大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

SOKENDAI

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

Huy H. Nguyen*, Fuming Fang,
Junichi Yamagishi, and Isao Echizen

BTAS 2019

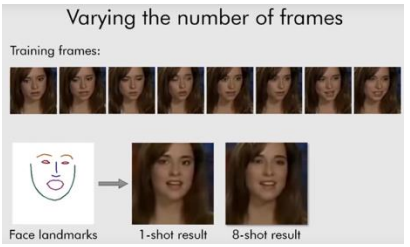
1. Motivation

It is very easy to create high-quality manipulated videos!

Face2Face: Real-time Face Capture and Reenactment of RGB Videos
(Thies et al. 2016)

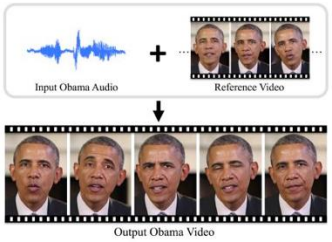
Deepfakes
(2017)

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



Few-Shot Adversarial Learning of 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)

1. Motivation

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)

2. Related Work in Manipulated Face Detection

2.1. Classification

- Evaluated using one or a few databases: Deepfakes (Afchar et al. 2018, Li et al. 2018, Korshunov and Marcel 2019,), FaceForensics & FaceForensics++ (Rossler et al. 2018, Rossler et al. 2019).

- Cross-database transferability evaluation (Cozzolino et al. 2018).

→ Conventional methods failed to detect unseen manipulation methods.

2. Related Work in Tampering Detection

2.2. Segmentation

- Most methods focus on 3 commonly used means of tampering: **removal**, **copy-move**, and **splicing**.

- Need to process full-scale images (Bappy et al. 2017, Bappy et al. 2019, Zhou et al. 2018).
- Using sliding window (Rahmouni et al. 2017, Nguyen et al. 2018, Rossler et al. 2018).

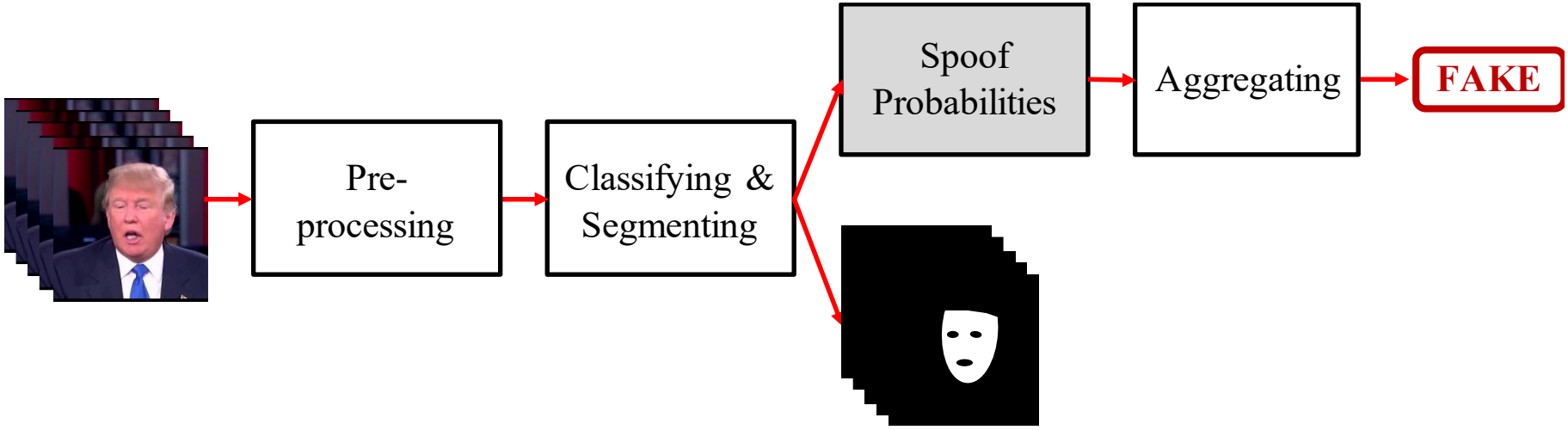
3. Proposed method

- 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.

