Why Do Facial Deepfake Detectors Fail?

Binh Le Sungkyunkwan University South Korea bmle@g.skku.edu

Shahroz Tariq CSIRO's Data61 Australia shahroz.tariq@data61.csiro.au

Alsharif Abuadbba CSIRO's Data61 Australia sharif.abuadbba@data61.csiro.au

Kristen Moore CSIRO's Data61 Australia kristen.moore@data61.csiro.au

Simon S. Woo Sungkyunkwan University South Korea swoo@g.skku.edu



Input image

Face detected from engine #2

Figure 1: Illustration of the face detected from different engines. While the first approach uses a fixed-size central cropped patch from the detected face, the second resizes it, resulting in stretched input image for detectors.

obstacles, we hope to provide insight into the limitations of current deepfake detection methods and stimulate further research in this critical area.

To address the challenges of deepfake detection, it is crucial to understand the limitations and pitfalls of existing approaches. Despite the emphasis on detection accuracy, there is a lack of consideration for explainability in deepfake detection. In this article, we will delve into two specific scenarios where using deepfake detectors may lead to unexpected results. Firstly, a mismatch between the pre-processing pipeline used in the deployment and the one used during training can compromise the detector's performance. Secondly, a lack of diversity in the datasets used during training can lead to biased and unreliable results.

Pre-processing: A newcomer to the field of deepfake detection may encounter challenges in obtaining accurate results when using an off-the-shelf detection tool, as they may not be familiar with the pre-processing pipeline of the tool. It is important to note that deepfake videos are not simply fed directly into the detector. This is because the deepfake detection models typically require the input of a certain size, whereas the original image/video may be a different size. To deal with this, pre-processing is performed, which may obscure the deepfake artifacts that the detector relies on. Worse, there is no standard pre-processing pipeline nor a standardized size for inputs to deepfake detectors. This paper aims to shed light on

ABSTRACT

Recent rapid advancements in deepfake technology have allowed the creation of highly realistic fake media, such as video, image, and audio. These materials pose significant challenges to human authentication, such as impersonation, misinformation, or even a threat to national security. To keep pace with these rapid advancements, several deepfake detection algorithms have been proposed, leading to an ongoing arms race between deepfake creators and deepfake detectors. Nevertheless, these detectors are often unreliable and frequently fail to detect deepfakes. This study highlights the challenges they face in detecting deepfakes, including (1) the pre-processing pipeline of artifacts and (2) the fact that generators of new, unseen deepfake samples have not been considered when building the defense models. Our work sheds light on the need for further research and development in this field to create more robust and reliable detectors.

CCS CONCEPTS

• Security and privacy \rightarrow Security services.

KEYWORDS

Deepfake Detection, Image manipulation, Adversarial noise, Selfsupervised learning

INTRODUCTION 1

Deepfakes, a term derived from "deep learning" and "fake" has gained popularity in recent years due to their ability to manipulate images, videos, and audio in a highly realistic manner using artificial intelligence (AI) algorithms. With the increasing sophistication of deepfakes, there is a growing need for effective methods of deepfake detection to combat their potential for harm [24, 30]. In response, an increasing number of deepfake detection methods have been proposed, employing techniques such as biometric analysis using appearance and behaviors [1], inter-frame inconsistencies using spatiotemporal data [26], texture enhancement with multi-attention maps [31], few shot-based [14] and continual learning-based [7, 8] methods for deepfake detection generalization, in addition to other deep learning algorithms [2, 12, 13, 25].

The robustness of deepfake detectors is crucial, particularly in cybersecurity applications such as Facial Liveness Verification, where their failure could have serious consequences [15]. In this paper, we aim to shed light on some of the challenges that deepfake detectors face when deployed in real-world situations. By exploring these

Table 1: Performance (ACC and AUC) of deepfake detector trained on raw FaceForensics++ datasets, and validated under different circumstances.

Test type	FF++ test set (dlib ₄)	Video comp.	dlib ₁₅	MTCNN ₄	Adv. noise	CelebDF-v2
ACC	0.980	0.615 (↓ .365)	0.951 (↓ .290)	0.970 (↓ .010)	0.002 (↓ .978)	0.526 (↓ .454)
AUC	0.994	0.788 (↓ . 206)	0.989 (↓ .005)	0.993 (↓ .001)	0.000 (↓ .994)	0.573 (↓ .421)

the impact of pre-processing on the detection process and provide explainability/guidance around one of the major preprocessing tasks — when to crop versus resize input.

