Practical Detection of Adversarial Face-Swap Deepfakes for Social Media Platforms

Jingming Dai 46272346, Supervisor: Associate Prof. Dan Kim

Background

What are Face-Swap deepfake videos?
A fake video that appears real and is created using face-swap AI techniques.

Applications of Deepfakes
Positive: Art & Film making, Education
Negative: Disinformation, Fake news, Identity theft

Deepfake Detection
- Content analysis methods
- Deep learning methods

Figure 1: Face-Swap Deepfake Image Examples

Proposed Work

Deepfake Video Dataset Selection
- Select 285,452 videos from DFDC (Deepfake Detection Challenge) and KoDF (Korean Deepfake Detection) datasets.

Image and Face Extraction
- 2 frames were extracted from each video.
- MTCNN extracted the face bounding box in each frame.

Multi-resolution and Augmentations
- Image Augmentations
  - Geometric & Colour, Noise, Compression, To-Grey
- Multi-resolution Training
  - 50%, 60%, 70%, 80%, 90% resolutions

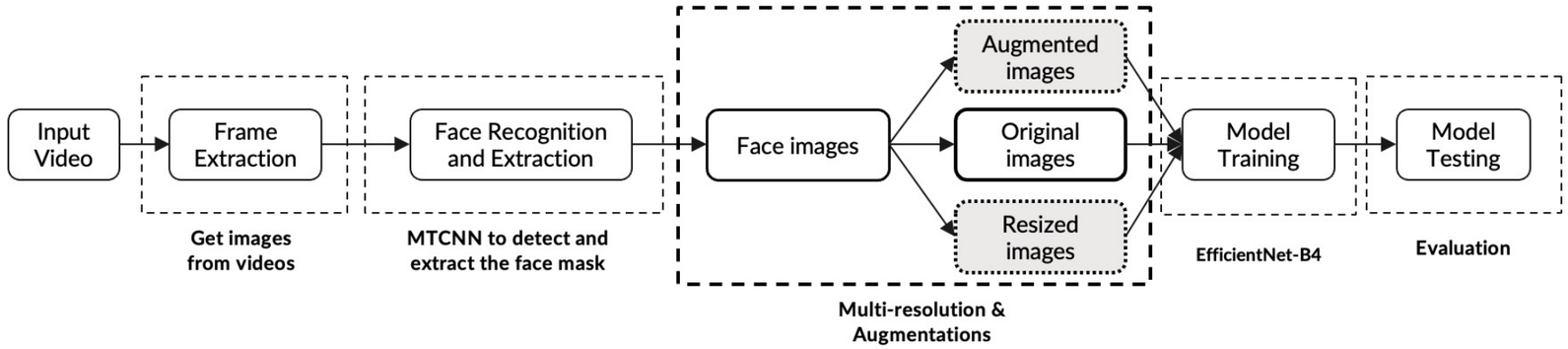

Figure 2: Overall Experiment Pipeline

Results

Multi-resolution & Image Augmentation Experiments

Environment: Intel i7-10700f CPU, 32GB RAM, RTX 2070s GPU

Reducing Training Samples
- Reduced images from 300,000 to 80,000
- Training time decreased from 25.5h to 6.5h
- Model learning slowed down after 40,000 images (performance remained similar beyond 40,000 images)

Multi-resolution Experiment
- 60% resolution maximizes model performance across resolutions
- 80% and 90% resolutions perform similarly to the baseline
- Models performed best at nearby resolutions (+-10%)

Image Augmentation Experiment
- Noise augmentation significantly improves model detection across augmented images
- Geometric & Colour and Compression augmentations perform similarly to baseline
- To-Grey augmentation decreased performance compared to the baseline

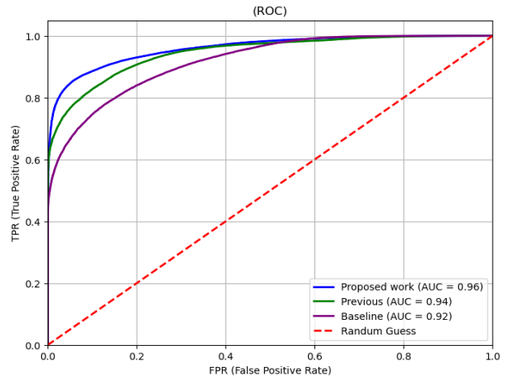

Figure 3: Comparison of AUC and ROC between the Combined Model, the Baseline Model, and the Previous Paper's Model

Integrated Experiment

- The combined model (60% resolution + noise augmentation) improves detection across resized and augmented images
- A 9% improvement in F1 score over the baseline model
- 8.5% higher F1 score compared to the random augmentation model

Conclusions

Training with multiple resolutions and image augmentation techniques improved the model's ability to detect images at different resolutions and under various augmentations.

Additional 60% of the original size combined with noise augmentation maximizes performance across resized and augmented images.