Multi-Modal Semantic Distillation for Image Forgery Detection

Hima Mariya T*, Sudeepth Sasikumar*, Karthika K*, Irine Rose V Chazoor*, Anjana B†

*Student, Vidya Academy of Science and Technology, Thrissur

† Assistant Professor, Vidya Academy of Science and Technology, Thrissur

***Abstract*—In today's digital era, multimedia content such as images, videos, and audio plays a crucial role in communication, entertainment, and information sharing. However, with the rapid advancement of digital editing tools, forgery and manipulation of multimedia data have become easier, posing serious challenges to authenticity, privacy, and trust. Multimedia forgery detection is a system designed to identify and prevent the misuse or alteration of digital media. This project aims to develop a secure and efficient platform that allows the detection of forged multimedia content while maintaining user interaction and complaint management. The system includes three major roles: Admin, Cybercell, and User. The Admin manages user verification, cybercell approvals, and complaint handling. The Cybercell acts as the investigation authority, analyzing multimedia data for forgery (image, video, and audio) and taking appropriate actions such as blocking or unblocking users. The User can register, share posts, interact through comments and chats, and file complaints against suspicious activities or fake content. The system integrates multimedia forgery detection modules to analyze suspicious media and generate authenticity reports. This ensures a secure digital environment and helps law enforcement authorities handle cyber-related cases more efficiently.**

***Index Terms*—Image Forgery, Artificial Intelligence, Deep Learning, Social Media Security, Audio Forgery**

I. INTRODUCTION

Advances in generative models such as Generative Adversarial Networks (GANs) and diffusion models have enabled the creation of realistic fake images and audio. Deepfakes, face swaps, and synthetic voice generation tools are now easily accessible, making digital forgery a major threat to online trust and security. As a result, forgery detection has become a crucial research area in multimedia forensics.

Early works in forgery detection focused primarily on visual artifacts in images or videos. More recently, audio deepfake detection has emerged as a parallel research domain due to the rise of synthetic speech and voice cloning technologies. Although multi-modal fusion methods exist, many practical systems adopt a modality-specific strategy, where image and audio forgeries are detected using separate pipelines. This sur-vey focuses on such modality-specific approaches and reviews their effectiveness in real-world applications.

Advancements in Artificial Intelligence (AI), Machine Learning (ML), and Computer Vision have enabled automated accident detection systems using surveillance cameras, traffic data, and intelligent transportation infrastructure. These systems aim to reduce detection time, assess accident severity, and initiate rapid emergency response.

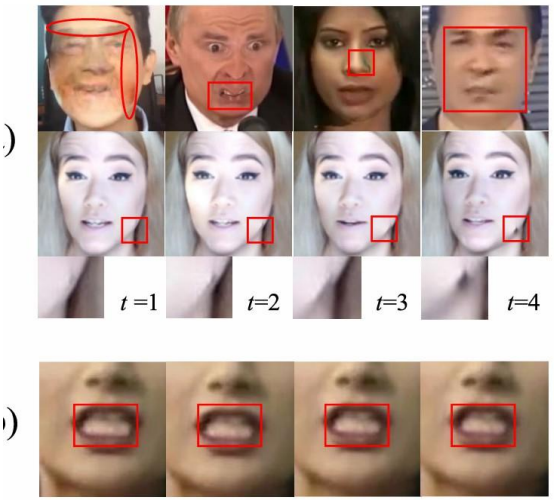

Fig. 1. The examples of low-level forgery artifacts and high-level semantic irregularities. The (a) row represents common low-level forgery artifacts including blending boundary, blur artifacts, color mismatch, mouth blurring, flickering, and so on. The (b) row represents high-level semantic inconsistencies. For instance, the speaker's teeth are not naturally part as the lips are opening.

II. LITERATURE SURVEY

A. Audio Spoofing Detection Using Capsule Networks

Reference [8]: A. Luo et al., "A capsule network based approach for detection of audio spoofing attacks," ICASSP 2021.

Luo et al. proposed an audio deepfake detection framework based on Capsule Networks, focusing on identifying spoofed and synthetic speech. Unlike traditional CNN-based methods, capsule networks preserve hierarchical relationships between acoustic features, allowing the model to better capture subtle artifacts introduced during speech synthesis and voice conversion. The approach utilized spectral features such as MFCCs and demonstrated improved robustness compared to conventional architectures. This work emphasized the importance of modality-specific audio analysis and showed that audio forgery detection requires specialized architectures distinct from visual-based models, supporting the design choice of handling audio detection separately.

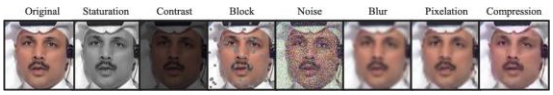

Fig. 2. Corruption examples. Examples of the corruptions considered in our robustness experiments at severity level 5; these corruptions were introduced in [79]. It consists of changes in saturation and contrast, block-wise distortions, white Gaussian noise, Gaussian blurring, pixelation, and video compression.

B. Lip-Synchronization-Based Audio-Visual Forgery Detection

Audio-visual forgery detection methods based on lip synchronization analyze the consistency between facial lip movements and corresponding speech signals. These techniques rely on high-level semantic cues rather than low-level pixel artifacts, resulting in improved robustness and cross-dataset generalization. By detecting mismatches between audio and visual streams, lip-sync based approaches effectively identify manipulated videos, even under compression. Despite their advantages, these methods require the availability of synchronized audio and video data and involve higher computational complexity due to audio-visual alignment. Their performance may degrade when forgeries preserve accurate lip synchronization or when audio data is missing.

