https://doi.org/10.2352/EI.2024.36.4.MWSF-333

Improving Video Deepfake Detection: A DCT-Based Approach with Patch-Level Analysis

Luca Guarnera, Salvatore Manganello, Sebastiano Battiato; Department of Mathematics and Computer Science, University of Catania, Italy

Abstract

A new algorithm for the detection of deepfakes in digital videos is presented. The I-frames were extracted in order to provide faster computation and analysis than approaches described in the literature. To identify the discriminating regions within individual video frames, the entire frame, background, face, eyes, nose, mouth, and face frame were analyzed separately. From the Discrete Cosine Transform (DCT), the β components were extracted from the AC coefficients and used as input to standard classifiers. Experimental results show that the eye and mouth regions are those most discriminative and able to determine the nature of the video under analysis.

Introduction

The rapid development of the deepfake technology poses significant challenges to the authenticity and trustworthiness of multimedia content [2]. Deepfakes are synthetic creations that manipulate or generate content using advanced generative models, making it increasingly difficult to differentiate them from genuine recordings. This has raised concerns regarding the potential misuse of deepfakes for malicious purposes, such as disinformation campaigns, identity theft, or defamation. The rapid advancement of deep learning techniques, coupled with the availability of vast amounts of data, has led to increasingly sophisticated deepfake algorithms.

The Deepfake phenomenon first emerged in 2017 on the website "Reddit", when an anonymous user named "deepfakes" uploaded a pornographic video that superimposed a celebrity's face onto another person's body. This event sparked the interest of other users, who began creating similar videos by replacing the faces of different celebrities. Research indicates that approximately 96% of Deepfake videos are pornographic in nature, while the remaining 4% cover various other genres. Deepfake technology primarily focuses on manipulating faces, as facial recognition plays a crucial role in identifying individuals. As a result, four main face manipulation modalities are employed: generation of entire synthetic faces, facial attribute manipulation, identity swapping, and expression swapping.

Examples of deepfake videos have demonstrated the potential for creating realistic yet fabricated scenarios. For instance, politicians have been depicted giving speeches they never delivered, celebrities have been inserted into explicit scenes, and individuals' faces have been swapped onto others in various contexts. Examples include the "Synthesizing Obama" project [1], which is a fake video showing former US President Barack Obama and

Figure 1. *Videos in the Faceforensics++ dataset manipulated with respect to the techniques of (a) Faceswap, (b) Face2Face, (c) Face Shifter, (d) DeepFakes, (e) DeepFake Detection, (f) Neural textures.*

American director Jordan Peele saying words and phrases they have never said using a technique called lip-syncing. The face of Nicolas Cage has also been put in films where the same actor did not act, such as "Matrix" or "Fight Club", as well as in music videos such as "Never Gonna Give You Up" [2]. These examples highlight the potential consequences of deepfake technology, including its potential to deceive, manipulate public opinion, and compromise personal privacy.

To address the risks associated with deepfakes, the development of reliable and efficient deepfake detection methods has become necessary. Researchers have explored various approaches, including deep learning models [23, 13, 4, 14, 22], statistical analysis [7, 8, 17, 6], and forensic [9] techniques, to detect the presence of deepfakes in multimedia content. However, the constant evolution of deepfake algorithms necessitates ongoing research and innovation in this field. In this study, we propose a novel deepfake detection algorithm that exploits the Discrete Cosine Transform (DCT) to extract discriminative features from video frames. By analyzing the β components derived from the AC coefficients, we aim to identify the most informative frequencies for differentiating between real and deepfake videos. Our methodology focuses on both accuracy and computational efficiency, with the goal of enabling real-time deepfake detection in practical scenarios. Moreover, to enhance the efficacy of our algorithm, we investigated the discriminative patches within individual frames. By analyzing I-frames, we aimed to identify specific regions, such as the eyes and mouth, that exhibit distinctive characteristics when

[1] `https://www.youtube.com/watch?v=cQ54GDm1eL0`, last accessed 10/03/2023.

