Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos

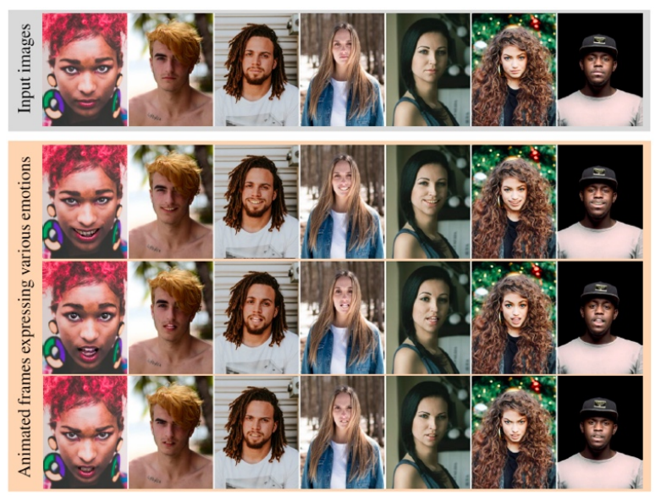
Huy H. Nguyen (SOKENDAI, Japan) Fuming Fang (NII, Japan) Junichi Yamagishi (NII, Japan) Isao Echizen (NII, Japan)

Generating of Fake Videos Impersonating a Person Using Deep Learning

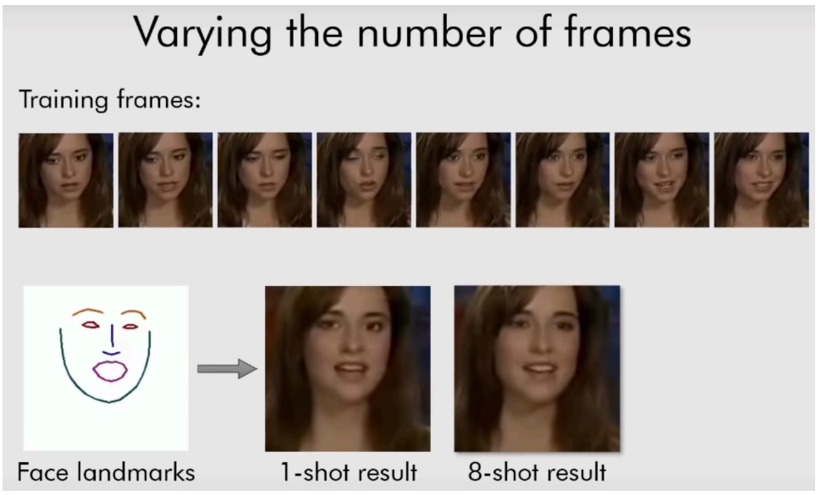
Real-time Facial Reenactment
Live capture using a commodity webcam

Face2Face: Real-time facial reenactment
(Thies et al. 2016)

Deepfakes
Video face swapping
(2017)

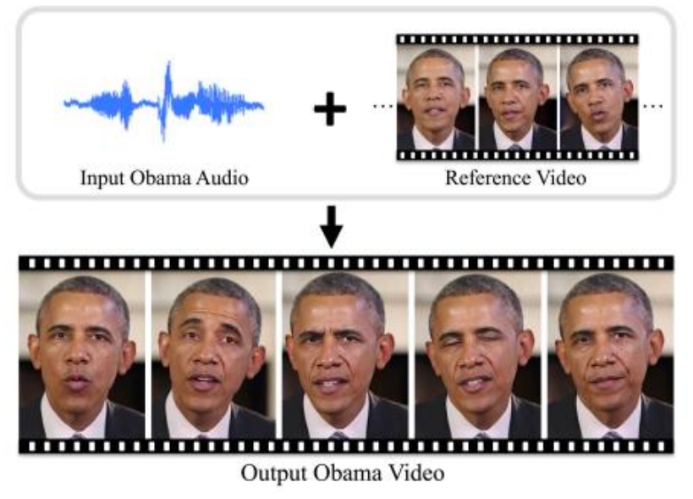


Bringing portraits to life
(Averbuch-Elor et al. 2017)

