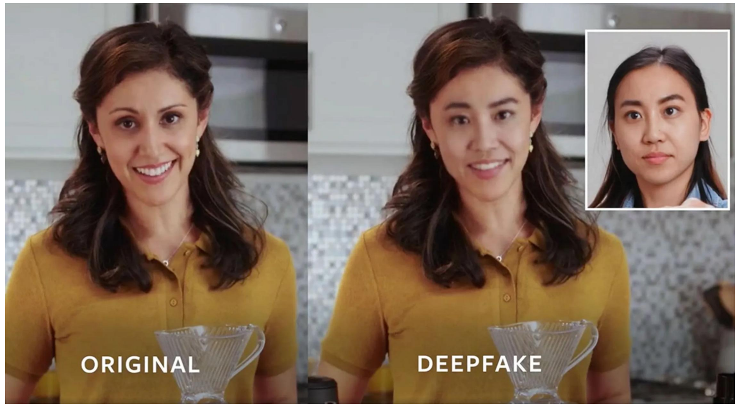
Research Introduction

On Developing Robust and Generalized DeepFakes Detection Algorithms

Xiaoman Lu

Ubiquitous & Visual Computing (UV) Lab, University of Warwick

➢ Introduction of Deepfake



Don't believe everything you see and hear in an internet video.

Deepfake = Deep Learning + Fake

Using **artificial intelligence methods (deep learning)** to generate **fake images** that closely resemble real effects.

The misuse of DeepFakes may lead to......

Fake News
Women's Safety Problem
Financial fraud
Political fraud
......

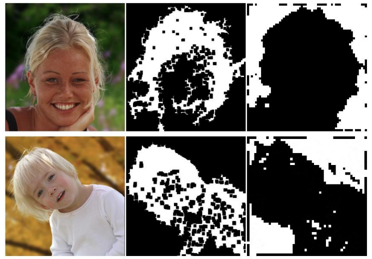
➡ It's crucial to develop effective deepfake detection methods

Background

Motivation

Overview

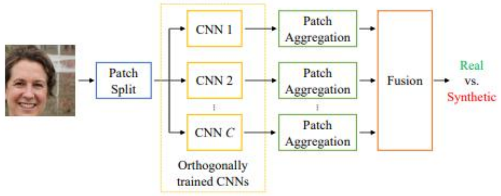
➢ Motivation of My Research


tampered image our method method of [16]

- **Traditional detectors**

➢ based on **intrinsic statistical information** like local noise

➢ highly dependent on the scene & Insufficient robustness in complex media environments



- **Deep Learning-based detectors**

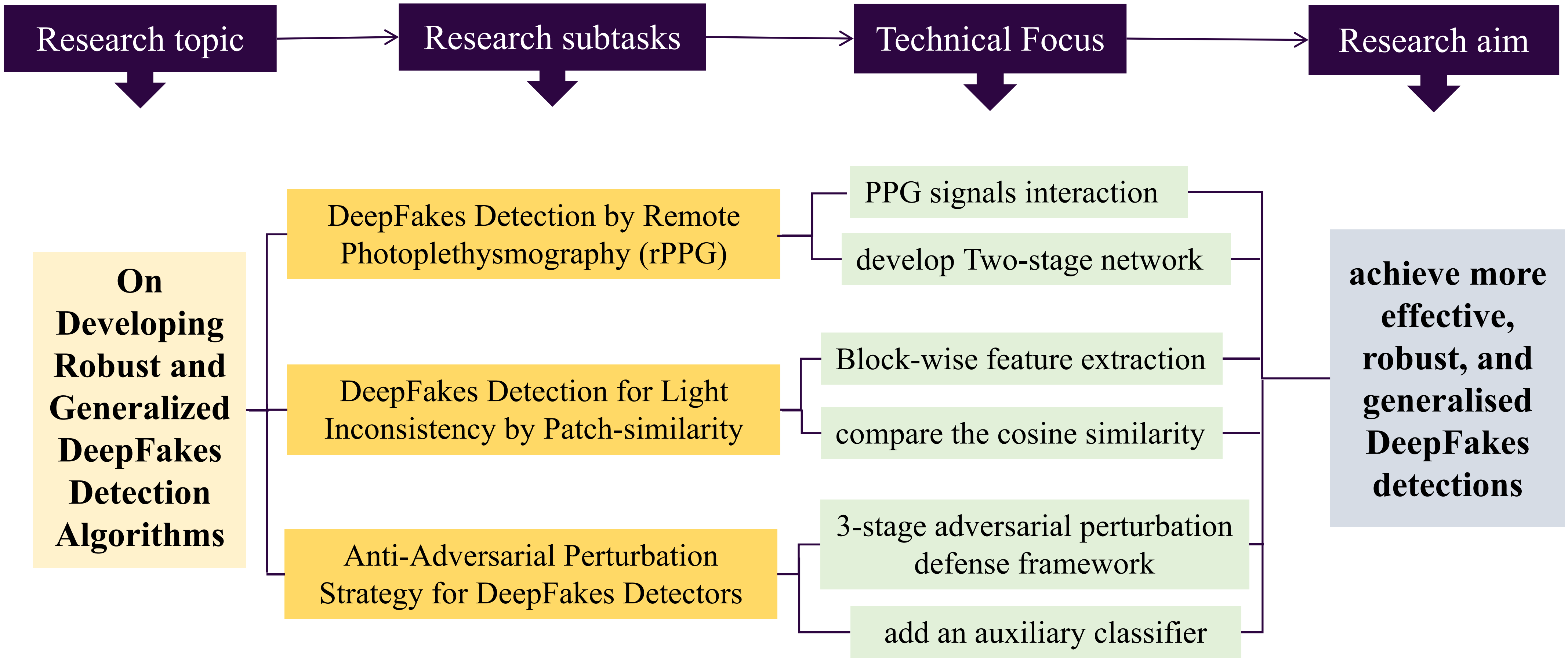
➢ based on spatial or frequency domain features