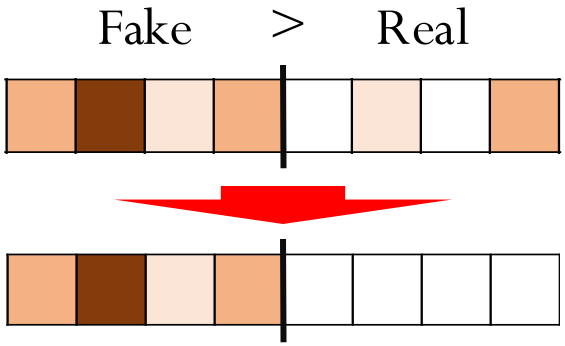
- Giving more information to judge the origin of the input (real or fake).

3. Proposed method

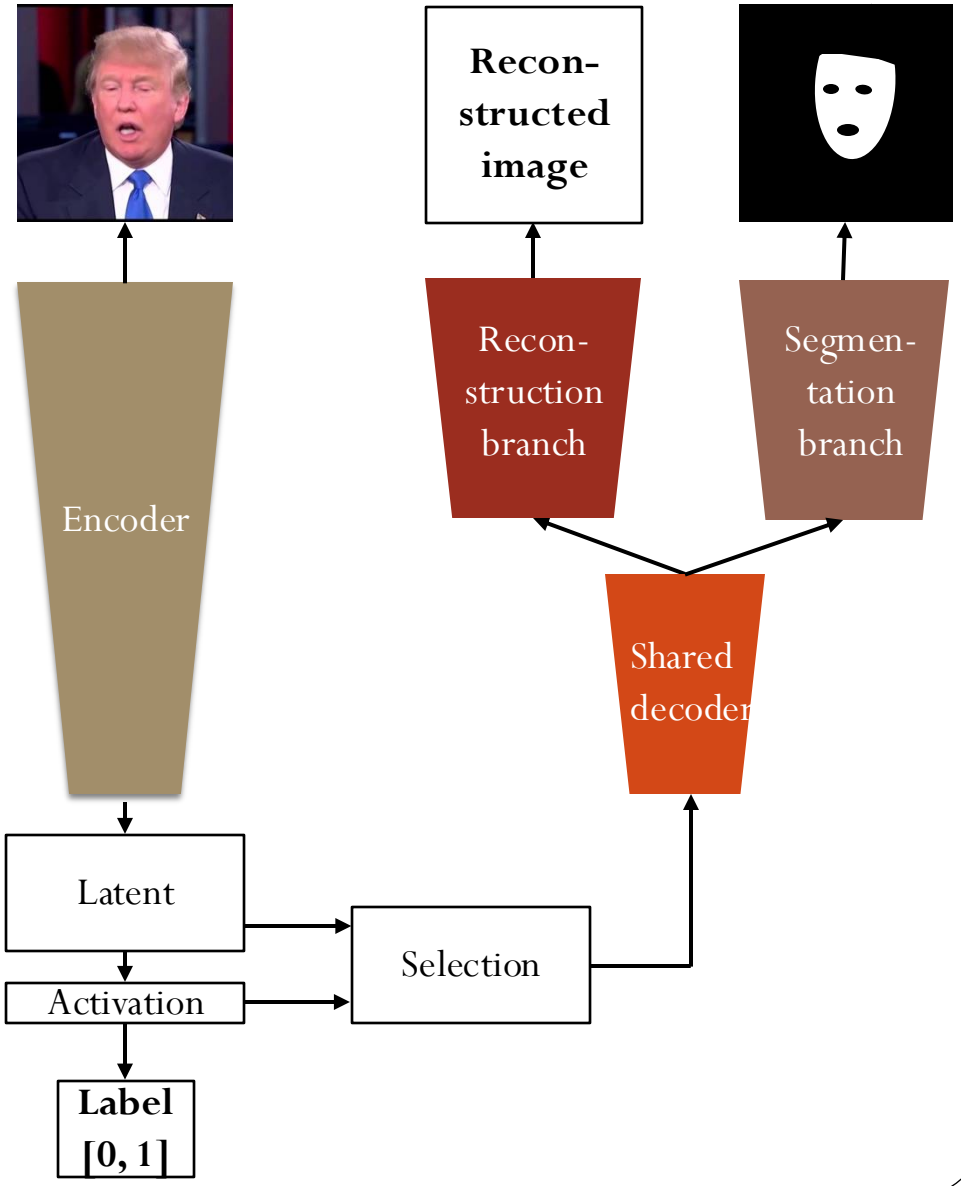


→ Infer the manipulated method

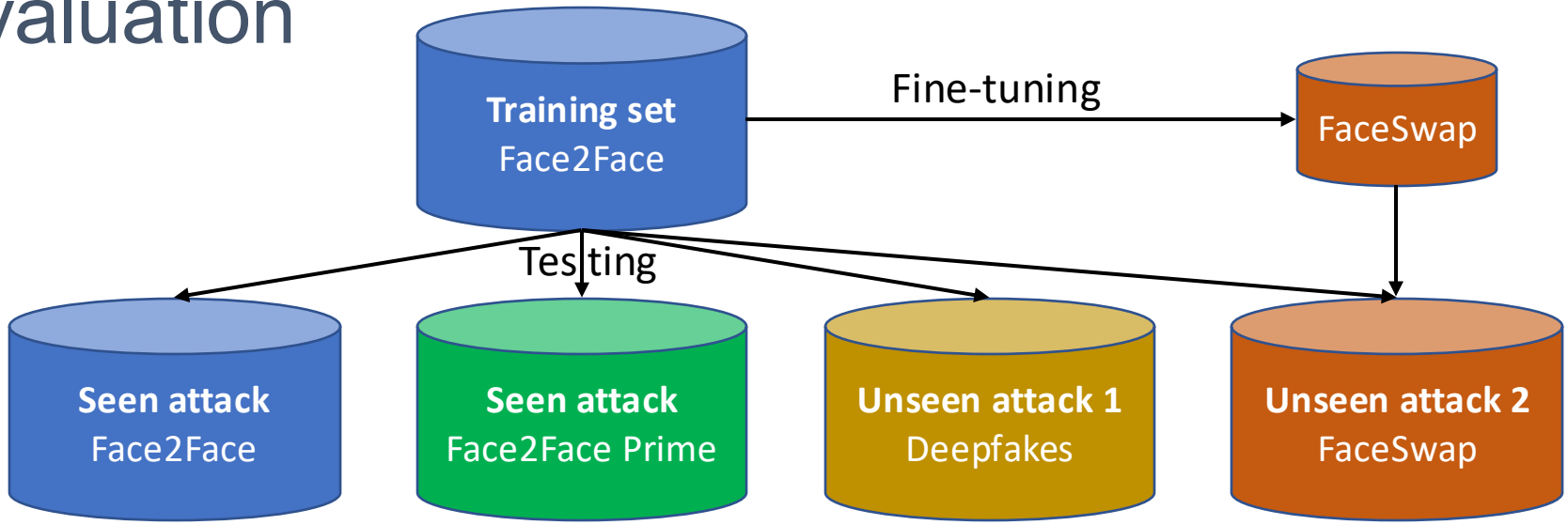
3. Proposed method



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



4. Evaluation



Type of attack	Classification EER (%)	Segmentation Acc. (%)
Match condition of seen attack	8.18	90.27
Mismatch condition of seen attack	8.07	90.20
Unseen attack 1 (without fine-tuning)	42.24	70.37
Unseen attack 2 (without fine-tuning)	34.04	84.67
Unseen attack 2 (fine-tuning on small data)	15.07	93.01

Thank you for your attention

Please come to my poster for more information and demo ☺