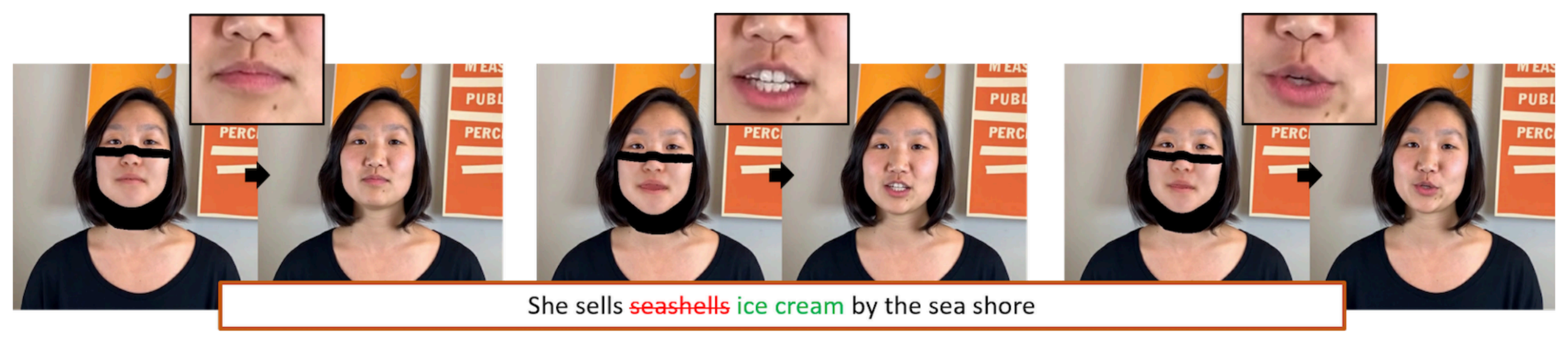


Realistic Neural Talking Head Models
(Zakharov et al. 2019)

Speech2Vid
(Chung et al. 2017)

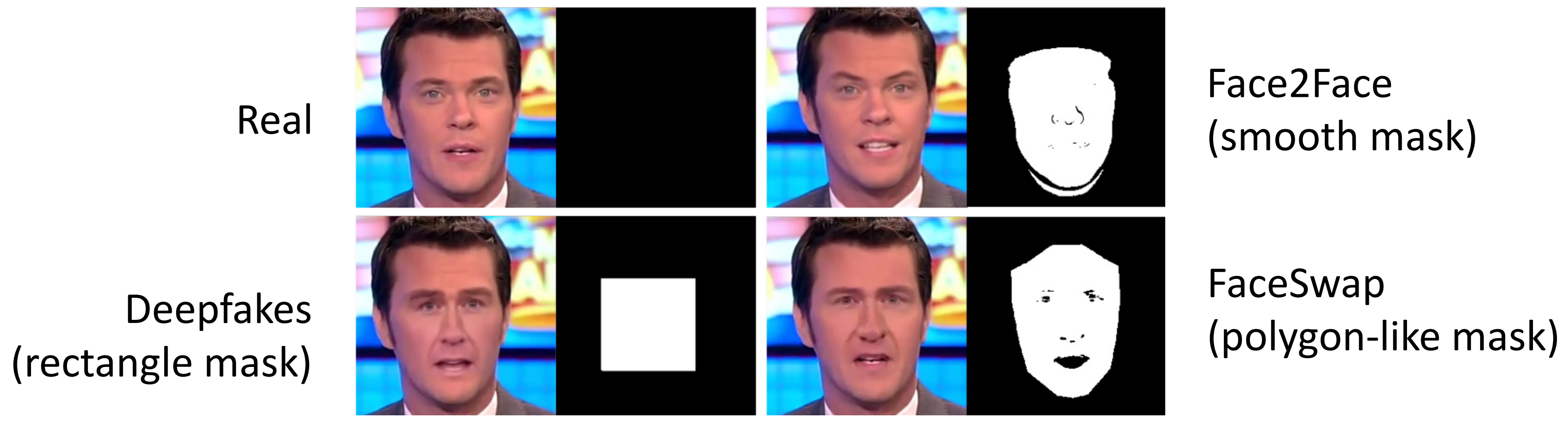


Synthesizing Obama: Learning lip sync from audio
(Suwajanakorn et al. 2017)



Text-based Editing of Talking-head Video
(Fried et al. 2019)

Overview



Example of a natural image and three corresponding manipulations: Deepfakes, Face2Face, and FaceSwap.
→ Based on segmentation result, we can infer the manipulated method.