[2] `https://www.youtube.com/watch?v=4soZciRrZRI`, last accessed 13/03/2023.

comparing real and deepfake videos. This patch-level analysis further enhances the explainability and reliability of our deepfake detection system.

To evaluate the performance of our algorithm, we conducted extensive experiments on widely used deepfake datasets, including Faceforensics++ and Celeb-DF (v2). Our results demonstrate the effectiveness of the proposed approach in accurately classifying real and deepfake videos. By presenting our methodology, experimental results, and analysis, this research contributes to the ongoing efforts in combating deepfakes. Our study aims to advance the state-of-the-art in deepfake detection, with a focus on speed, interpretability, and accuracy. The insights gained from this work can help mitigate the potential harmful effects of deepfakes and contribute to the development of robust defense mechanisms against their malicious use.

Related Works

Deepfake video detection has become a crucial research area due to the increasing prevalence and potential harm associated with manipulated videos. Several methods have been proposed to tackle this challenge. An overview on Media forensics with a particular focus on Deepfakes has been proposed in [21, 11, 1].

Various existing methods have been developed for Deepfake detection, aiming to distinguish genuine multimedia content from Deepfakes. Many approaches focus on identifying traces left by generative models during the creation of Deepfakes. Matern et al. [16] exploit the presence of artifacts, such as differently colored eyes or differently shaped ears, which are often produced by generative models. Their method involves training two classifiers that extract features to describe these anomalies. However, this technique is limited to synthetic faces with open eyes or visible teeth.

Recent studies have shown that artifacts can also be detected using Convolutional Neural Networks (CNNs) [22, 10]. Nguyen et al. [18] propose a multi-task learning approach, training a CNN for both manipulation detection and segmentation of manipulated areas. Nirkin et al. [19] introduce a CNN-based method that analyzes the face and its relationship with the surrounding context, comparing the extracted features to identify any discrepancies. Haliassos et al. [12] present the LipForensics method, which leverages CNNs to identify irregularities in lip movements commonly found in synthetic videos. They extract lip features using a CNN and compare them with the rest of the face. Zhang et al. [24] propose the use of a 3-Dimensional Convolutional Neural Network (3DCNN) to capture spatio-temporal information from videos, enabling differentiation between original and Deepfake videos.

Zheng et al. [25] propose a two-phase method for Deepfake detection. In the first phase, they introduce a novel architecture called the Fully Temporal Convolution Network (FTCN), which reduces the spatial convolutional matrix dimension to 1 while maintaining the temporal convolutional matrix dimension. In the second phase, a Temporal Transformer network is employed to verify long-term temporal coherence. Ge et al. [5] propose the Latent Pattern Sensing (LPS) model, which captures semantic change features for Deepfake video detection. LPS involves an analyzer to detect faces, an encoder to extract spatial semantic features, an aggregator to obtain spatiotemporal features, and a classifier to distinguish real videos from Deepfake videos. Also, two additional models, namely LPS*ri* and LPS*ssd*, have been introduced.

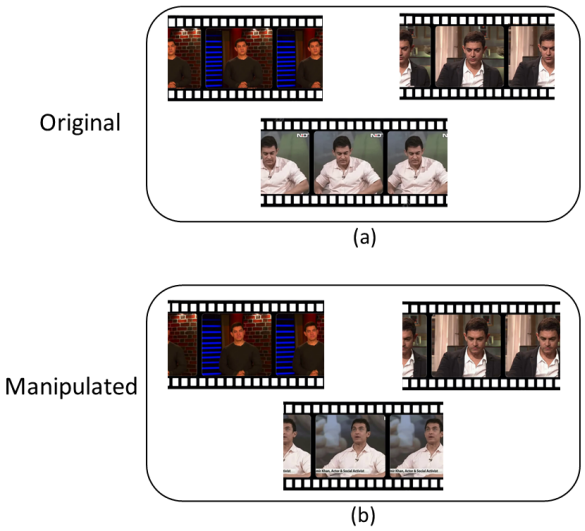

Figure 2. *Examples of real (a) and manipulated (b) videos in the Celeb-DF dataset (v2).*

