WARWICK

DiffFake: Exposing Deepfakes using Differential Anomaly Detection

Sotirios Stamnas Victor Sanchez

University of Warwick

WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

Outline

WARWICK

1. Introduction
2. Related works
3. Our method: Differential Anomaly Detection

Introduction

Face Forgery in Videos

The two most common facial manipulations in videos. Left: Identity Swap, Right: Facial Reanactement [1]

Other common types of facial manipulations include

- Entire face synthesis
- Attribute manipulation

SOTA and human performance

WARWICK

Early SOTA methods deal with the problem as a binary classification problem

Method	DF	F2F	FS	NT
Human	77.6	49.6	76.1	32.3
XceptionNet	99.6	99.6	99.1	99.4

The in-dataset performance of SOTA XceptioNet vs human observers [1]

Training set	DF	F2F	FS	NT
DF	99.4	75.1	49.1	80.4
FS	70.1	61.7	99.4	68.7

The cross-manipulation generalization performance of XceptionNet [2]

Related works

Generalisation to unseen Deepfake generation methods

WARWICK

Most of the recent work on Deepfake detection focuses on improving the generalisation performance on the cross-manipulation and cross-dataset cases:

- Extraction of features from images in the frequency domain [3]
- Detection of irregularities in face: natural mouth movement [4]
- Use augmentation techniques to synthesize pseudo-deepfakes and train a binary classifier [5]
- Formulate the detection problem as an out of distribution anomaly detection (AD) task [6]

Our method: Differential Anomaly Detection

Overview

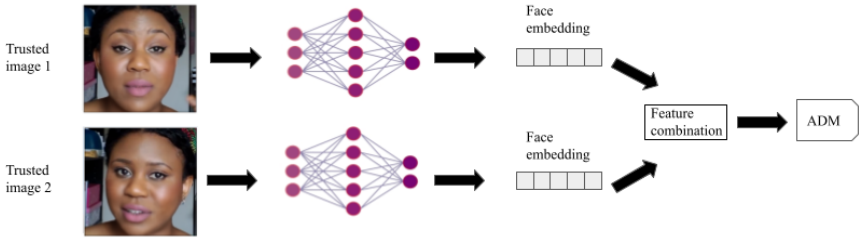
WARWICK

- Used in previous works [7] for identity attack detection
- Idea: Learn natural changes (i.e.change of head pose, illumination, face boundary consistency) that occur between two real images of the same subject
- Unnatural and extreme changes not observed in real images will be attributed to the presence of a manipulation method

First column: Real video, Second and third columns: Fake videos

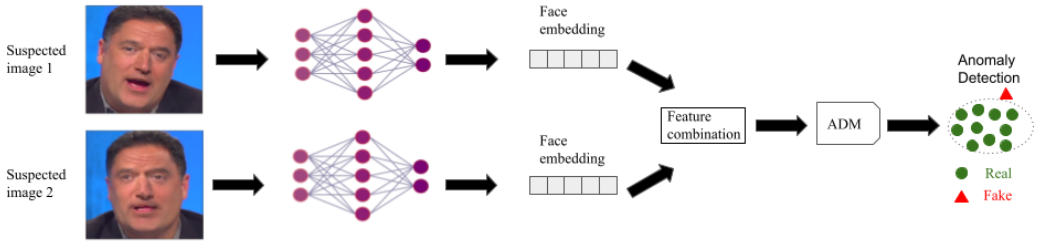
Training

- Idea: extract deep face embeddings from image pairs $h : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{d}$, and combine them to train an anomaly detection model (ADM)



Inference

- Idea: extract deep face embeddings from image pairs $h : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{d}$, and combine them to train an anomaly detection model (ADM)