➢ detection performance sensitive to the **datasets** and **pre-trained model**

⇓

Let DeepFakes detectors acquire classification features for the low-level attributes of facial images, thus improving **generalization** and **robustness**

➢ Research Overview



Research topic → Research subtasks → Technical Focus → Research aim

On Developing Robust and Generalized DeepFakes Detection Algorithms

DeepFakes Detection by Remote Photoplethysmography (rPPG)

- PPG signals interaction
- develop Two-stage network

DeepFakes Detection for Light Inconsistency by Patch-similarity

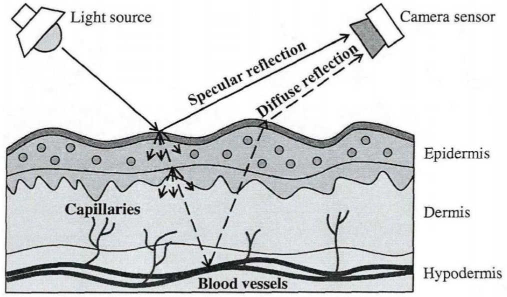
- Block-wise feature extraction
- compare the cosine similarity

Anti-Adversarial Perturbation Strategy for DeepFakes Detectors

- 3-stage adversarial perturbation defense framework
- add an auxiliary classifier

achieve more effective, robust, and generalised DeepFakes detections

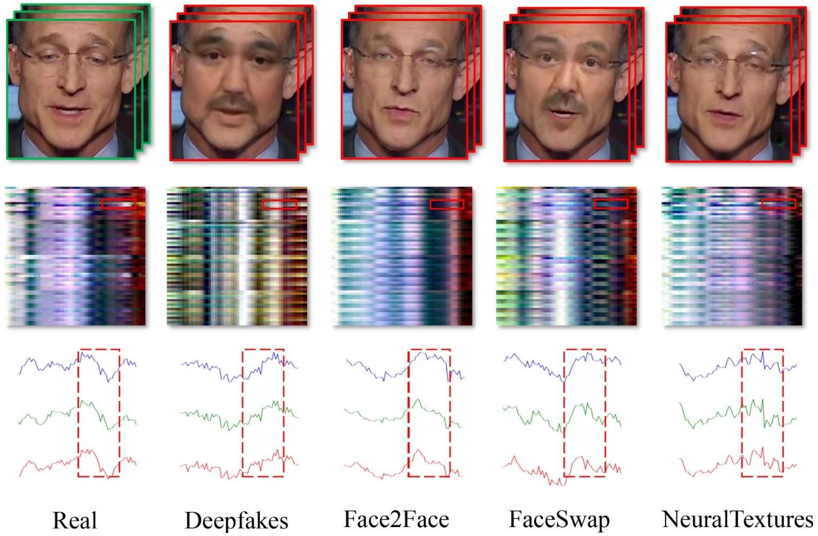
➢ DeepFakes Detection by Remote Photoplethysmography (rPPG)



rPPG Signal

Each heartbeat causes periodic changes in skin microvessels, resulting in a periodic signal of light reflection.

Since the rPPG signal is a biometric signal with unique information,
we can use this signals to detect deepfakes.