Dataset	Type		#Videos
Faceforensics++	Manipulated Sequence	DeepFakeDetection	3084
		Deepfakes	1001
		Face2Face	1001
		NeuralTextures	1002
		FaceSwap	1000
		FaceShifter	1001
	Original Sequance	Actor	364
		Youtube	1003
Celeb-DF (v2)	Original	Celeb-real	590
		Youtube	300
	Manipulated	Celeb-synthesis	5639

Tab.1: Number of videos present in each folder of the utilized dataset.

Dataset

The experiments were performed using two different types of Deepfake video datasets: *Faceforensics++* [20] and *Celeb-DF (v2)* [15]. Both datasets consist of a combination of real and synthetic videos. The real videos included in both Celeb-DF (v2) and Faceforensics++ were sourced from the internet, primarily from platforms like YouTube, or captured using real actors. The synthetic videos in Faceforensics++ were generated using two types of manipulations: computer graphics-based approaches such as Face2Face, FaceSwap, and learning-based approaches like Deepfakes and NeuralTextures. Figure 1 shows some examples of video frames of the FaceForensics++ dataset. The Celeb-DF (v2) dataset contains synthetic videos created using publicly available algorithms specifically designed for Deepfake generation. Figure 2 shows some examples. Table 1 provides an overview of the number of videos included in each dataset.

In our study, we wanted to conduct a thorough evaluation of deepfake detection methods using a variety of standard classifiers (e.g., k-NN, SVM and others). In the experiments carried out, we trained all classifiers considering both datasets just described. Specifically, that dataset was divided into Training (50%), Validation (20%), and Test (30%), ensuring that the classes of real that