Aim: Solving **3 problems simultaneously**:

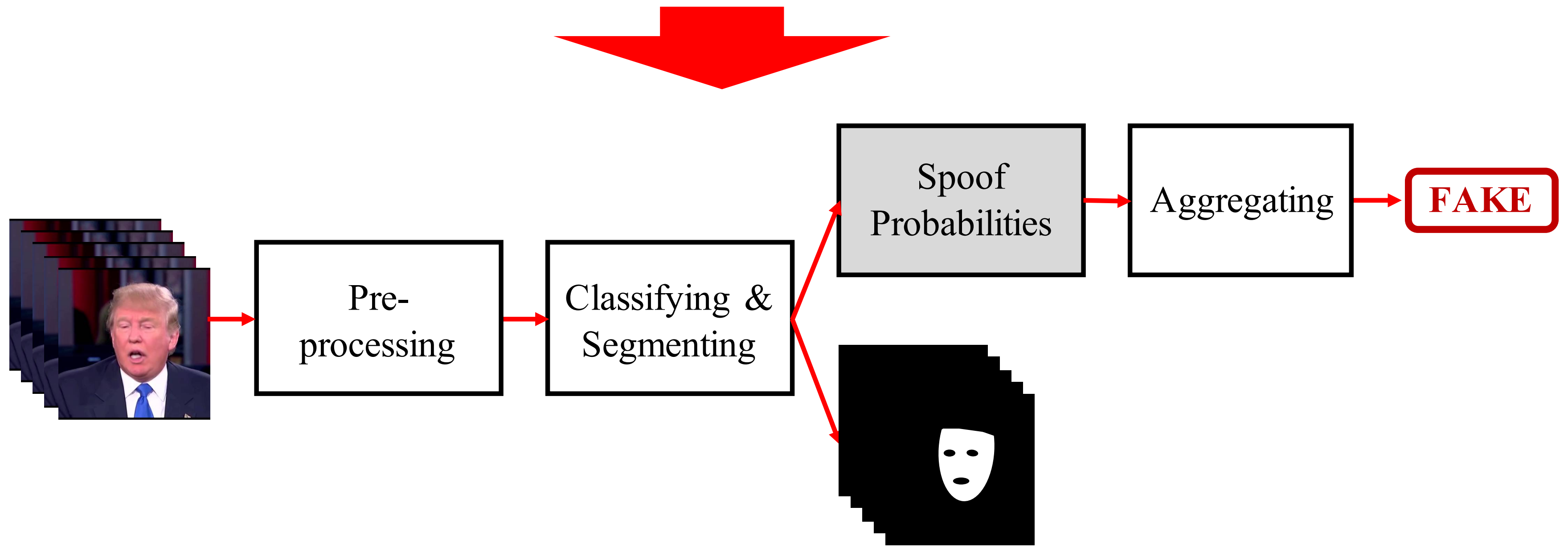
1. Identifying manipulated images/videos (PAD → classification).
2. Specifying manipulated regions (tampering detection → segmentation).
3. Detecting unseen attacks (transferability/cross-database detection).

Solution:

- Combining classification (real or fake), segmentation (tampering detection), and image reconstruction in a single network → multi-task learning.

