Forensic AI: A Novel Multi-Granular Approach for Detecting Synthetic Media Manipulation

Aakash Mor
University of the Arts London
aakashmor@gmail.com

***Abstract*—With the rapid evolution of AI-driven synthetic media, deepfake technology has emerged as a potent tool for deception, raising concerns about misinformation, security, and digital authenticity. This study introduces a novel multi-granular artifact detection framework to distinguish between real and AI-generated images with high precision. By leveraging deep learning techniques and a custom convolutional architecture, the approach identifies intrinsic and extrinsic inconsistencies left by synthetic media generation processes. Extensive experiments across diverse datasets validate the robustness of the model, demonstrating superior generalization across both seen and unseen deepfake samples. The research provides a step forward in forensic AI, ensuring enhanced media integrity and mitigating the growing threats posed by deepfake technology.**

***Index Terms*—Deepfake Detection, Synthetic Media, Digital Forensics, Convolutional Neural Networks, AI-generated Images, Media Integrity, Artifact Detection, Deep Learning**

I. Introduction

Recent breakthroughs in generative modeling, particularly through deep learning architectures, have markedly advanced the synthesis of facial imagery. Among these innovations, the technique commonly referred to as DeepFake has gained prominence due to its capacity to generate lifelike facial swaps and its accessibility for practical use. In essence, DeepFake technology involves replacing an individual's facial appearance in a genuine video sequence with a digitally rendered representation of another person's face, all while preserving intricate details such as facial movements, expressions, and head orientation.

Despite its potential in fields such as entertainment and digital communication, DeepFake technology presents considerable ethical and security challenges. Malicious actors can exploit it to fabricate scenarios and visual content that never transpired, undermining public confidence and posing significant risks to digital authenticity, privacy, and social order. As such, the development of robust and adaptable mechanisms to accurately identify synthetic media has emerged as a critical area of research.

To address this challenge, multiple detection strategies have been formulated, leveraging various sources of information, such as manually extracted visual artifacts, high-level semantic inconsistencies, or learned representations through deep neural networks. Nevertheless, the problem remains far from resolved, owing to two prominent difficulties. First, the continuous refinement of generative techniques has significantly diminished the perceptual and statistical differences between manipulated and real footage. Second, there exists a marked degradation in the generalizability of detection algorithms when evaluated against previously unseen datasets or DeepFake variants, reflecting a strong domain dependency and sensitivity to distribution shifts.

This paper contributes to the ongoing efforts by analyzing contemporary detection frameworks and identifying key shortcomings in their design. It also proposes a novel strategy that combines temporal and spatial feature modeling to improve resilience across different forgery styles. By systematically evaluating this approach under multiple conditions and datasets, the study aims to push the boundary of general-purpose DeepFake detection solutions.

II. Motivation

The proliferation of DeepFake technologies has introduced profound concerns across multiple sectors, including the domains of politics, entertainment, journalism, and personal digital identity. These advanced generative systems possess the capability to create hyper-realistic visual content—both still and moving images—wherein human faces are convincingly synthesized or replaced. Such media fabrications are increasingly indistinguishable from genuine recordings, complicating efforts to discern authenticity and amplifying the risk of public deception.

The potential misuse of such synthetically altered content extends to numerous malicious scenarios: the spread of disinformation campaigns, the deliberate alteration of historical records, digital impersonation, and targeted character attacks. As these threats intensify, the necessity for resilient and intelligent detection systems becomes a priority to ensure the credibility of online visual content and defend against coordinated manipulative efforts.

Given these escalating challenges, the design and implementation of reliable DeepFake identification frameworks are essential. These systems must go beyond superficial analysis and leverage deep structural or forensic cues to pinpoint anomalies and inconsistencies introduced during generative processes. By enabling automated validation of media content, such frameworks will support digital authenticity, strengthen public trust, and help mitigate the propagation of visually deceptive narratives.

The current research initiative aims to make a meaningful contribution to this pressing issue by examining advanced methodologies for the identification of tampered visual data.

Emphasis is placed on both theoretical innovation and practical evaluation to better understand the adversarial landscape of synthetic media, ultimately fostering the development of robust countermeasures against the evolving tactics used in facial forgery systems.

III. Objective

A. Formulate an Innovative Detection Architecture

Propose and develop a novel framework—referred to as the Bi-granularity Artifacts (BiG-Arts) model—designed to identify both low-level and high-level discrepancies embedded within manipulated facial data. This dual-resolution artifact detection strategy targets intrinsic inconsistencies (e.g., texture-level noise) and extrinsic distortions (e.g., geometric misalignments), both of which are frequently introduced during DeepFake synthesis procedures. The approach utilizes a multi-task learning paradigm to jointly optimize the recognition of these diverse artifact types, promoting increased precision and generalization capacity.

