Advancements in Deepfake Detection and Disinformation

Dan-Cristian Stanciu
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
dan.stanciu1203@upb.ro

Bogdan Ionescu
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
bogdan.ionescu@upb.ro

***Abstract*—Artificial Intelligence (AI) is more popular than ever, with millions of people having access to open-source, paid and free models that can generate and evaluate large amounts of data. This raise in popularity of deep learning technology has as many drawbacks as it has advantages. Nowadays, information and disinformation can be very hard to distinguish, with many multimodal disinformation posts appearing everyday on the internet. An important example of this are deepfakes: images and videos generated by AI that portray real people and can misinform, skew the public perception towards them or even frame them. Therefore, there is an urgent need of good, reliable technology that can detects all forms of disinformation, with deepfakes being at the forefront. This paper presents our work in the field of disinformation detection, along with the current state of the art and the open challenges that come with the current technology. In addition, this paper will showcase our efforts to combat disinformation with the Multimedia Against Disinformation Campaign.**

***Index Terms*—deepfake detection, generative ai, disinformation, generalization, evaluation**

I. Introduction

Artificial Intelligence has seen a rise in popularity with the mainstream public over the last few years, with the release of Large Language Models like ChatGPT. Machine learning models are used more widely than ever and more accessible than ever. For example, nearly half of young people use AI on a daily basis. At the same time, it is important to note that deep learning models can create as much harm as good. A clear example of that are deepfakes - videos and images generated by artificial intelligence. The interest in fake news, disinformation and deepfakes has seen a substantial rise in just the last five years. For example, the term ”deepfake” is 70% more searched now than 4-5 years ago. At the same thime, the efforets to fight disinformation have increased as well.

This paper outlines the current state of disinformation and disinformation detection, focusing on:

- The current state of the art of deepfake detection
- The open challenges in deepfake detection and in disinformation classification in general
- Our efforts in the fight against disinformation
- A conclusion regarding the state of the field, the needs and further steps

2nd Workshop on Artificial Intelligence for Multimedia, AI Multimedia Lab, CAMPUS Research Institute, POLITEHNICA Bucharest, Bucharest, Romania, 15 July, 2025

II. Related Work

A. Deepfake Detection Approaches

Recent advancements in deepfake detection have led to the development of several high-performing methods that leverage deep learning architectures to identify forged media. Among the most popular and robust methods is the 3D R50-FTCN architecture proposed in [1], targeting artifacts and inconsistencies in the video stream and using 3D convolution, as well as Video Transformers to leverage the information in video deepfakes. This approach is particularly effective because many deepfakes are created frame-by-frame rather than as continuous video. As a result, temporal inconsistencies—such as unnatural facial movements or flickering between frames—can be detected using spatiotemporal cues. Another way of approaching deepfake detection is presented in [2], which combines data augmentation with robust inconsistency detection. This model generates new identity swap deepfakes during training and learns to identify intra-image inconsistencies—mismatches within a single frame that may occur between the synthesized face and the real background, lighting, or facial contours. A third very popular approach that benchmarks the best in popular datasets such as FaceForensics++, CelebDF or DFDC [3]–[5] is leveraging blending artifacts in face-swap deepfakes. The method presented in [6], called EFNB4 + SBIs, targets blending artifacts, which are subtle visual cues that emerge when a synthesized face is pasted onto a real person’s body, a technique often used in identity swap deepfakes. Identity manipulation is one of the most common forms of deepfakes, where the face of one individual is replaced with that of another, while the body and background remain largely unchanged.

Overall, deepfake detection methods are very varied, targeting a wide variety of particularities of deepfakes. At the same time, no method is necessarily better than the others. Some methods are even too specific, working on just some types of deepfakes.

