[Deepfake Videos in the Wild : Analysis and Detection]

Intro & Background

- DF-W : collection of deep fake data in the wild
- DF-W ≠ dataset from research community [distribution / generation method
- evaluate detection schemes → work poorly on DF-W
  → existing detection schemes are not ready for real-world model deployment.
  → distributional difference [training / deployment
- integrated gradients
  → improve detection schemes
  → create more evasive deepfakes
- approaches to improve detection performance
  → transfer learning-based domain adaptation scheme

the source face : the face to be swapped in

the target face : the face that will be replaced.

Building DF-W Dataset

1. search & identify potential deepfake videos
2. Filter & download videos

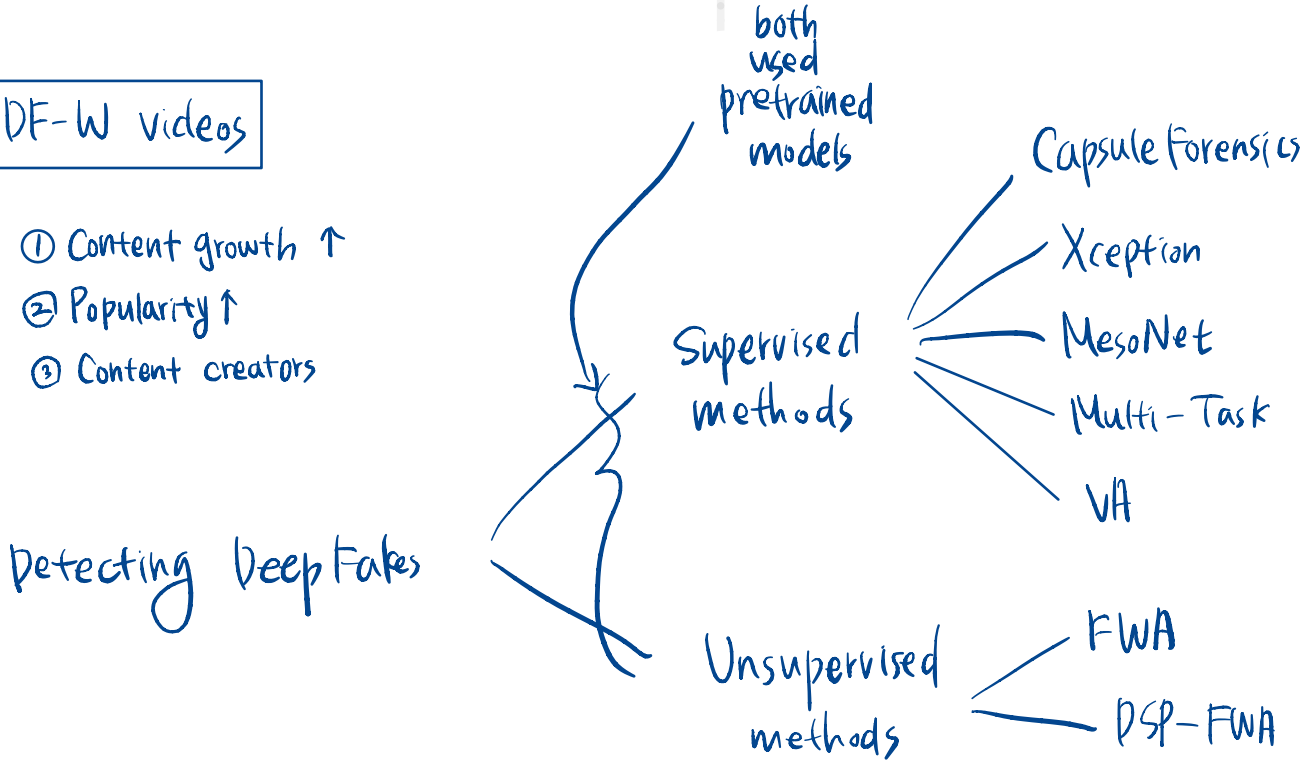
Research Community Datasets

DF-W Dataset

Research community	vs.	In the Wild
every frame contains fake content		only a fraction of the frames contains fake content
single face		multiple faces
shorter in duration		longer in duration

DF-W videos

① Content growth ↑
② Popularity ↑
③ Content creators



Performance of existing detection schemes

- fail to generalize to a variety of real-world deep fake videos.
- poor on DF-W (same lack of generalization capabilities)

Analyzing detection failures

- Racial bias : F1 score for Asian faces is considerably low.

Model Interpretation schemes

- Integrated Gradients → saliency map (attribution score of every pixel)

↓

a systematic, manual annotation process

< Annotation Process >

① Sampling representative images for explanation

② Using IntGrad to obtain saliency masks

③ Manually annotating saliency maps

Towards improving detection performance

- Improving detection results via transfer learning

  : Starting from the pre-trained weights, we retrain a given model using a limited amount of additional new data

Without freezing layers
- ① Domain adaptation retraining : ⊕ in-the-wild deepfake videos
- ② Source expansion retraining: ⊕ academic deepfake dataset

Results

① DA retraining ⇝ Promising

② CapsuleForensics > MesoNet

③ SE retraining ⇝✗ DF-W dataset

<Seferbekov model performance>

① X high detection performance

② DA or SE > seferbekov model

③ Seferbekov good on DF-R or DF-W