B. Enhance Generalization and Accuracy Across Datasets

Pursue a significant improvement in detection accuracy for both in-domain (within the same dataset) and out-of-domain (across diverse datasets) conditions. The proposed method seeks to exploit the complementary nature of bi-granularity features to strengthen its adaptability to unseen manipulation techniques and varied visual contexts. Comprehensive evaluations are conducted using benchmark datasets to affirm performance gains and establish resilience against domain drift.

C. Evaluate Against Existing Detection Benchmarks

Conduct comparative analyses to measure the efficacy of the BiG-Arts method relative to leading approaches in DeepFake identification. By benchmarking results using standard metrics—such as Area Under the Curve (AUC), precision-recall values, and F1-score—the study aims to underscore the strengths and limitations of the proposed system. Particular attention is given to its placement in global challenges such as the DeepFake Detection Challenge (DFDC), where competitive evaluation against state-of-the-art systems provides insight into real-world applicability.

D. Address Real-World Implementation Barriers

Investigate solutions to address common obstacles encountered during deployment, including the rapid evolution of generative architectures and the reduced transferability of current detection methods across different forgery styles and video compression levels. A special focus is placed on enhancing model robustness through architecture regularization, adversarial training, and domain adaptation mechanisms.

E. Support Public Safety and Ethical Digital Practices

Deliver an effective and transparent forensic tool capable of identifying manipulated facial content at scale, thereby contributing to larger initiatives in cybersecurity, media verification, and digital policy enforcement. The intention is to develop technologies that can be adopted by legal authorities, media verification entities, and digital platform providers to prevent the exploitation of synthetic media for unethical purposes.

F. Promote Future Research and Open Challenges

Identify potential research gaps and offer directions for further innovation based on the outcomes and boundaries of the BiG-Arts framework. By making datasets, implementation protocols, and evaluation results publicly accessible, the project aims to stimulate collaborative progress and support the continued evolution of DeepFake detection methodologies in the broader scientific community.

IV. Related Work

A. Advancements in Face Synthesis

In recent years, the evolution of deep generative architectures has significantly enhanced face editing methodologies, including but not limited to Generative Adversarial Networks (GANs), facial attribute transformation, and identity replacement. One notable method among these is the DeepFake technique, which has garnered extensive interest due to its capacity to seamlessly overlay the facial identity of a target onto the source face while preserving features such as facial orientation and emotional expression. This process commonly utilizes a Variational Autoencoder (VAE)-based model where the encoder is designed to eliminate identity-dependent components. At the same time, the decoder reconstructs the facial characteristics of the designated individual.

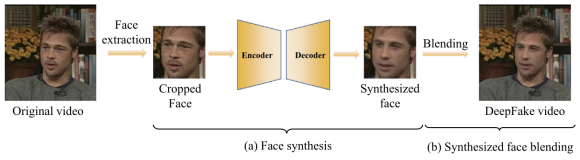

Fig. 1. Deepfake Generation

B. Detection Techniques for DeepFake Media

Numerous detection frameworks have emerged in response to the rise of synthetic visual media. These methodologies can be broadly grouped into f the distinct categories: learning-based strategies, spectral domain analysis, artifact-centric models, and temporal coherence-based techniques.

In learning-driven approaches, detectors are trained end-to-end on both genuine and fabricated visual content. Models such as MesoNet, XceptionNet, and CapsuleNet have demonstrated effectiveness by leveraging deep feature extraction from both domains. Enhancements have been proposed using hybrid components like Central Difference Convolution (CDC)

and Atrous Spatial Pyramid Pooling (ASPP), as seen in the MTD-Net architecture.

In the frequency-domain arena, methods like those by Agarwal et al. have introduced dual-path architectures capturing discrepancies in both spatial and spectral characteristics. Additionally, Luo et al. employed SRM filtering techniques to extract noise-specific features, developing a multi-scale module for diverse high-frequency signal extraction.

V. Dataset and Feature Preprocessing

This research utilizes a curated subset of the DeepFake Detection Challenge (DFDC) dataset, accessible via Kaggle. While the full dataset exceeds 470GB, a carefully selected 100GB segment was used for model development. Within this subset, a substantial proportion—approximately 83%—consists of synthetically generated videos. This disproportion is due to the generation of multiple deepfake variations (ranging from 1 to 22) from each authentic clip, averaging about 5.19 manipulations per source.