Background

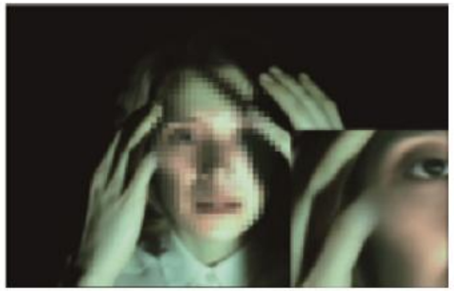
Motivation

Overview

➢ DeepFakes Detection for Light Inconsistency by Patch-similarity

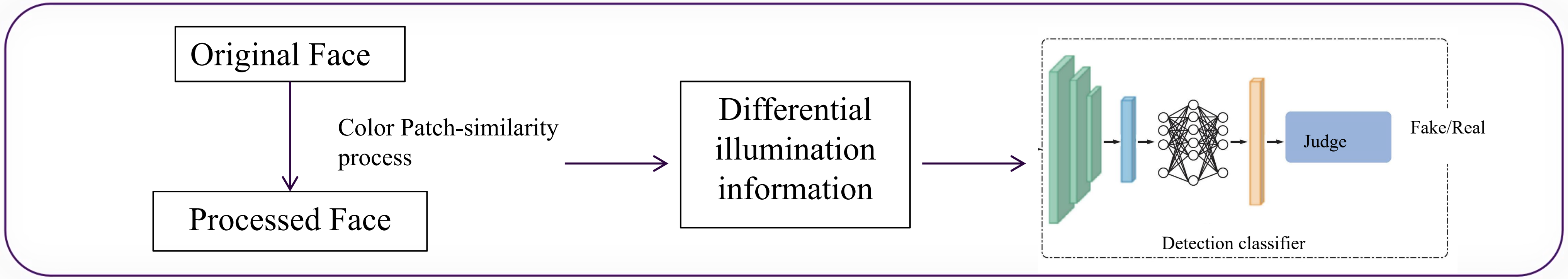
original

Target face



exchanged face

- ➢ deepfakes cannot handle the illumination refinement problem
- ➢ leaving an obvious **illumination inconsistency** between the exchanged face and the target face.



Original Face

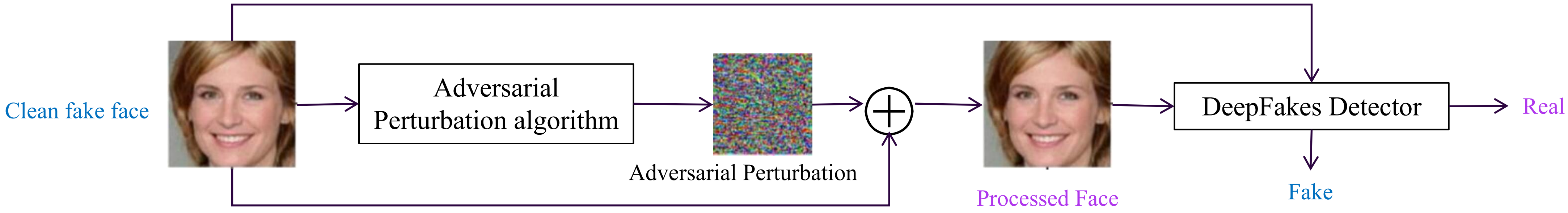
Color Patch-similarity process

Processed Face

Differential illumination information

Deepfake detecting network with illumination information

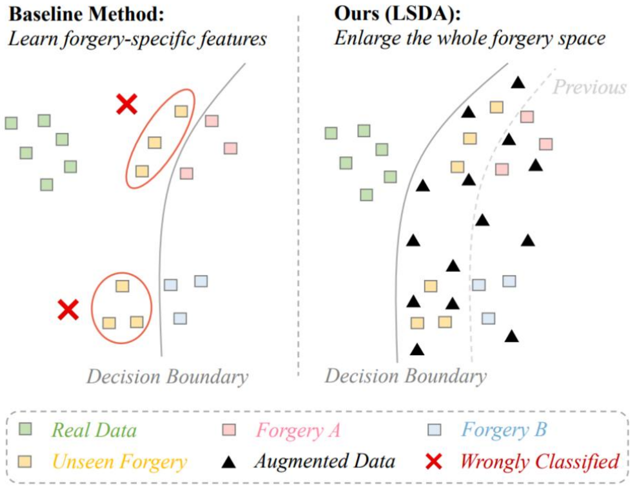
➢ Anti-Adversarial Perturbation Strategy for DeepFakes Detectors



Add adversarial perturbations to fake face images disables DeepFakes Detecor

Generalization Issue

Latent Space Augmentation



Enlarging the forgery space through **interpolating samples**

- ➢ encourages models to learn **a more robust decision boundary**
- ➢ helps **alleviate the forgery-specific overfitting**