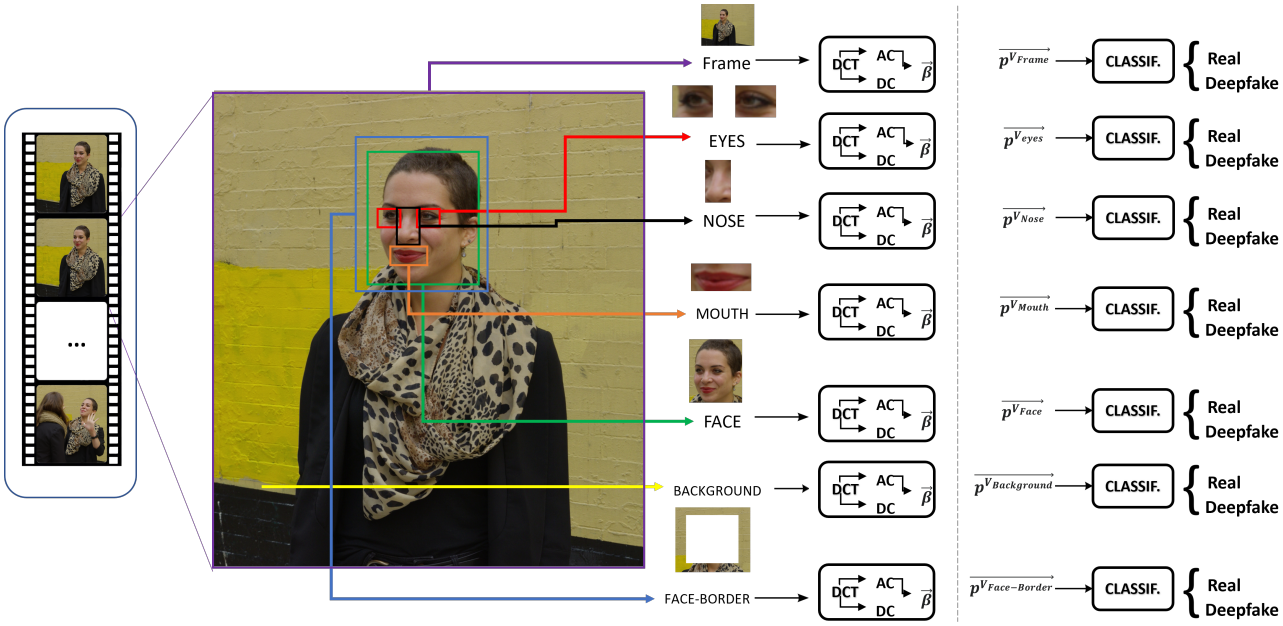

Figure 3. *Proposed approach: (a) For each patch in A of the I-frames of the video V, the DCT is calculated and the β components of the 63 AC coefficients are extracted. (b) The final feature vectors p^{V_a} of video V, are used in the various classifiers to solve the Real Vs Deepfake task and identify the most discriminative regions.*

of deepfakes were balanced in terms of the number of videos in each category.

Overall, the inclusion of these diverse datasets allows us to evaluate the effectiveness and robustness of our proposed deepfake detection methods across a wide range of scenarios and manipulation techniques.

Proposed Approach

The proposed pipeline in this paper consists of working on frames of type I (I-frame) of the datasets under analysis. The main objective was to search for anamalies in the Discrete Cosine Transform (DCT) domain by analysis of the β parameters extracted from the 63-AC coefficients, which were found to be highly discriminative in distinguishing Deepfake images from real ones [3, 6]. In order to achieve a fast method able to identify the most discriminative area in a video frame, various regions were analyzed separately:

$$A = \{\textit{entire frame, face, face contour, eyes, nose, mouth, background}\}$$

Each region of interest in each frame was initially converted to grayscale and then divided into 8x8 blocks. DCT was applied to each block by considering the Formula 1.

$$F_{u,v} = \frac{2}{N}\left[\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} C(u)C(v)f(x,y) \cos\frac{(2x+1)v\pi}{2*N}\cos\frac{(2y+1)v\pi}{2*N}\right] \quad (1)$$

where $N = 8$, $C(u)$ and $C(v)$ are defined as follows:

$$C(u) = \begin{cases} \frac{1}{\sqrt{2}} & u = 0 \\ 1 & u > 0 \end{cases}, C(v) = \begin{cases} \frac{1}{\sqrt{2}} & v = 0 \\ 1 & v > 0 \end{cases} \quad (2)$$

DCT coefficients are generated and ordered in zig-zag order from the top-left element to the bottom-right element. From the DCT, it is possible to analyze the DC and AC coefficients, where the DC coefficient represents the average luminance value and is considered a redundant value. All other coefficients are AC and represent specific frequency bands. The DC coefficient follows a Gaussian distribution while the AC coefficients follow a Laplacian distribution centered at 0, according to the following Formula 3:

$$P(x) = \frac{1}{2\beta} exp\left(\frac{-|x-\mu|}{\beta}\right) \quad (3)$$

where $\mu = 0$, $\beta = \sigma/\sqrt{2}$ and σ corresponds to the standard deviation of the AC coefficient.

Given a generic video V, the following feature vector was then created for each patch p_i ($i = 1,...,k$, k = *total number of patches extracted in video V*) of the area of interest $a \in A$:

$$p_i^{V_a} = \{\beta_1, \beta_2, \ldots, \beta_{63}\} \quad (4)$$

In order to obtain a unique descriptor $\overrightarrow{p^{V_a}}$ of a for video V, the average of each β component was computed, thus obtaining the following feature vector:

$$p^{V_a} = \{\bar{\beta_1}, \bar{\beta_2}, \ldots, \bar{\beta_{63}}\} \quad (5)$$

