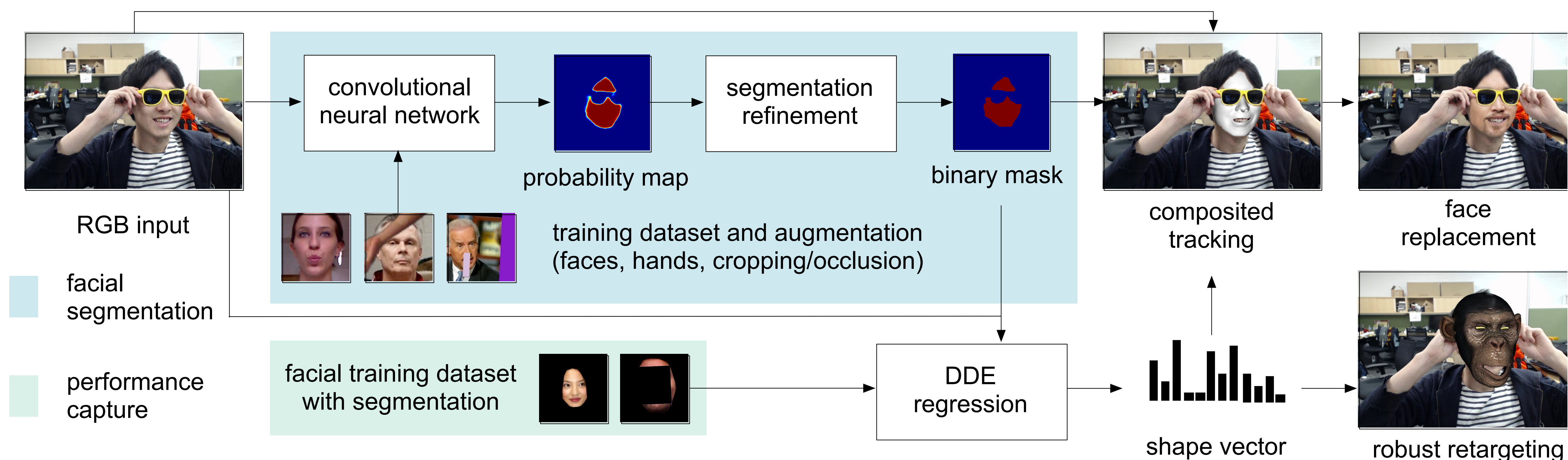# REAL-TIME FACIAL SEGMENTATION AND PERFORMANCE CAPTURE FROM RGB INPUT

ECCV'16

Shunsuke Saito **Tianye Li** **Hao Li**

Pinscreen University of Southern California

RGB input

facial segmentation

performance capture

convolutional neural network · probability map · segmentation refinement · binary mask

training dataset and augmentation (faces, hands, cropping/occlusion)

facial training dataset with segmentation · DDE regression · shape vector

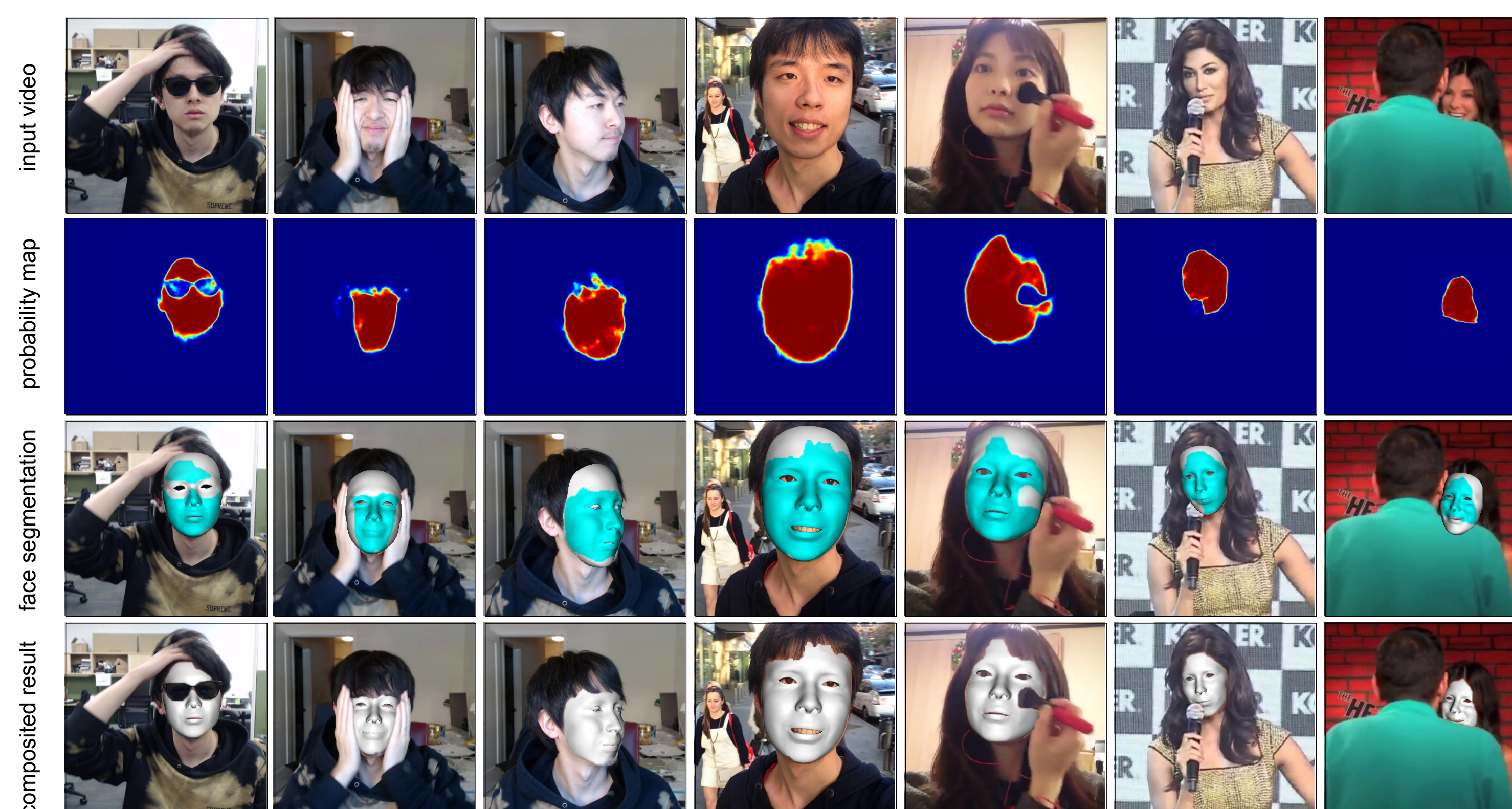composited tracking · face replacement · robust retargeting