B. Open Challenges in Deepfake Detection

Deepfake detection is still a new field, with a lot of challenges for researchers to overcome. Some of those challenges are:

- **Generalization**. The majority of the deepfake detection models are deep learning-based, and learn without much outside input. Additionally, every deepfake is generated by another deep learning model. A deepfake detector can learn to detect the generator's "fingerprints", rather than the general concept of "deepfake". This results in a generalization problem: deepfake classifiers have a very high accuracy on the training sets, but a low accuracy on new data. Because new deepfake generators are created every day, this becomes a cat and mouse game, where the already known models can be detected, but new ones elude deepfake detectors.
- **Deepfake Datasets.** The most deepfake detection datasets, like FaceForensics++. CelebDF, DFDC etc. are trained for identity swap detection. At the same time, more and more models like ChatGPT or SORA AI can generate images from scratch, especially for very well known people.
- **Dataset Fairness and Distribution.** Many deepfake detection datasets have some bias, because the deepfake detection datasets are not necessarily geared towards fairness, as the data is very hard to obtain. For example, many deepfake detection datasets are geared towards celebrities.
- **Data Processing.** When uploading a video on real world websites, some kind of preprocessing occurs (compression, resizing etc.). While the deepfake detectors are trained with this in mind, the evaluation data is not geared towards this level of complexity. Therefore, while the raw videos can be correctly classified as deepfakes, the models have difficulties classifying videos that have been compressed or changed in any meaningful way.

III. Advancements in Disinformation detection

In this chapter, we will present the advancements we made towards improving the state of disinformation detection, focusing on 3 aspects: (i) Improving generalization in deepfake detection, evaluating and implementing state-of-the art deepfake detectors in a fair way, focusing on generalization and encouraging collaboration in the research community with the Multimedia Against Disinformation workshop.

A. Improving Generalization in Deepfake Detection

Our work on deepfake detection was implemented with 2 key aspects in mind: improving generalization and performance.

For improving performance, we proposed two complementary approaches to improve deepfake detection performance and efficiency. First, we developed a modified Capsule Network (CapsNet) architecture [7] that preserves spatial hierarchies by removing pooling layers, increasing the number of primary capsules, and refining the routing algorithm. This model achieved an AUC of 99.88% on the CelebDF dataset, and maintained strong performance (99.56% AUC) even in a reduced version with only 6.4 million parameters—highlighting its scalability and suitability for resource-constrained environments. Second, we introduced a CNN-LSTM-based temporal detection framework [8] that focuses on key facial features—specifically the mouth, eyes, and nose—to capture subtle temporal inconsistencies across video frames. This targeted approach proved especially effective in identifying localized manipulations, demonstrating robustness against more nuanced deepfakes.

To address the challenge of generalization in deepfake detection, we proposed two novel augmentation-based strategies aimed at reducing overfitting and improving cross-dataset performance.

First, we developed an autoencoder-based augmentation technique [9] designed to minimize the model's reliance on generator-specific artifacts (commonly referred to as "fingerprints"). By regenerating training images using over 80 distinct autoencoder configurations, we introduced controlled variations that help suppress these artifacts. This approach led to substantial improvements in cross-dataset generalization, increasing AUC by nearly 10% on CelebDF and 2% on DFDC when trained on FaceForensics++. It also enhanced robustness to common perturbations such as lossy compression and adversarial noise. Second, we introduced a recurrent adversarial augmentation framework [10] that generates synthetic deepfake-like samples from real images using adversarial perturbations. These challenging examples are iteratively incorporated into the training process, enabling the model to adapt continuously and improve its detection capabilities. Without requiring additional real or fake data, our method achieved significant generalization gains—up to 10% AUC improvement on CelebDF and 9% on DFDC Preview.

Together, these augmentation strategies offer flexible, model-agnostic solutions that significantly improve deepfake detection performance across diverse datasets and conditions, advancing the development of real-world-ready detection systems.

