Identifying Deepfake images from real images

$https://github.com/sohil3002z/dip-project$

1st Sohil Agarwal
UG Student VIT Chennai

2nd Abhijay Dhodapkar
UG Student VIT Chennai

***Abstract*—The rapid advancement of Deepfake technology has introduced significant current and potential future challenges to our daily lives. As Deepfake images increasingly resemble real image, various detection methods employing deep learning models have been explored. However, while some existing techniques show promising performance in statistical evaluations, they often overlook the underlying forensic traces unique to Deepfakes. In this research, we investigate these distinctive noise traces within Deepfake image frames and propose a detection approach based on analyzing such noise patterns using a deep neural network. Our method involves training a Siamese noise extractor using a novel strategy that distinguishes between synthesized facial areas and unaltered backgrounds. Additionally, we introduce a similarity matrix module to compare the forensic noise traces of cropped facial and background regions in candidate image frames, thereby enhancing Deepfake detection accuracy.**

I. Introduction

The advent of Deepfake technology has ushered in a new era of multimedia manipulation, presenting both unprecedented opportunities and profound challenges. Deepfakes, synthetic media generated by deep learning algorithms like generative adversarial network (GAN), have the ability to convincingly alter audiovisual content, often superimposing individuals' faces onto existing footage with remarkable realism. While this technology holds promise for creative expression and entertainment, its misuse poses significant threats to the integrity of information and the fabric of society.

One of the most pressing concerns surrounding Deepfakes is their potential to deceive and manipulate individuals, organizations, and entire communities. As Deepfake videos become increasingly difficult to distinguish from genuine content, the spread of misinformation, slander, and propaganda poses grave risks to public discourse, political stability, and personal reputation. The rapid proliferation of Deepfake technology has underscored the urgent need for robust detection mechanisms capable of identifying and mitigating the spread of synthetic media.

In response to this imperative, researchers have explored various approaches to detecting Deepfakes, leveraging advancements in deep learning and computer vision. While existing methods have demonstrated promising results in statistical evaluations, many overlook the intrinsic forensic traces embedded within Deepfake image frames. These traces, which arise from the synthetic generation process, offer unique insights that can enhance the accuracy and reliability of detection algorithms.

In this study, we delve into the realm of noise traces within Deepfake imagery, seeking to extract their characteristics and exploit them for detection purposes. Our research aims to address the gap in existing literature by proposing a novel approach that leverages deep neural networks to analyze and distinguish between authentic and synthetic media based on these noise traces. Specifically, we focus on developing a Siamese noise extractor trained using a novel face-background strategy, which enables the differentiation of synthesized facial areas from unaltered backgrounds.

Furthermore, we introduce a similarity matrix module designed to quantify the differences in forensic noise traces between cropped facial and background regions within candidate image frames. By systematically analyzing these traces, our approach seeks to provide a more robust and reliable method for detecting Deepfakes, thereby mitigating their harmful impacts on society.

In the subsequent sections of this paper, we detail the methodology, experimental setup, and results of our investigation, culminating in a comprehensive evaluation of our proposed detection approach. Through our research, we aim to contribute to the ongoing efforts to combat the proliferation of Deepfake technology and safeguard the integrity of multimedia content in the digital age.

II. Literature Review

Younus et al. [1] proposed detecting deepfakes using Haar wavelet transform followed by some edge detection algorithm as deepfakes has inconsistencies in high frequency regions and edges. Li et al. [2] found that deepfakes can be identified using forensic symmetry between front face images and side face images. Zhang et al. [5] revealed that the deepfakes can be found by training DL on facial landmark points and analysing the on the characteristics like spectrum and texture while Kolagati et al. [3] used Multi Layer Perceptron to expose deepfakes from the real images. Deepfakes can also be found by exploring the varying color spaces through representative forgery learning as done by Amin et al. [4]. Convolutional Neural network is also used to solve the same issue, binary classification can be done, can detect subtle distortions in image to classify [10][11]. Ganguly et al. [6] used ViXNet that combines a Vision Transformer with an Xception Network to capture both local and global inconsistencies in facial regions. It aims to identify imperceptible artifacts left by deepfaking methods. Kingra et al. [7] analyzed facial texture irregularities using the Local Binary Pattern(LBP) texture analysis method and Deep learning-based model called LBPNet. Using spatiotemporal convolution network, exploiting prediction error inconsistencies through LSTM-based classifiers , and generalizing GAN

image forensics is done by Nguyen et al.[8] while John et al.[9] used semi supervised GAN to solve this problem

III. Methodology

The Siamese noise extractor extracts the Deepfake forensic noise traces from the face and background squares, respectively, and the noise traces are then analyzed via similarity matrix for the Deepfake detection results.