Pre-processing techniques, such as resizing and cropping, are widely used in the pipeline of deepfake detectors, but their effects on detection performance are often overlooked. Our analysis reveals that resizing elongates the face to match the specified size, while cropping does not have this effect. This can lead to issues if the model is trained on stretched faces but deployed on naturally proportioned images. On the other hand, shrinking the input size through resizing may result in crucial features being lost. Therefore, we have found that selecting resizing over cropping when reducing the input size would negatively impact detection accuracy.

Dataset/Deepfake generator diversity: In recent years, the diversity of deepfake datasets and deepfake generators has proliferated, supporting research in the area. Some of the most popular deepfake datasets include DeepFake Detection Challenge [5], Face-Forensics++ [21], and Celeb-DF [17]; and they vary in terms of quality and methods used for generating them. Alongside this, a wide range of generators have been published, aiming to be more accessible and user-friendly, including DeepFaceLab, GAN-based and AutoEncoder-based generators.

Nevertheless, most state-of-the-art (SoTA) deepfake detectors have been developed to detect a specific type of deepfake dataset, leading to performance degradation on new, unseen deepfakes. Their generalization limitations also include the variation in deepfake quality, as demonstrated by their poor performance on compressed or manipulated inputs. Our study illuminates the correlation between deepfakes datasets. In particular, we employ both frequency transformation and deep-learning embeddings to visualize their interdependence and distribution. In this way, we highlight undisclosed reasons that may lead to the poor performance of a biased detector that was learned from limited types of deepfakes or limited generation toolkits.

2 BACKGROUND

While the term deepfakes can be used to refer to any artificial replacement using AI, we limit ourselves in this work to facial deepfakes [19]. In general, there are two categories of facial manipulation approaches: face reenactment and face-swapping. Face reenactment involves changing the facial expressions, movements, and speech of a person in a video to make it appear as though they are saying or doing things they never actually did. In face-swap deepfakes, the face of a person in a video or image is replaced with someone else's face, making it appear as if the latter person was present in the original footage.

In this study, we utilize the FaceForensics++ dataset [21], which is a well-known deepfake dataset that was created for validating different deepfake detection algorithms. From 1000 real videos, the authors generated a corresponding 1000 synthesized videos using DeepFakes [3], Face2Face [28], FaceSwap [4], NeuralTextures [27], and FaceShifter [16] algorithms. Among these, Neural-Textures and Face2Face are reenactment methods; the others are face-swapping algorithms. In addition, to increase the diversity, we include CelebDF-v2 [17] dataset, which is created by several published deepfake apps for face swapping, and fine-tuned by a sequence of post-processing steps, making it a highly realistic dataset.

3 METHODOLOGY

3.1 Data-preprocessing

For FaceForensics++ datasets, we follow the same preprocessing step as in ADD [11]: 92,160, 17,920, and 17,920 images for training, validation, and testing, respectively. Each set has a balanced number of real and fake images, and the fake images are derived from all five deepfake datasets. For the CelebDF-v2, we used 16,400 for solely validating the pre-trained model.

In order to detect faces from a video, we used the dlib [9] toolkit with padding factor of 15% and 3%, respectively, and MTCNN with padding of 4% to simulate different face detection engines.

3.2 Training and validation

We utilized ResNet50 as our detector and built a binary classifier. All the input images were resized to 224×224 , and we used detected faces with padding of 3% for training. The models were trained with the Adam [10] optimizer with a learning rate of 2e - 3, scheduled by one-cycle strategy [23]. Only random horizontal flip is applied during training. We used a mini-batch size of 192. During every epoch, the model was evaluated ten times, and we saved the best weight based on the validation accuracy. Early stopping [20] was applied when the model didn't not improve after 10 consecutive validation times.

4 EVALUATION

In this section, we evaluate the pre-trained model under the following different scenarios discussed below. The experimental results are presented in Table 1.

4.1 Video compression

Deepfakes are detected by their subtle artifacts, represented by high-frequency components. Various methods of lossy compression, including video compression and JPEG compression, can successfully eliminate these fine-grained artifacts, leading high rate of false-positive prediction. As shown in the second column of Table 1, we applied the H.264 codec with constant rate quantization parameters of 23 to raw videos. As a result, the pre-trained detector drastically dropped its accuracy from 98% to 61.2%. Why Do Facial Deepfake Detectors Fail?



Figure 2: High-frequency discrepancies of central cropped face (Left; Crop: 156) vs. resize large cropped face (Right; Resize: 156).



Figure 3: Similarities between frequency density lines from Figure 2. The higher values indicate the higher similarities between datasets (Left; Crop: 156 and Right; Resize: 156).