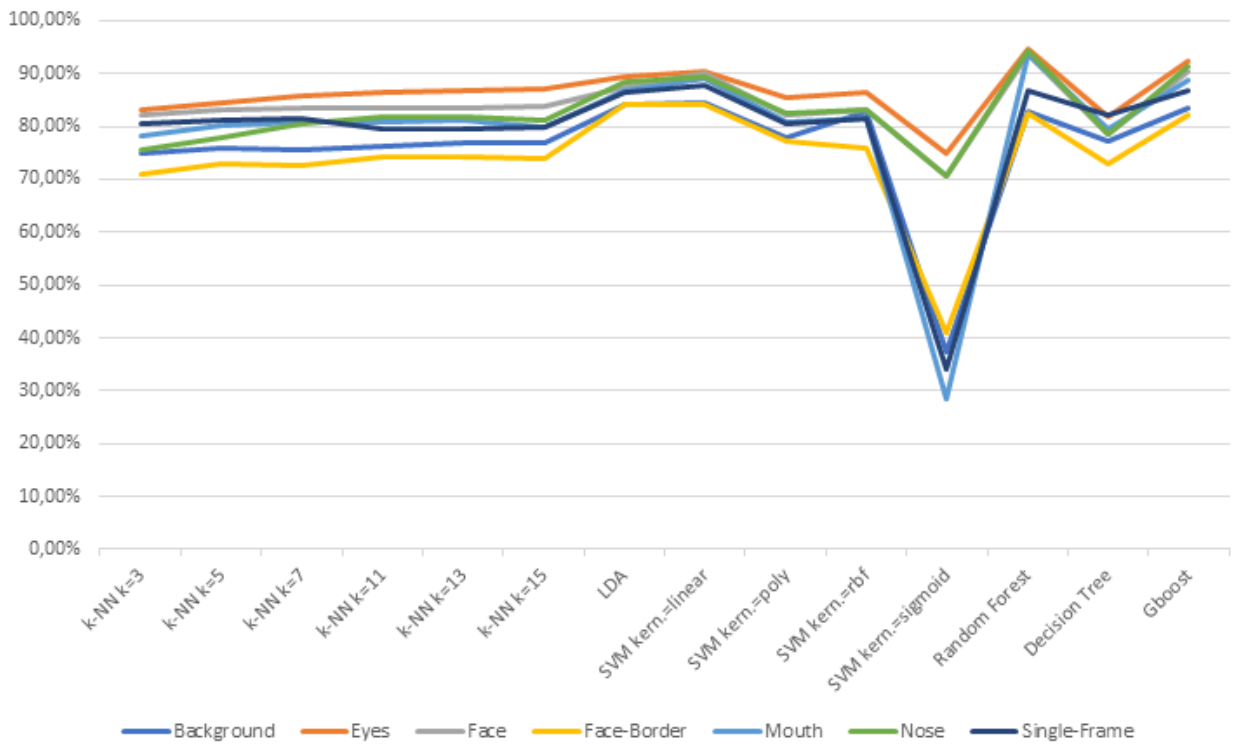

Figure 4. *Values of AUC metric (%) across classifiers and regions under analysis.*

Different classifiers were trained on each a area, giving feature $\overrightarrow{p^{V_a}}$ as input in order to identify the most discriminative region capable of solving the task in question. The classifiers used are:

- k-NN with k = {3, 5, 7, 11, 13, 15};
- Linear Discriminative Analysis (LDA);
- Support Vecotr Machine (SVM) with kernel = {linear, poly, RBF, sigmoid}
- Random Forest;
- Decision Tree;
- GBoost.

Figure 3 summarizes the proposed approach.

The proposed method was implemented using the Python language, OpenCV and the Pytorch framework.

Experimental Results and Comparison

In the experiments carried out, all feature vectors $\overrightarrow{p^{V_a}}$ (described in the previous section) were collected and given as input to the various classifiers. As an evaluation metric, we chose AUC (Area Under the Curve).

Figure 4 shows the results obtained for each classifier with respect to the various areas A under analysis. It can be seen from the Figure that the best results in the task in question were obtained using the Random Forest classifier. Specifically, the most discriminative areas in both datasets turn out to be the eyes with an AUC value equal to **94.55%** and the nose with an AUC value of **94.28%**. The least discriminative regions turn out to be the facial border and the background.