All videos adhere to a uniform specification of 10-second duration with a consistent frame rate of 30 frames per second. Subjects span diverse demographics and environments, from well-lit interiors to dimly lit outdoor backdrops. To transition the problem from video classification to still image analysis, representative frames were extracted at intervals of 60 frames (i.e., every 2 seconds). Each selected frame underwent resizing to 224×224 pixels, pixel-wise normalization (scaling by 255), and random data augmentation involving changes in contrast, brightness, and saturation. For each genuine frame, an average of 3–4 fake counterparts were included to enrich the dataset.

VI. Proposed Framework

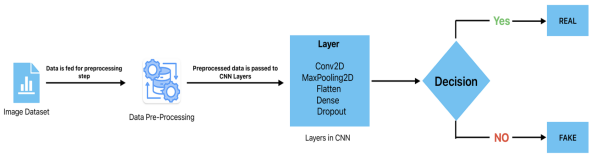

Fig. 2. Proposed Model

A. Architectural Composition

The detection model is anchored on a Convolutional Neural Network (CNN) structure comprising three convolutional blocks interleaved with max-pooling layers. The convolutional layers utilize the ReLU activation function to enable non-linear transformations, while pooling layers reduce the dimensionality of intermediate representations, thereby retaining essential features and improving computational efficiency. The output feature maps are flattened and processed through two fully connected layers, both activated by ReLU, culminating in a final sigmoid layer for binary classification.

B. Optimization Strategy

Training is conducted using the Adam optimization algorithm due to its adaptive learning rate and momentum capabilities. Given the binary nature of the task, the loss function employed is the binary cross-entropy:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$$

Here, $y_i \in \{0,1\}$ represents the ground truth label, and $\hat{y}_i \in [0,1]$ denotes the predicted probability for each instance. The model is trained over 10 full epochs with an appropriate batch size, using the augmented image data for both training and validation stages. Training progress is monitored through step-wise backpropagation, with each epoch consisting of the total training samples divided by the batch size.

C. Performance Assessment

Following the completion of training, model effectiveness is gauged on a reserved test set. Evaluation metrics include classification accuracy, confusion matrix, precision, and recall scores. The confusion matrix is denoted as:

$$\text{CM} = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$$

Where TP and TN represent true positives and true negatives respectively, and FP and FN denote false positives and negatives. Derived metrics are computed using:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

D. Prediction Pipeline

The trained CNN model is employed to infer the authenticity of unseen images. The input image undergoes identical preprocessing steps as the training data (resize, normalization), followed by classification using the model's prediction output. Based on a thresholded probability, the sample is labeled as either genuine or manipulated.

This approach harnesses the power of deep learning to discern tampered imagery and contributes to the collective effort in defending digital ecosystems against malicious synthetic media.

VII. Results

Upon completion of the training phase, the model underwent evaluation on a separate test subset. The classifier attained a final accuracy of approximately 88%, demonstrating its capacity to reliably differentiate between authentic and manipulated facial imagery. To gain a more nuanced understanding of its

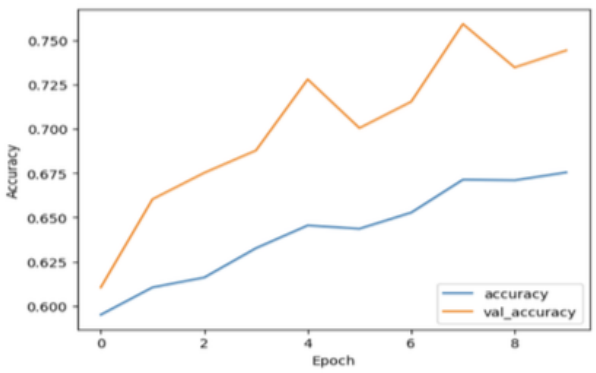

Fig. 3. Accuracy v/s Epoch

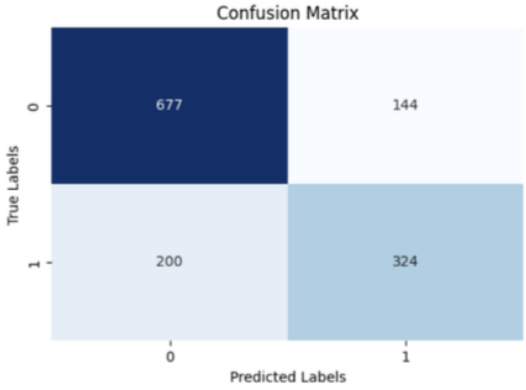

Fig. 4. Confusion Matrix

classification behavior, additional statistical indicators, such as sensitivity and specificity-based precision, were examined.