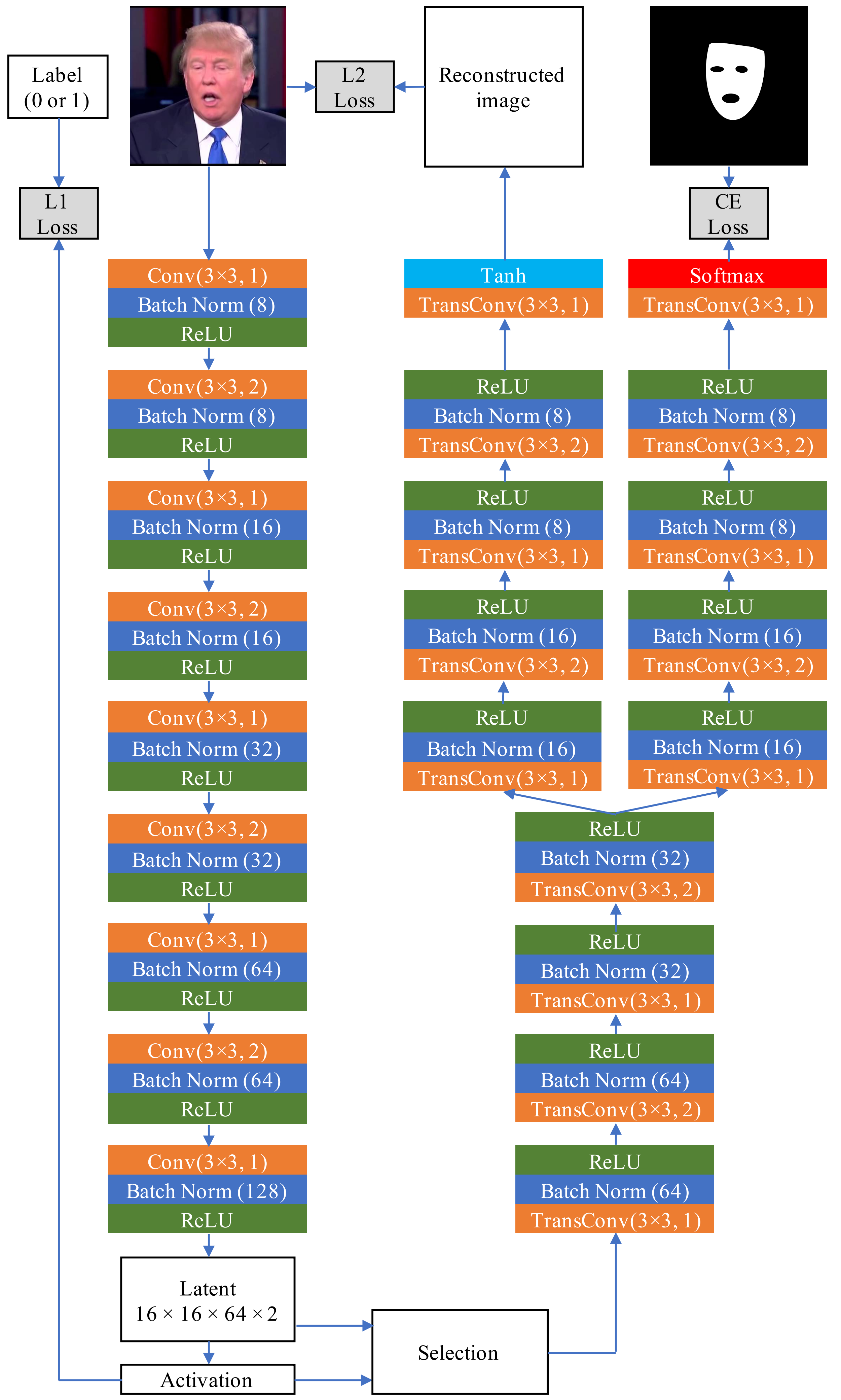
→ Sharing mutual information between tasks to improving the overall performance.

- Giving more information to judge the origin of the input (real or fake).