ABSTRACT

We introduce the concept of unconstrained real-time 3D facial performance capture through explicit semantic segmentation in the RGB input. To ensure robustness, cutting edge supervised learning approaches rely on large training datasets of face images captured in the wild. While impressive tracking quality has been demonstrated for faces that are largely visible, any occlusion due to hair, accessories, or hand-to-face gestures would result in significant visual artifacts and loss of tracking accuracy. The modeling of occlusions has been mostly avoided due to its immense space of appearance variability. To address this curse of high dimensionality, we perform tracking in unconstrained images assuming non-face regions can be fully masked out. Along with recent breakthroughs in deep learning, we demonstrate that pixel-level facial segmentation is possible in real-time by repurposing convolutional neural networks designed originally for general semantic segmentation. We develop an efficient architecture based on a two-stream deconvolution network with complementary characteristics, and introduce carefully designed training samples and data augmentation strategies for improved segmentation accuracy and robustness. We adopt a state-of-the-art regression-based facial tracking framework with segmented face images as training, and demonstrate accurate and uninterrupted facial performance capture in the presence of extreme occlusion and even side views. Furthermore, the resulting segmentation can be directly used to composite partial 3D face models on the input images and enable seamless facial manipulation tasks, such as virtual make-up or face replacement.
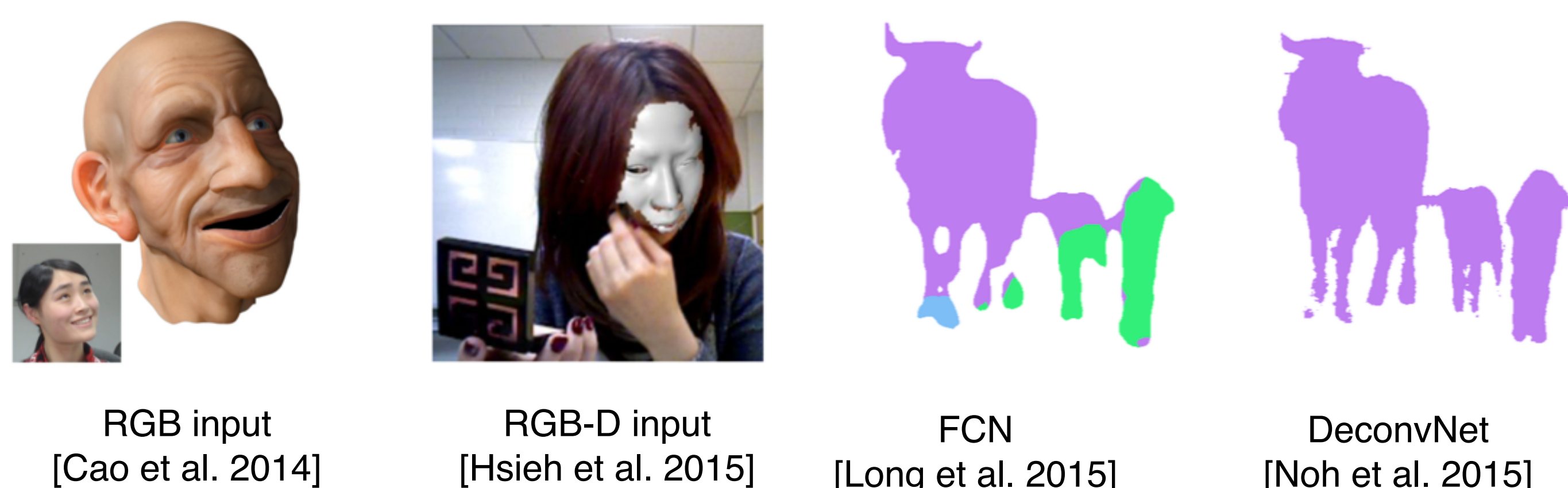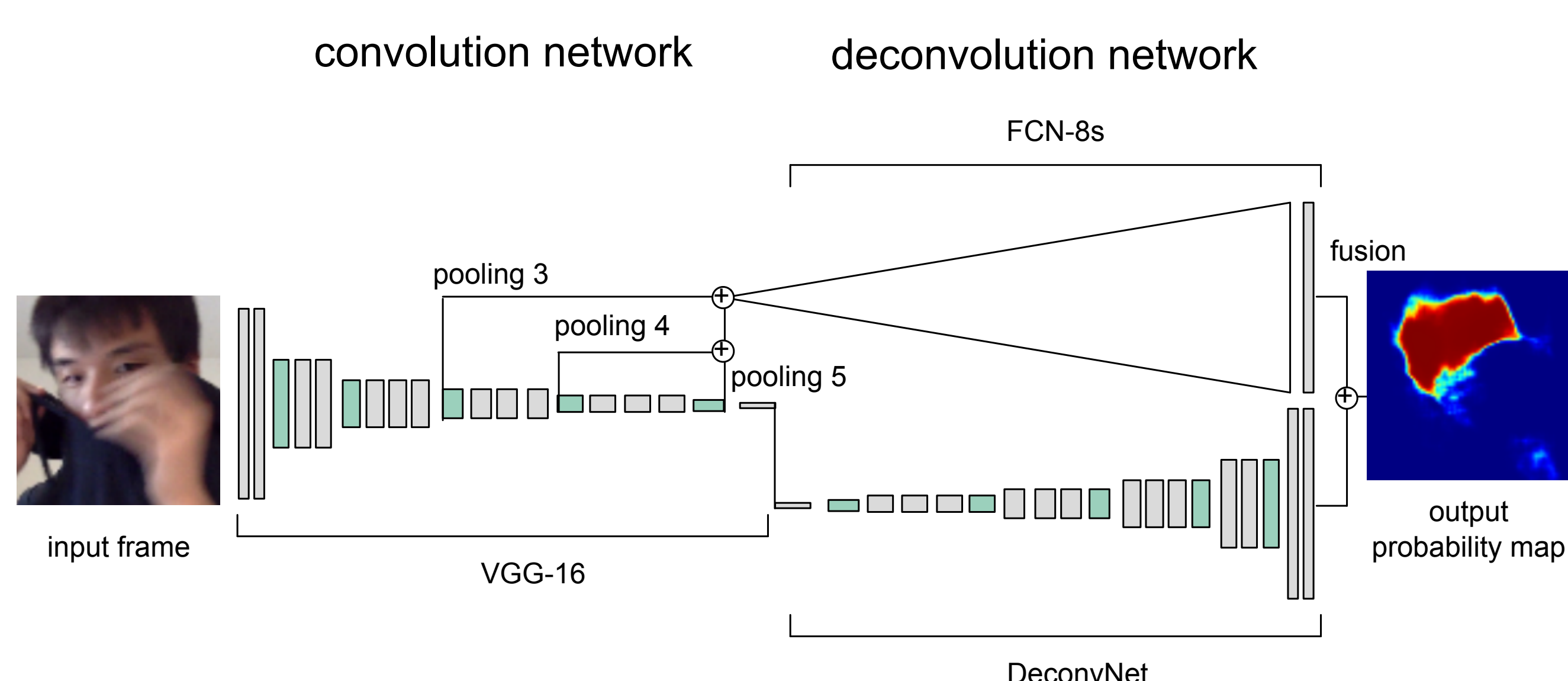
PRIOR WORK

facial performance capture

RGB input [Cao et al. 2014]

RGB-D input [Hsieh et al. 2015]

semantical segmentation

FCN [Long et al. 2015]

DeconvNet [Noh et al. 2015]

SEGMENTATION NETWORK

convolution network · deconvolution network

FCN-8s

pooling 3 · pooling 4 · pooling 5 · fusion

input frame · VGG-16 · DeconvNet · output probability map

TRAINING DATA & AUGMENTATION

input image · labeled segmentation · input image · with occlusion augmentation · face data · hand data · cropping / occlusion · negative samples

RESULTS

input video · probability map · face segmentation · composited result

EVALUATION

input video · no augmentation · occlusion augmentation · occlusion+hand augmentation

input video · FCN · DeconvNet · EDeconvNet · our method

expression

rotation

translation

tracking error (mm) · frame

occlusion · Cao et al. 2014 · our method · our method without data augmentation