Training the feature extractor with pseudo-deepfakes

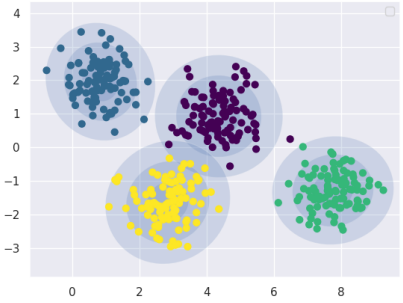
WARWICK

- Pseudo-deepfake: blend the face from a source image to a target image
  $\boldsymbol{I}_B = \boldsymbol{I}_s \odot \boldsymbol{M} + \boldsymbol{I}_t \odot (1 - \boldsymbol{M})$
- We introduce both global and local artifacts through four different types of mask generations
  1. Convex hull of all facial landmarks
  2. Convex hull of eye region landmarks
  3. Convex hull of lower jaw, mouth and nose apex landmarks
  4. Convex hull of entire jawline and nose tip landmarks
- More information in [5]

Anomaly Detection model: Gaussian Mixture Model (GMM)

WARWICK

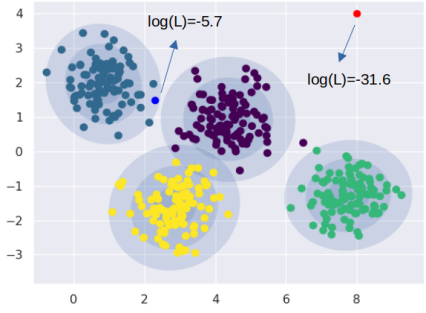
- A GMM models data points as a mixture of N Gaussian distributions $p(\boldsymbol{x}) = \sum_{i=1}^{N} \pi_i \mathcal{N}(\boldsymbol{X}|\mu_i, \Sigma_i)$
- It is effectively a clustering technique which assigns a probabilistic rather than a hard label to new datapoints
- In AD we can use the log likelihood of new datapoints as anomaly scores



Anomaly Detection model: Gaussian Mixture Model (GMM)

WARWICK

- A GMM models data points as a mixture of N Gaussians distributions $p(\boldsymbol{x}) = \sum_{i=1}^{N} \pi_i \mathcal{N}(\boldsymbol{X}|\mu_i, \Sigma_i)$
- It is effectively a clustering technique which assigns a probabilistic rather than a hard label to new datapoints
- In AD we can use the log likelihood of new datapoints as anomaly scores



Datasets

WARWICK

- Carried out extensive experiments using 6 different open-source deepfake datasets: FaceForensics++ (FF), Celeb-DF (CDF), DeeperForensics-1.0 (DF1.0), ForgeryNet (FNet), FaceShifter (FSh) and DeepFakeDetetion (DFD)

Dataset	Manipulation method
FaceForensics++	DF, FS, NT, F2F
Celeb-DF	Improved DeepFake
DeeperForensics-1.0	DF-VAE
ForgeryNet	8 different approaches
FaceShifter	AEI-Net + HEAR-Net
DeepFakeDetection	Unknown

Experimental setup

WARWICK

- Given two face embeddings A and B we consider the feature combinations

$$ABS = |A - B|, \;\; SUB = A - B, \;\; (SUB)^2 = (A - B)^2, \;\; (SUB)^3 = (A - B)^3$$

- We train an Efficientnet as our feature extractor using real images and pseudo-deepfakes
- The ADM (GMM with $k = 3$) is trained on pairs of only pristine images
- In all of our experiments, only real videos from FF++ are used for training
- At testing, pairs of suspected (real or fake) images from the same video are constructed
- A baseline AD model trained only on single frames is given for comparison

Results

WARWICK

Cross-manipulation evaluation: Test on different manipulations methods of FF++

Feature Comb.	Test Set AUC (%)				
	DF	F2F	FS	NT	Avg.
ABS	100	99.6	97.5	98.6	98.9
SUB	100	99.6	97.9	98.6	99.0
$(SUB)^2$	100	99.6	98.2	98.6	99.1
$(SUB)^3$	100	99.6	**98.6**	**98.9**	**99.3**

Performance of DiffFake with different feature combinations, under the cross-manipulation setting.

Method	Test Set AUC (%)				
	DF	F2F	FS	NT	Avg.
UNTAG [8]	–	–	–	–	81.8
OC-FakeDect2 [6]	88.4	71.2	86.1	97.5	85.8
Face X-ray [2]	99.2	98.6	98.2	98.1	98.5
PCL+I2G [9]	100	99.0	**99.9**	97.6	99.1
SBI† [5]	99.7	99.3	98.8	98.4	99.0
Baseline (ours)	99.6	99.3	96.8	98.2	98.5
DiffFake (ours)	100	**99.6**	98.6	**98.9**	**99.3**

Cross-manipulation evaluation results on FF++. DiffFake achieves the best performance on F2F and NT. Note that SBI† was re-evaluated using the official code.

