現実世界とサイバー空間の境界で生体情報を守るには

# Deepfake Detection and Segmentation

Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen

NII

## Motivation

Deepfake contents could be used to:

- Breaking authentication systems
- Impersonating people, creating fake news or porn videos.

→ Need to detect them and specify the manipulated regions.

## Objective

**Solving 3 problems simultaneously:**

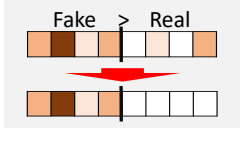
1. Identifying manipulated images/videos (real or fake → classification)
2. Specifying manipulated regions (segmentation)
3. Detecting unseen attacks (transferability / cross-database detection)

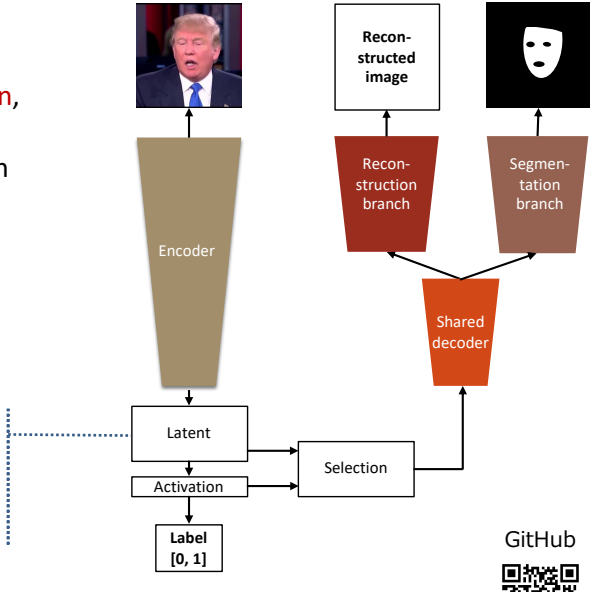
→ Heading toward **explainable AI**

## Methodology

Combining classification, segmentation, and image reconstruction in a single network

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

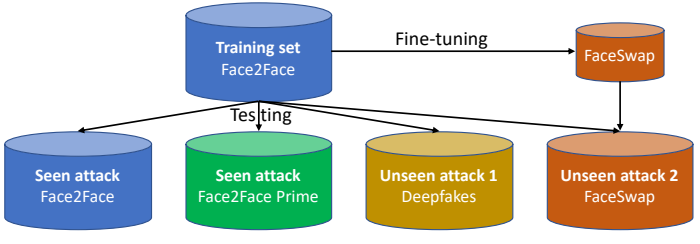



GitHub

Reference: **H. H. Nguyen**, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos," BTAS 2019.

## Results




| 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 |