4.2 Face extraction approach

The effects of using different face extraction approaches are indicated in the fourth and fifth columns of Table 1. While the MTCNN detector can slightly reduce the performance of the ResNet50 model, using dlib with larger padding can substantially decrease its performance in terms of both accuracy and AUC scores. We argue that since the model had learned only from the facial features, the complex background in some contexts affected the model's attention.

Our second exercise in this category was inspired by a recent live-face detection algorithm [29] which proposed to use fixedsize patches cropped from the original faces instead of resizing the input. The explained reason is that resizing step can distort the discriminative features. To further examine this hypothesis, we conducted a pilot study in which we selected 5,000 images from each deepfake dataset and performed central crop and resizing steps, respectively, on them, as illustrated in Figure 1. Next, we applied Fourier transformation and extracted the average density representation of each dataset in the frequency domain [6]. The results are provided in Figure 2. As one may observe, central cropping results in high-frequency differences between datasets. Resizing, on the other hand, pushes the high-frequency representations of datasets close together, making it difficult to distinguish.

4.3 Adversarial noise

Deep neural networks are well known for their adversarial nature showing through their vulnerability against adversarial examples. The adversarial samples are created by adding small, often imperceptible, perturbations to the original inputs. To validate this property, we apply one-step L_{∞} white-box PGD attack [18] with a small perturbation size of 1/255 and step size $\epsilon = 1/255$. As indicated in the sixth column of Table 1, almost all the predictions are flipped, as demonstrated by an accuracy score close to *zero*. Therefore, deploying deepfake detectors in practice should consider this aspect and have proper pre-processing steps or defense mechanisms to eliminate the effect of adversarial samples.

4.4 Data shift

Data shift refers to changes in the statistical properties of the data distribution used to train the detection model compared to the distribution of data the model encounters in deployment. In fact, data shift in deepfake detection can be a result of different factors: ethnicity (*e.g.*, Asian vs. African), environment (*e.g.*, indoor vs. outdoor), generating method (*e.g.*, Neural texture vs. FaceSwap). We show the results of cross-dataset validation in the final columns of Table 1. Although the model was trained over five datasets of FaceForensics++, it still struggles to distinguish deepfake from the



Figure 4: t-SNE visualization of various deepfake datasets using a pre-trained self-supervised learning embedding model

CelebDF-v2 dataset, indicated by its performance of approximately random guesses.

To explain this phenomenon, we perform two experiments to visualize the relationships between datasets. First, from the density representations of datasets from Figure 2, we use negative distance to indicate the closeness between datasets that are formulated as $max - ||a - b||_2^2$. As we can observe from Figure 3, the cropping step introduces less relationship between datasets compared to resizing. Nevertheless, in both approaches, there is less relation between FaceForensics++ datasets and CelebDF-v2, both in real and deepfake parts. In our second experiment, we utilize a pre-trained "self-supervised learning" model, SBI [22], with EfficientNet-B4 backbone to get the intermediate representations of each deepfake dataset. As illustrated in Figure 4, each deepfake dataset has its own distribution in the latent space. Therefore, if a detection model solely learns a single dataset, its decision boundary may lose its generalization for others, leading to the degradation of its performance.

5 REMARKS

Despite a plethora of ongoing research aimed at improving the accuracy of deepfake detectors, there is also a multitude of factors that hinder their performance. These include pre-processing steps, intended manipulation from attackers, and ongoing advancement of deepfake technology induces the low generalization of pre-trained detectors. In this paper, we quantially and visually expose these factors from the explainability viewpoints. This study also raises the awareness of researchers of not only developing effective deepfake detectors but also putting their effort into mitigating those crucial factors, reducing false positive and negative rates in practice.

ACKNOWLEDGMENTS

This work was partly supported by Institute for Information & communication Technology Planning & Evaluation (IITP) grants funded by the Korean government MSIT: (No. 2022-0-01199, Graduate School of Convergence Security at Sungkyunkwan University), (No. 2022-0-01045, Self-directed Multi-Modal Intelligence for solving unknown, open domain problems), (No. 2022-0-00688, AI Platform to Fully Adapt and Reflect Privacy-Policy Changes), (No. 2021-0-02068, Artificial Intelligence Innovation Hub), (No. 2019-0-00421, AI Graduate School Support Program at Sungkyunkwan University), and (No. 2023-00230337, Advanced and Proactive AI Platform Research and Development Against Malicious Deepfakes).

