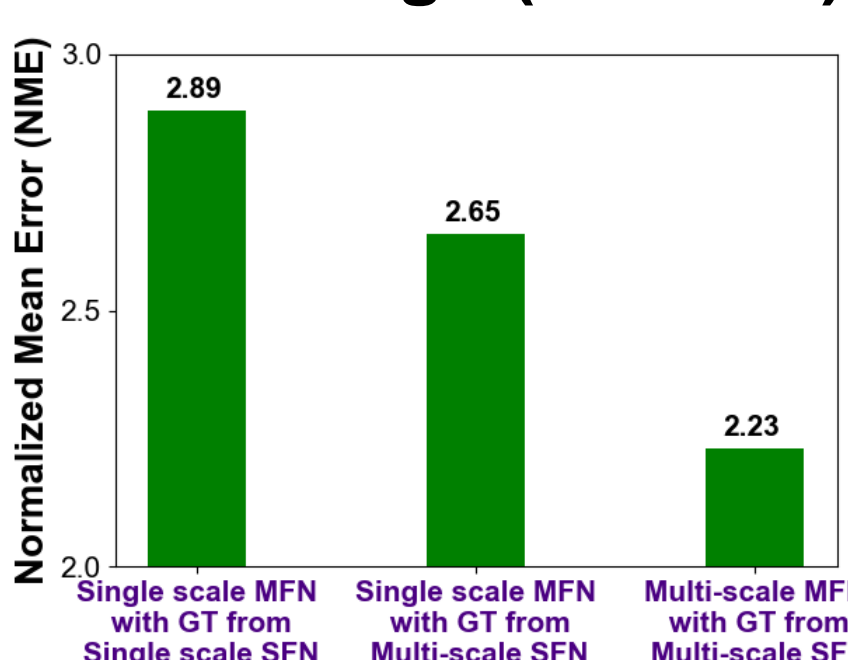## Qualitative Results

Single Face Retargeting (custom test set)

Multi Face Retargeting (network outputs for AFW and WIDER test set)

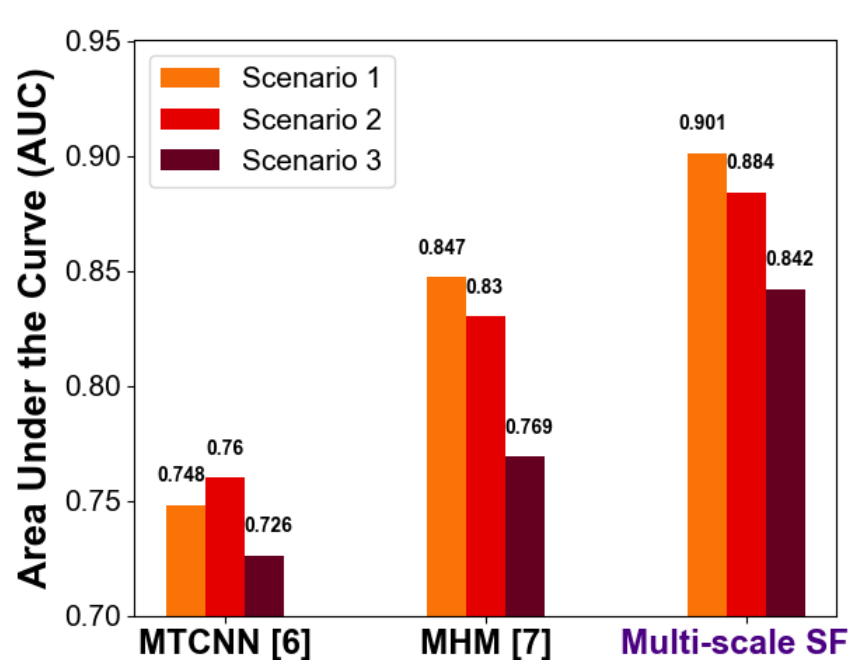Multi Face Retargeting (live performance capture using webcam and CPU)

REFERENCES: (1) Iandola et al.,"SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size", *arXiv*, 2016
(2) Hu et al.,"Squeeze-and-Excitation Networks", *CVPR*, 2018
(3) Joseph Redmon and Ali Farhadi, "YOLO9000: Better, Faster, Stronger", *CVPR*, 2017
(4) Ranjan et al.,"Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition", *TPAMI*, 2017
(5) Yang et al.,"From Facial Parts Responses to Face Detection: A Deep Learning Approach", *ICCV*, 2015
(6) Zhang et al.,"Joint face detection and alignment using multitask cascaded convolutional networks", *IEEE Signal Processing Letters*, 2016
(7) Deng et al.,"Joint multi-view face alignment in the wild", *arXiv*, 2017