B. Evaluation Deepfake Detection Approaches and Open Challenges

At the moment, we are working on a benchmark evaluation paper, that aims to implement and compare the most important state-of-the-art approaches in deepfake detection. The overview paper aims to see whether the implemented algorithms can maintain their performance in a variety of situations (against adversarial attacks, perturbations, edits, compression, processing algorithms), as well as evaluate whether the current models have biases. Below are a few initial conclusions drawn from the early experiments:

- Both light and heavy compression affect the current state-of-the-art models, significantly lowering performance. More, image perturbations as color changes, resolution changes can affect the performance of those models.
- Some state-of-the-art models are hard to implement due to the lack of details, and their performance can very drastically compared to the reported values.
- Most deepfake detection algorithms are trained on FaceForensics++ or CelebDF and evaluated on the rest of

the available datasets. The 2 aforementioned datasets might not be the most suitable for this task (especially FaceForensics++, as it is an "easy" dataset, compared to today's standards.

- The biggest problem in deepfake detection, besides generalization, is the lack of explainability, especially in the case of video. There is a limited number of papers that approach this issue, with the majority focusing on approaches like GradCam. However, it is very hard for the public to trust the level of explainability in the models right now.

C. The Multimedia Against Disinformation Workshop

To help combat disinformation and foster connections between researchers, we are the organizers of the Multimedia Against Disinformation Workshop (MAD) [11]. This year, the 4th edition of the Multimedia Against Disinformation Workshop (MAD) was held at the International Conference on Multimedia Retrieval, in Chicago, USA. Over 4 editions of the workshop, 38 papers were presented out of a total of 54 submissions. The objective of the MAD workshop is to provide a collaborative platform for researchers and practitioners working on disinformation in multimedia content. It aims to foster interdisciplinary dialogue, promote the exchange of innovative ideas, and accelerate progress in AI-powered detection and analysis of disinformation. By encouraging the sharing of methodologies, tools, and challenges, MAD seeks to advance the development of robust, trustworthy, and scalable solutions to combat the growing threat of misleading and manipulated media across digital platforms. Papers presented at the MAD workshop focus on AI-driven methods to detect and analyze disinformation in multimedia content. Topics include deepfake and synthetic media detection, multimodal verification, social media analysis, robustness against adversarial attacks, fairness, cultural aspects, dataset creation, and real-world disinformation campaign studies.

IV. Conclusion

This paper presents the state of disinformation detection and deepfake detection, as well as our work to fight disinformation over the years. Overall, there are a lot of open challenges in the deepfake detection field. Our work sims to solve the ones we consider the most important. For example, our work in deepfake generalization aims to provide robust models that are not sensitive to changes in the data. Our work in evaluating and benchmarking state-of-the-art deepfake detection models aims to provide an overview of the current best models and provide insights regarding the strengths and weaknesses of those models. Lastly, we presented an overview of the Multimedia Against Disinformation Workshop, now in its 4th edition.

In conclusion, while there are many challenges in the field of disinformation, there are as many opportunities to evaluate and improve the current approaches, as well as to innovate in the future.

References

[1] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 044–15 054.

[2] 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.

[3] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *International Conference on Computer Vision (ICCV)*, 2019.

[4] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3207–3216.

[5] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[6] 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.

[7] D.-C. Stanciu and B. Ionescu, "Uncovering the strength of capsule networks in deepfake detection," in *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*, ser. MAD '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 69–77. [Online]. Available: https://doi.org/10.1145/3512732.3533581

[8] ——, "Deepfake video detection with facial features and long-short term memory deep networks," in *2021 International Symposium on Signals, Circuits and Systems (ISSCS)*, 2021, pp. 1–4.

[9] ——, "Autoencoder-based data augmentation for deepfake detection," in *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, ser. MAD '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 19–27. [Online]. Available: https://doi.org/10.1145/3592572.3592840

[10] ——, "Improving generalization in deepfake detection via augmentation with recurrent adversarial attacks," in *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation*, ser. MAD '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 46–54. [Online]. Available: https://doi.org/10.1145/3643491.3660291

[11] D.-C. Stanciu, B. Ionescu, S. Papadopoulos, G. Kordopatis-Zilos, A. Popescu, R. Caldelli, M. Gerhardt, and V. Schmitt, "Mad'25: 4th acm international workshop on multimedia ai against disinformation," in *Proceedings of the 2025 International Conference on Multimedia Retrieval*, ser. ICMR '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 2148–2150. [Online]. Available: https://doi.org/10.1145/3731715.3734512