DeepFake Detection Based on the Discrepancy Between the Face and its Context

# Yuval Nirkin; Lior Wolf; Yosi Keller et al.

## 2020

We propose a novel detection cue which utilizes the commonalities of all recent face identity manipulation methods

# Scholarcy Highlights

* We present a novel signal for identifying fake images based on comparing the inner face region – the one that is directly manipulated – with its outer context, which is left unaltered by all face manipulation methods we are aware of
* (3) The proposed approach is shown to outperform existing state-of-the-art schemes when applied to FaceForensics++ [5], Celeb-DF-v2 [17], and DeepFake Detection Challenge (DFDC) [18]
* (4) We show further results on two additional face swapping benchmarks, created using the FaceForensics++ data and additional swapping techniques, not included in FaceForensics++
* Deep neural networks are extensively used for face identification, and we focus on the contributions of two very specific facial regions, dictated by the desired application: the segmented face and its surrounding context
* Our results show that our method significantly outperforms the baseline on both unseen methods
* We propose a novel detection cue which utilizes the commonalities of all recent face identity manipulation methods

# Scholarcy Summary

## Introduction

P HOTOGRAPHY is widely perceived as offering authentic evidence of actual events, including, in particular, the presence and actions of human subjects in images and videos.

This perception is slowly shifting, contemporary technology allows far easier and more accessible manipulation of images than many realize.

The face of a person captured in a crime-scene can be manipulated and replaced by another.

Both of these examples are referred to as face swapping.

A third scenario involves the reenactment of a person’s face to change expression or lip motion

## Objectives

Our goal is to capture facial manipulation cues caused by face swapping, where the apparent identity is changed

## Methods

Two-stream [51] 70.1 Meso4 [8].

MesoInception4 [8] 83.0 HeadPose [50] FWA [42] DSP-FWA [42] VA-MLP [46].

XceptionNet-raw [5] 99.7 XceptionNet-c23 [5] 99.7 XceptionNet-c40 [5] 95.5 Multi-task [61]

## Results

We evaluated our proposed scheme using three recent, challenging benchmarks: FaceForensics++ [5], DFDC [18], and Celeb-DFv2 [17].

The recently announced, industry-backed, preview of the DFDC benchmark [18] offers a total of 5,244 videos of 66 actors: 4,464 training videos and 780 test videos, 1,131 of them are real videos and 4,113 are fakes generated by two different, unknown, face swapping methods.

The accuracy in each of the different categories, on its own, is not a direct indication of detection performance, since there is a threshold-dependent trade-off between the accuracy on real and fake images.

These results hint at the relative detection difficulty of each class and are provided for completeness.

We verify the accuracy of our proposed scheme in detecting fakes produced by methods that were not part of its training set

## Conclusion

While the ability to manipulate faces in images and video has increased dramatically in the last few years, all recent methods follow similar patterns.

We propose a novel detection cue which utilizes the commonalities of all recent face identity manipulation methods.

It is complementary to conventional real/fake classifiers and can be used alongside them

Overcoming this approach would require a much broader integration of the new identity into the image, making our contribution hard to circumvent without additional technological breakthroughs.

This is in contrast to artifact detection methods, which are susceptible to the constant progress in the visual quality of generated images.

It is our hope that by further analyzing the design principles of face swapping techniques, additional methods of identifying fake images and videos would be discovered, leading to effective mitigation of the societal risks of such media