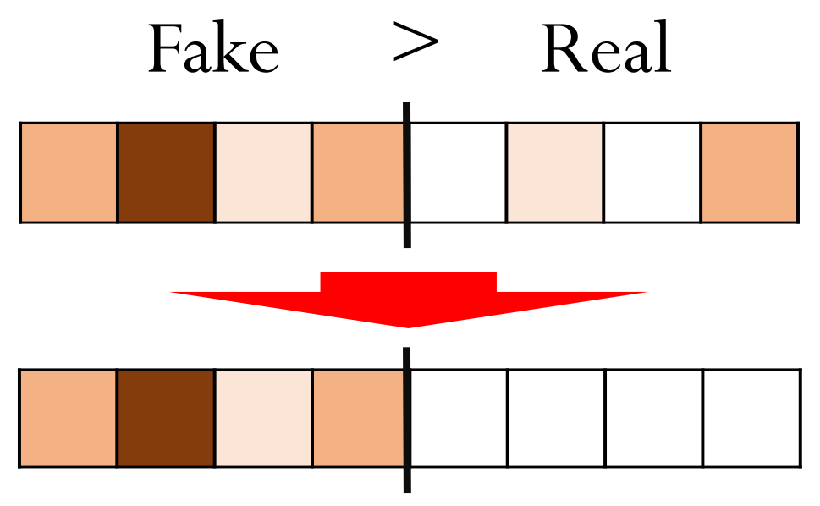


Pre-processing
Classifying & Segmenting
Spoof Probabilities
Aggregating
FAKE

Network Architecture

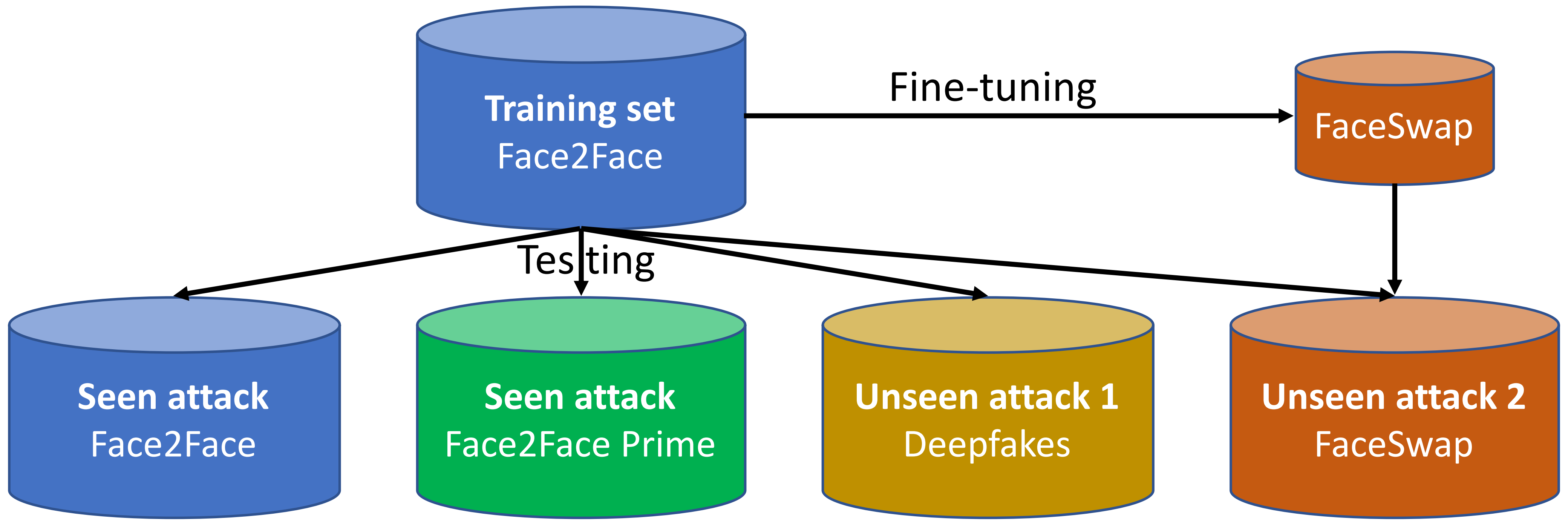


Label (0 or 1)
L1 Loss
L2 Loss
Reconstructed image
CE Loss
Conv(3×3, 1) Batch Norm (8) ReLU
Conv(3×3, 2) Batch Norm (8) ReLU
Conv(3×3, 1) Batch Norm (16) ReLU
Conv(3×3, 2) Batch Norm (16) ReLU
Conv(3×3, 1) Batch Norm (32) ReLU
Conv(3×3, 2) Batch Norm (32) ReLU
Conv(3×3, 1) Batch Norm (64) ReLU
Conv(3×3, 2) Batch Norm (64) ReLU
Conv(3×3, 1) Batch Norm (128) ReLU
Latent 16 × 16 × 64 × 2
Activation
Selection
TransConv(3×3, 1) Batch Norm (64) ReLU
TransConv(3×3, 2) Batch Norm (64) ReLU
TransConv(3×3, 1) Batch Norm (32) ReLU
TransConv(3×3, 2) Batch Norm (32) ReLU
TransConv(3×3, 1) Batch Norm (16) ReLU
TransConv(3×3, 2) Batch Norm (16) ReLU
TransConv(3×3, 1) Batch Norm (8) ReLU
TransConv(3×3, 2) Batch Norm (8) ReLU
TransConv(3×3, 1) Tanh
TransConv(3×3, 1) Batch Norm (16) ReLU
TransConv(3×3, 2) Batch Norm (16) ReLU
TransConv(3×3, 1) Batch Norm (8) ReLU
TransConv(3×3, 2) Batch Norm (8) ReLU
TransConv(3×3, 1) Softmax



Latent features are divided into two halves. The one with stronger activation will go through the decoder. The other one will be silent.