Methods	AUC (%)
Visual Artifacts [16]	66,55
Multi-task [18]	66,3
Face+Context [19]	70,5
LipForensics [12]	89,75
TD-3DCNN [24]	80,52
FTCN [25]	93.3
Our	**94,55**

Tab.2: AUC Score (%) Comparisons with State-of-the-Art Approaches on FaceForensics++, and Celeb-DF(v2) Datasets.

In general, we can therefore affirm that by exploiting the β values extracted from the AC coefficients from the nose and eyes and analyzing only the type I frames, the proposed method is able to discriminate well and solve the task in question, despite not being a deep learning approach.

The best results obtained from the proposed approach were compared with state-of-the-art methods, many of which are based on deep learning algorithms. The AUC matric was calculated for each method. Table 2 shows the obtained results. It can be seen from the obtained results that, the proposed method outperforms, albeit slightly, the state of the art. It should be highlighted, however, that the method we propose is fully analytical, explicable and fast in execution, compared to the various works in the literature. In addition, under these aspects, it is possible to use this method for real-time applications in order to counter the illicit use

of the powerful deepfake technology.

Conclusion

In this study, we have proposed a novel deepfake detection algorithm that leverages the Discrete Cosine Transform (DCT) to extract discriminative features from video frames. By analyzing the β components derived from the AC coefficients, we have identified informative frequencies that allow us to differentiate between real and deepfake videos effectively. Our approach prioritizes both accuracy and computational efficiency, aiming to enable real-time deepfake detection in practical scenarios.

To further enhance the efficacy of our algorithm, we have investigated discriminative patches within individual frames, focusing on regions like the eyes and mouth. Our analysis of these specific areas, particularly in I-frames, has revealed distinctive characteristics that differentiate real videos from deepfake ones. This patch-level analysis enhances the interpretability and reliability of our deepfake detection system. Through extensive experiments on well-established deepfake datasets, such as Faceforensics++ and Celeb-DF (v2), we have demonstrated the effectiveness of our proposed approach in accurately classifying real and deepfake videos. By presenting our methodology, experimental findings, and analysis, this research significantly contributes to the field of deepfake detection. Our focus on speed, interpretability, and accuracy aims to advance the state-of-the-art in deepfake detection techniques. The insights gained from this study can aid in mitigating the potential harmful effects of deepfakes and contribute to the development of robust defense mechanisms against their malicious use. While our proposed algorithm shows promising results, the constant evolution of deepfake algorithms calls for continued research and innovation in this domain. Future work should explore additional features and leverage advancements in machine learning and computer vision to further enhance the detection accuracy and efficiency of deepfake detection systems.

Acknowledgments

This research is supported by Azione IV.4 - “Dottorati e contratti di ricerca su tematiche dell’innovazione” del nuovo Asse IV del PON Ricerca e Innovazione 2014-2020 “Istruzione e ricerca per il recupero - REACT-EU”- CUP: E65F21002580005.

References

[1] Deepfake Generation and Detection, a Survey, author=Zhang, Tao, journal=Multimedia Tools and Applications, pages=1–18, year=2022, publisher=Springer.

[2] Sebastiano Battiato, Oliver Giudice, and Antonino Paratore. Multimedia Forensics: Discovering the History of Multimedia Contents. In *Proceedings of the 17th International Conference on Computer Systems and Technologies*, pages 5–16, 2016.

[3] Sara Concas, Gianpaolo Perelli, Gian Luca Marcialis, and Giovanni Puglisi. Tensor-Based Deepfake Detection In Scaled And Compressed Images. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3121–3125. IEEE, 2022.

[4] Apurva Gandhi and Shomik Jain. Adversarial Perturbations Fool Deepfake Detectors. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[5] Shiming Ge, Fanzhao Lin, Chenyu Li, Daichi Zhang, Weiping Wang, and Dan Zeng. Deepfake Video Detection via Predictive Representation Learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2s):1–21, 2022.