Results

WARWICK

Cross-dataset evaluation: Test on CDF, DF1.0, FNet and FSh

Feature Comb.	Test Set AUC (%)				
	CDF	DF1.0	FNet	FSh	Avg.
ABS	74.5	87.8	80.9	90.7	83.5
SUB	75.1	89.8	80.0	92.4	84.3
$(SUB)^2$	**76.1**	91.0	**83.7**	**92.5**	**85.8**
$(SUB)^3$	75.7	91.0	83.0	91.1	85.2

Performance of DiffFake with different feature combinations, under the cross-dataset setting.

Method	Real Only	Test Set AUC(%)			
		CDF	DF1.0	FNet	FSh
Face X-ray [2]	Yes	74.8	-	-	-
SBI† [5]	Yes	**85.6**	83.3	82.2	94.0
SLAAD [10]	No	79.7	88.9	-	-
UNTAG [8]	Yes	74.7	-	77.0	-
Baseline (ours)	Yes	74.0	88.0	81.0	91.4
DiffFake (ours)	Yes	76.1	**91.0**	**83.7**	92.5

Cross-dataset evaluation results on various datasets. DiffFake achieves the best performance on DF1.0 and FNet.

Results

WARWICK

Cross-quality evaluation: Training and testing on compressed videos of FF++

Method	Test Set AUC (%)			
	c40		c23	
	DF	FS	DF	FS
Xception [1]	58.7	51.7	77.0	71.8
Face X-ray [2]	57.1	51.0	58.5	77.9
F3Net [3]	58.3	51.9	80.5	61.2
RFM [11]	55.8	51.6	79.8	63.9
SRM [12]	55.5	52.9	83.8	**79.5**
SLAAD [10]	62.8	56.8	84.6	72.1
Baseline (ours)	74.9	55.4	87.9	66.1
DiffFake (ours)	**78.5**	**58.2**	**89.3**	68.9

Cross-quality evaluation results on FS and DF. DiffFake achieves the best performance in three out of the four settings. Note that the results from all other methods are taken from [10].

Conclusions

WARWICK

- The main challenge in deepfake detection is the generalization performance across unseen generation methods
- We proposed a differential anomaly detection framework that leverages unnatural changes between frames of the same subject
- We proposed a pseudo-deepfake generation method that introduces both global and local artifacts through four different mask generation cases
- Our extensive experiments show that our method can match or exceed the performance of existing SOTA methods

Future work

WARWICK

- Test DiffFake on completely synthetic videos generated by Text-to-Video (T2V) and Image-to-Video (I2V) models
- Learn the best feature combination instead of defining it mathematically

Website:
https://warwick.ac.uk/fac/sci/dcs/research/siplab

Emails:
Sotirios.Stamnas@warwick.ac.uk
V.F.Sanchez-Silva@warwick.ac.uk

References I

[1] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics: A large-scale video dataset for forgery detection in human faces," *arXiv preprint arXiv:1803.09179*, 2018.

[2] L. Li, J. Bao, T. Zhang, *et al.*, "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5001–5010.

[3] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *European conference on computer vision*, Springer, 2020, pp. 86–103.

[4] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5039–5049.

[5] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 720–18 729.

[6] H. Khalid and S. S. Woo, "Oc-fakedect: Classifying deepfakes using one-class variational autoencoder," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 656–657.

[7] M. Ibsen, L. J. González-Soler, C. Rathgeb, P. Drozdowski, M. Gomez-Barrero, and C. Busch, "Differential anomaly detection for facial images," in *2021 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, 2021, pp. 1–6.

References II

[8] N. Mejri, E. Ghorbel, and D. Aouada, "Untag: Learning generic features for unsupervised type-agnostic deepfake detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[9] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 023–15 033.

[10] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 710–18 719.

[11] C. Wang and W. Deng, "Representative forgery mining for fake face detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 923–14 932.

[12] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 317–16 326.