The sensitivity metric—commonly referred to as recall—quantifies the proportion of actual positive samples that were successfully recognized by the classifier. This measure was calculated to be 0.86, suggesting the framework is proficient in identifying manipulated media. Conversely, the specificity of detection, quantified through the precision score, assesses the model's capability to limit false positive predictions and was computed at 0.89. These findings imply a strong balance between correctly classifying synthetic content and minimizing incorrect assignments of genuine samples as forgeries.

The mathematical formulations used for deriving these evaluation metrics are as follows:

$$\text{Detection Accuracy} = \frac{C_{tp} + C_{tn}}{C_{tp} + C_{tn} + C_{fp} + C_{fn}}$$

$$\text{Recall Rate} = \frac{C_{tp}}{C_{tp} + C_{fn}}, \quad \text{Precision Ratio} = \frac{C_{tp}}{C_{tp} + C_{fp}}$$

Here, C_{tp} and C_{tn} denote the instances correctly recognized as synthetic and genuine, respectively. C_{fp} represents the number of genuine images incorrectly labeled as manipulated, while C_{fn} accounts for manipulated images misclassified as authentic.

The model's behavior was also visualized through a confusion matrix and epoch-wise performance plots. The confusion matrix delivers a clear summary of classification outcomes across both categories, highlighting the distribution of correct and incorrect predictions. In parallel, the training and validation accuracy trajectories across multiple epochs reveal the learning progression and convergence of the model over time. These graphical representations are instrumental in diagnosing overfitting, underfitting, or performance plateaus, offering key insights for further architectural enhancements or training refinements.

VIII. Conclusion

To summarize, the developed framework for identifying synthetically altered visual content has exhibited commendable performance in accurately distinguishing between genuine and tampered imagery. The integration of convolutional neural networks, in conjunction with strategic image augmentation, has contributed to the system's capability to detect inconsistencies within facial features that are commonly indicative of manipulative generation techniques. These findings reinforce the model's utility as a reliable instrument for addressing the proliferation of fabricated visual content in digital ecosystems.

The robustness of the approach and its promising classification metrics underscore its potential application in real-world scenarios where authenticity of visual data is critical. The methodology employed lays the groundwork for scalable, efficient solutions to combat the rise of maliciously generated media artifacts, especially in contexts such as journalism, digital forensics, and content authentication platforms.

Despite the effectiveness of the current architecture, there remain several opportunities for enhancement and expansion. One promising direction involves the adoption of more sophisticated image transformation strategies during training to enrich the diversity of the dataset. Approaches such as CutMix, RandAugment, and mixup augmentation—each capable of introducing novel image compositions and variations—could further bolster the generalization power of the classifier and reduce susceptibility to overfitting.

Moreover, incorporating temporal coherence across sequential frames in video datasets or leveraging attention-based transformer architectures may introduce deeper semantic understanding, potentially enabling the detection of subtle manipulations that evade conventional pixel-level models. As deepfake synthesis techniques evolve with increasing realism, continuous adaptation and innovation in detection frameworks will be essential for maintaining the credibility and trustworthiness of digital media.

In future iterations, emphasis may also be placed on real-

time inference capabilities, model interpretability, and domain adaptation for multilingual or multicultural datasets. Such improvements would collectively enhance the system's adaptability and resilience across diverse application domains.

IX. References

1) J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Nießner, *Face2Face: real-time face capture and reenactment of RGB videos*, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2387–2395.
2) D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, *MesoNet: a compact facial video forgery detection network*, in: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7, doi:10.1109/WIFS.2018.8630761.
3) A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, *FaceForensics++: learning to detect manipulated facial images*, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1–11.
4) F. Chollet, *Xception: deep learning with depthwise separable convolutions*, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.
5) Y. Li, S. Lyu, *Exposing DeepFake videos by detecting face warping artifacts*, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019.
6) H.H. Nguyen, J. Yamagishi, I. Echizen, *Capsule-forensics: using capsule networks to detect forged images and videos*, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2307–2311, doi:10.1109/ICASSP.2019.8682602.
7) H. He, E.A. Garcia, *Learning from imbalanced data*, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.
8) D. Güera, E.J. Delp, *DeepFake video detection using recurrent neural networks*, in: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018, pp. 1–6, doi:10.1109/AVSS.2018.8639163.
9) X. Li, Y. Lang, Y. Chen, X. Mao, Y. He, S. Wang, H. Xue, Q. Lu, *Sharp multiple instance learning for DeepFake video detection*, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1864–1872.