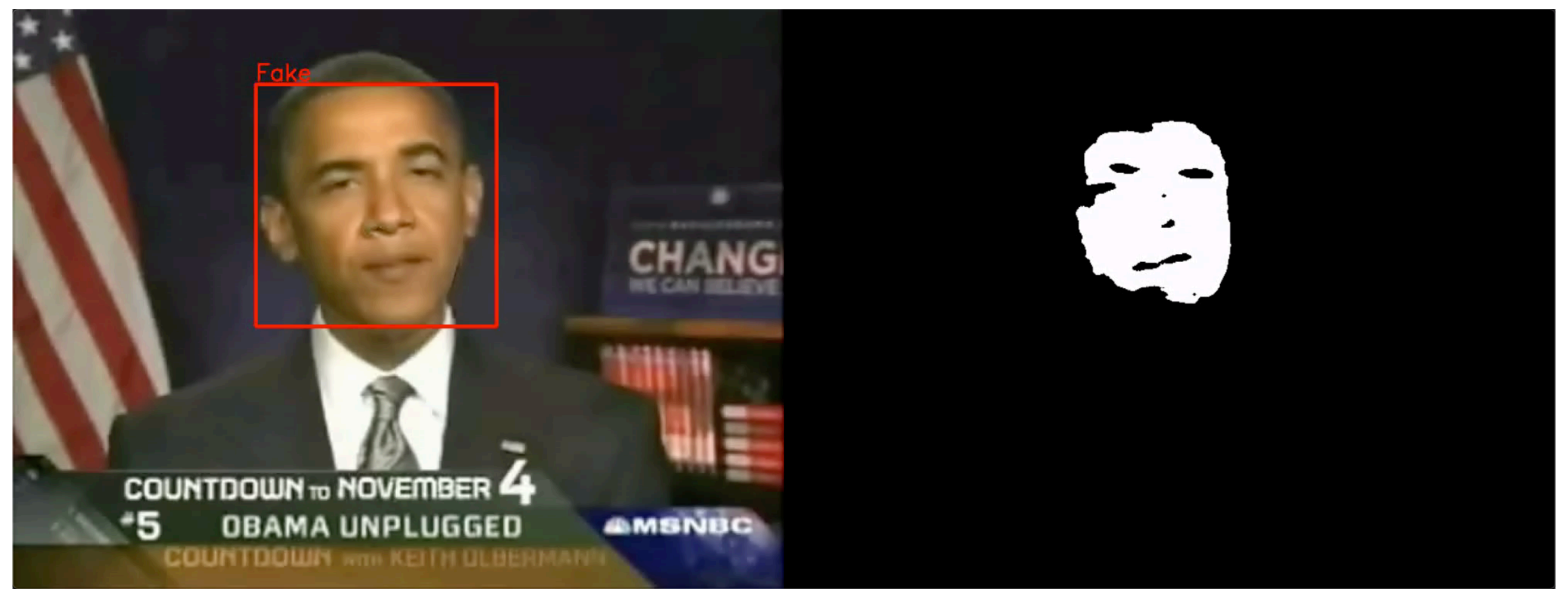
Evaluation



Training set Face2Face
Fine-tuning
FaceSwap
Testing
Seen attack Face2Face
Seen attack Face2Face Prime
Unseen attack 1 Deepfakes
Unseen attack 2 FaceSwap

Type of attack	Database (Medium compression)	Classification		Segmentation
		Accuracy (%)	EER (%)	Accuracy (%)
Match condition of seen attack	FaceForensics (Face2Face) Source-to-target	92.77	8.18	90.27
Mismatch condition of seen attack	FaceForensics (Face2Face) Self-reenactment	92.50	8.07	90.20
Unseen attack 1 (without fine-tuning)	FaceForensics++ Deepfakes	52.32	42.24	70.37
Unseen attack 2 (without fine-tuning)	FaceForensics++ FaceSwap	54.07	34.04	84.67
Unseen attack 2 (fine-tuning on small data)	FaceForensics++ FaceSwap	83.71	15.07	93.01

An example of detection and segmentation result on a video frame of the former US president Barack Obama modified by Face2Face method.



JST-ANR VoicePersonae Project

GitHub

Email: nhhuy@nii.ac.jp