A. Face Backgound strategy

Deepfake usually only modifies the face area when performing face-swapping, and most of the background area remains un-changed. For each keyframe, we locate the face position using the dlib library2 and crop the face square and a background square that has the largest Euclidean distance from the face square. This face-background strategy of locating the furthest background area guarantees to crop the background square that is the least likely to be modified by Deepfake even though the background area close to the face may be modified along with the target face. In other words, for each face-background pair, the cropped background square is always unmodified while the face square may be manipulated by Deepfake with noise traces left behind

B. Siamese noise trace extraction

Siamese design for noise trace extraction from the face squares and background squares where the two branches share the same weights. In particular, for a face-background pair, the face square and the background square are each passed through one branch of the Siamese architecture. In each Siamese branch, a pre-trained DnCNN denoiser is adopted and improved for Deepfake forensic noise trace extraction. For a real image frame that is unmodified, it contains the same noise pattern everywhere within the image, in other words, no Deepfake forensic noise trace. On the other hand, a Deepfake image frame has the face area synthesized such that the underlying noise pattern of the face area is different from that of the unmodified background area. Since the two branches of the Siamese architecture share the weights on noise trace extraction, different noise patterns are extracted under the same noise trace extraction process from the face and background squares of the Deepfake synthesized videos, while the same noise patterns can be found within the face and background squares of the real videos. The Siamese architecture is coded as one single network in implementation since both branches share the same set of network weights. The extracted forensic noise traces from the face and background squares are further fed to a similarity matrix design for noise trace pattern comparison and Deepfake detection decision making.

C. Noise similarity analysis

Considering the forensic noise traces for the unmodified area are always clean, the Deepfake manipulated faces are expected to have complicated noise traces. We use the similarity matrix on the extracted face and background noise traces for the noise similarity analysis. In specific, the similarity matrix is implemented as the inner product of the two noise representations to find out the correspondence between every two vector entries. Following the convention, multiplication is more powerful to find relations between deep neural network feature matrices than summations. When performing matrix summation, only entries in the same position can be summed up and thus the correspondence is weaker than conducting a product operation.

References

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first . . .”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

References

[1] Younus, Mohammed Akram, and Taha Mohammed Hasan. ”Effective and fast deepfake detection method based on haar wavelet transform.” 2020 International Conference on Computer Science and Software Engineering (CSASE). IEEE, 2020.

[2] Li, Gen, Xianfeng Zhao, and Yun Cao. ”Forensic symmetry for DeepFakes.” IEEE Transactions on Information Forensics and Security 18 (2023): 1095-1110.

[3] Kolagati, Santosh, Thenuga Priyadharshini, and V. Mary Anita Rajam. ”Exposing deepfakes using a deep multilayer perceptron–convolutional neural network model.” International Journal of Information Management Data Insights 2.1 (2022): 100054.

[4] Amin, Muhammad Ahmad, et al. ”Exploring Varying Color Spaces through Representative Forgery Learning to Improve Deepfake Detection.” Digital Signal Processing (2024): 104426.

[5] Zhang, Jixin, et al. ”A Heterogeneous Feature Ensemble Learning based Deepfake Detection Method.” ICC 2022-IEEE International Conference on Communications. IEEE, 2022.

[6] Ganguly, Shreyan, et al. ”ViXNet: Vision Transformer with Xception Network for deepfakes based video and image forgery detection.” Expert Systems with Applications 210 (2022): 118423.

[7] Kingra, Staffy, Naveen Aggarwal, and Nirmal Kaur. ”LBPNet: Exploiting texture descriptor for deepfake detection.” Forensic Science International: Digital Investigation 42 (2022): 301452.

[8] Nguyen, Thanh Thi, et al. ”Deep learning for deepfakes creation and detection: A survey.” Computer Vision and Image Understanding 223 (2022): 103525

[9] John, Jerry, and Bismin V. Sherif. ”Comparative Analysis on Different DeepFake Detection Methods and Semi Supervised GAN Architecture for DeepFake Detection.” 2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC). IEEE, 2022.

[10] Jellali, Ameni, Ines Ben Fredj, and Kais Ouni. ”Data Augmentation for Convolutional Neural Network DeepFake Image Detection.” 2023 IEEE International Conference on Advanced Systems and Emergent Technologies (IC-ASET). IEEE, 2023.

[11] Gummadi, Sai Dheeraj, and Anirban Ghosh. ”Deep residual learning based discriminator for identifying deepfakes with cut-out regularization.” 2022 IEEE world conference on applied intelligence and computing (AIC). IEEE, 2022.