[6] Oliver Giudice, Luca Guarnera, and Sebastiano Battiato. Fighting Deepfakes by Detecting GAN DCT Anomalies. *Journal of Imaging*, 7(8), 2021.

[7] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. DeepFake Detection by Analyzing Convolutional Traces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 666–667, 2020.

[8] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Fighting Deepfake by Exposing the Convolutional Traces on Images. *IEEE Access*, 8:165085–165098, 2020.

[9] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake Style Transfer Mixture: a First Forensic Ballistics Study on Synthetic Images. In *Image Analysis and Processing–ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part II*, pages 151–163. Springer, 2022.

[10] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Level Up the Deepfake Detection: a Method to Effectively Discriminate Images Generated by GAN Architectures and Diffusion Models. *arXiv preprint arXiv:2303.00608*, 2023.

[11] Luca Guarnera, Oliver Giudice, Cristina Nastasi, and Sebastiano Battiato. Preliminary forensics Analysis of DeepFake Images. In *2020 AEIT International Annual Conference (AEIT)*, pages 1–6. IEEE, 2020.

[12] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja 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*, pages 5039–5049, 2021.

[13] Chih-Chung Hsu, Yi-Xiu Zhuang, and Chia-Yen Lee. Deep Fake Image Detection Based on Pairwise Learning. *Applied Sciences*, 10(1), 2020.

[14] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face X-ray for More General Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020.

[15] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A Large-Scale Challenging Dataset for Deepfake Forensics. pages 3207–3216, 2020.

[16] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92. IEEE, 2019.

[17] Scott McCloskey and Michael Albright. Detecting GAN-generated Imagery using Saturation Cues. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4584–4588. IEEE, 2019.

[18] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-Task Learning for Detecting and Segmenting Manipulated Facial Images and Videos. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8. IEEE, 2019.

[19] Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. Deepfake Detection Based on Discrepancies Between Faces and Their Context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6111–6121, 2021.

[20] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–

11, 2019.

[21] Luisa Verdoliva. Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020.

[22] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020.

[23] Xinsheng Xuan, Bo Peng, Wei Wang, and Jing Dong. On the Generalization of GAN Image Forensics. In *Chinese Conference on Biometric Recognition*, pages 134–141. Springer, 2019.

[24] Daichi Zhang, Chenyu Li, Fanzhao Lin, Dan Zeng, and Shiming Ge. Detecting Deepfake Videos with Temporal Dropout 3DCNN. In *IJCAI*, pages 1288–1294, 2021.

[25] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring Temporal Coherence for More General Video Face Forgery Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15044–15054, 2021.

Author Biography

Luca Guarnera was born in Catania on October 26, 1992. Since January 1, 2022, he is a research fellow in Computer Science at the University of Catania. He graduated as Ph.D. in Computer Science (XXXIII cycle, PON number E37H18000330006) on October 14, 2021, discussing the thesis entitled "Discovering Fingerprints for Deepfake Detection and Multimedia-Enhanced Forensic Investigations" at the Department of Mathematics and Computer Science, University of Catania. His main research interests are Computer Vision, Machine Learning, Multimedia Forensics and its related fields with a focus on the Deepfake phenomenon.

Salvatore Manganello was born in San Cataldo on October 14, 1998. He completed her bachelor's degree in Computer Science at the University of Catania. Her primary research areas focus on the forensic study of video deepfake detection.

Sebastiano Battiato is a full professor of Computer Science at the University of Catania. He received his degree in Computer Science (summa cum laude) in 1995 from the University of Catania and his Ph.D. in Computer Science and Applied Mathematics from the University of Naples in 1999. He has been Chairman of the Undergraduate Program in Computer Science (2012-2017), and Rector's delegate for Education: postgraduates and Phd (2013-2016). He is currently the Scientific Coordinator of the PhD Program in Computer Science (XXXIII-XXXVI cycles) and Deputy Rector for Strategic Planning and Information Systems at the University of Catania. His research interests include Computer Vision, Imaging technology and Multimedia Forensics.