REFERENCES

- Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. 2020. Detecting deep-fake videos from appearance and behavior. In 2020 IEEE international workshop on information forensics and security (WIFS). IEEE, 1–6.
- [2] Young Oh Bang and Simon S Woo. 2021. DA-FDFtNet: dual attention fake detection fine-tuning network to detect various AI-generated fake images. arXiv preprint arXiv:2112.12001 (2021).
- [3] DeepFakes Community. 2017. DeepFakes. https://github.com/deepfakes/ faceswap. Accessed: 2021-01-01.
- [4] FaceSwap Community. 2016. FaceSwap. https://github.com/MarekKowalski/ FaceSwap. Accessed: 2021-01-01.
- [5] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397 (2020).
- [6] Ricard Durall, Margret Keuper, and Janis Keuper. 2020. Watch your upconvolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 7890–7899.
- [7] Minha Kim, Shahroz Tariq, and Simon S Woo. 2021. Cored: Generalizing fake media detection with continual representation using distillation. In Proceedings of the 29th ACM International Conference on Multimedia. 337–346.
- [8] Minha Kim, Shahroz Tariq, and Simon S Woo. 2021. Fretal: Generalizing deepfake detection using knowledge distillation and representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 1001–1012.
- [9] Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. Journal of Machine Learning Research 10 (2009), 1755–1758.
- [10] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).
- [11] Binh M Le and Simon S Woo. 2021. ADD: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images. arXiv preprint arXiv:2112.03553 (2021).
- [12] Binh M Le and Simon S Woo. 2021. Exploring the Asynchronous of the Frequency Spectra of GAN-generated Facial Images. arXiv preprint arXiv:2112.08050 (2021).
- [13] Sangyup Lee, Jaeju An, and Simon S Woo. 2022. BZNet: Unsupervised Multiscale Branch Zooming Network for Detecting Low-quality Deepfake Videos. In Proceedings of the ACM Web Conference 2022. 3500–3510.
- [14] Sangyup Lee, Shahroz Tariq, Junyaup Kim, and Simon S Woo. 2021. Tar: Generalized forensic framework to detect deepfakes using weakly supervised learning. In ICT Systems Security and Privacy Protection: 36th IFIP TC 11 International Conference, SEC 2021, Oslo, Norway, June 22–24, 2021, Proceedings. Springer, 351–366.
- [15] Changjiang Li, Li Wang, Shouling Ji, Xuhong Zhang, Zhaohan Xi, Shanqing Guo, and Ting Wang. 2022. Seeing is living? rethinking the security of facial liveness verification in the deepfake era. In 31st USENIX Security Symposium (USENIX Security 22). 2673–2690.
- [16] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2019. Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457 (2019).
- [17] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 3207–3216.
- [18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017).
- [19] Yisroel Mirsky and Wenke Lee. 2021. The creation and detection of deepfakes: A survey. ACM Computing Surveys (CSUR) 54, 1 (2021), 1–41.
- [20] Lutz Prechelt. 1998. Early stopping-but when? In Neural Networks: Tricks of the trade. Springer, 55–69.
- [21] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated

facial images. In Proceedings of the IEEE/CVF international conference on computer vision. 1–11.

- [22] Kaede Shiohara and Toshihiko Yamasaki. 2022. Detecting deepfakes with selfblended images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 18720–18729.
- [23] Leslie N Smith and Nicholay Topin. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, Vol. 11006. International Society for Optics and Photonics, 1100612.
- [24] Shahroz Tariq, Sowon Jeon, and Simon S Woo. 2022. Am I a Real or Fake Celebrity? Evaluating Face Recognition and Verification APIs under Deepfake Impersonation Attack. In Proceedings of the ACM Web Conference 2022. 512–523.
- [25] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S Woo. 2019. Gan is a friend or foe? a framework to detect various fake face images. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. 1296– 1303.
- [26] Shahroz Tariq, Sangyup Lee, and Simon Woo. 2021. One detector to rule them all: Towards a general deepfake attack detection framework. In *Proceedings of*

the web conference 2021. 3625-3637.

- [27] Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred neural rendering: Image synthesis using neural textures. Acm Transactions on Graphics (TOG) 38, 4 (2019), 1–12.
- [28] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2387–2395.
- [29] Chien-Yi Wang, Yu-Ding Lu, Shang-Ta Yang, and Shang-Hong Lai. 2022. Patch-Net: A simple face anti-spoofing framework via fine-grained patch recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 20281–20290.
- [30] Mika Westerlund. 2019. The emergence of deepfake technology: A review. Technology innovation management review 9, 11 (2019).
- [31] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. 2021. Multi-Attentional Deepfake Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2185–2194.