C. CNN-Based Face Forgery Detection Using XceptionNet

Convolutional Neural Network-based models, particularly XceptionNet, have been widely adopted for face forgery detection due to their strong feature extraction capabilities. These methods detect low-level visual artifacts such as tex-ture inconsistencies and blending errors introduced during manipulation. XceptionNet-based approaches have achieved high accuracy on benchmark datasets like FaceForensics++. However, their reliance on dataset-specific artifacts limits their generalization ability. Performance often declines when faced with compressed videos, noise, or unseen forgery techniques, as these models lack semantic-level understanding of facial behavior.

III.RESEARCH GAP

Despite significant progress in AI-based multimedia forgery detection, several critical research gaps remain evident in the existing literature.

Most existing forgery detection approaches focus on isolated modalities, such as image-based or audio-based detection, without providing a unified yet flexible framework that can independently analyze different media types. Many systems attempt tightly coupled multi-modal fusion, which increases architectural complexity and reduces interpretability. The lack of modular, modality-aware frameworks limits scalability and real-world deployability of forgery detection systems.

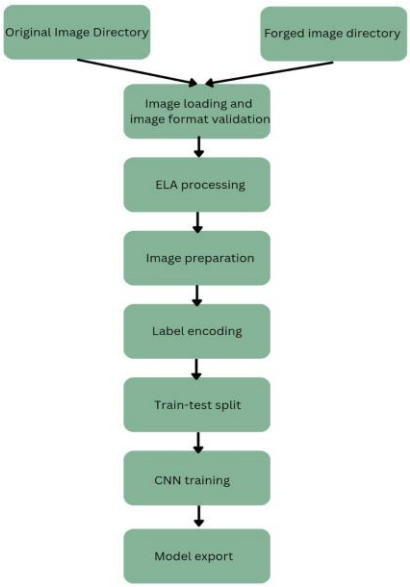

Fig. 3. image detection framework

The availability of large-scale, diverse, and real-world forgery datasets remains limited. A majority of existing models are trained and evaluated on curated datasets generated under controlled conditions. This restricts the generalization capability of detection models when deployed in real-world environments involving compression, noise, varying recording devices, and unseen manipulation techniques.

Deep learning models used for image and audio forgery detection often operate as black-box systems, providing limited explainability. The absence of interpretable decision-making mechanisms poses a significant challenge in security-critical applications, where transparency, accountability, and user trust are essential for adoption by cyber authorities and legal systems.

There is insufficient emphasis on real-time and edge-based deployment of forgery detection systems. Most existing studies validate their models in offline environments without address-ing computational constraints, latency, and scalability issues associated with continuous real-time monitoring on social media platforms or embedded devices.

Furthermore, current research places limited focus on independent confidence-aware decision mechanisms for different modalities. In scenarios where either audio or visual data is missing, corrupted, or unreliable, tightly fused multi-modal systems may fail. This highlights the need for modality-specific detection pipelines that can function independently while still contributing to a reliable overall decision.

IV.FUTURE RESEARCH DIRECTIONS

- Modality-Specific End-to-End Frameworks: Future systems should focus on developing modular yet integrated architectures that independently analyze image and audio forgeries while supporting flexible end-to-end decision-making for multimedia authenticity verification.
- Large-Scale Real-World Forgery Datasets: There is a strong need for publicly available, large-scale multi-media forgery datasets captured under diverse real-world conditions, including different compression levels, noise, recording devices, and unseen manipulation techniques, to improve model robustness and generalization.
- Explainable AI for Digital Forensics: Future research should incorporate Explainable AI (XAI) techniques to improve transparency and interpretability of forgery detection models, enabling better trust, accountability, and acceptance in forensic, legal, and cybercrime investigation applications.
- Real-Time and Edge-Based Deployment: Lightweight deep learning models and edge-computing solutions should be explored to support low-latency, real-time image and audio forgery detection on social media plat-forms, surveillance systems, and embedded devices.
- Independent and Confidence-Aware Decision Mechanisms: Designing confidence-based decision strategies for individual modalities can improve reliability in scenarios where either audio or visual data is missing, corrupted, or unreliable.
- Scalable Multi-Modal Analysis: While maintaining modality independence, future systems may explore scalable and interpretable strategies for selectively combining audio and image detection results to enhance overall detection accuracy.
- Institutional and Platform-Level Readiness: Research should address standardization, ethical guidelines, and platform-level integration to ensure effective real-world deployment of multimedia forgery detection systems across social media, cybersecurity, and digital governance infrastructures.

V.CONCLUSION

This work highlights the growing importance of robust multimedia forgery detection in the context of rapidly advancing image and audio manipulation techniques. Through an analysis of existing literature, it is evident that while significant progress has been made using deep learning–based approaches, current systems continue to face challenges related to generalization, computational complexity, and real-world deployability.

Most existing methods rely on tightly coupled multi-modal fusion or modality-specific models that operate as black-box systems, limiting

interpretability and scalability. To address these limitations, the proposed approach emphasizes modality-specific forgery detection, where image and audio forgeries are analyzed independently using dedicated detection pipelines. This design improves robustness against unseen attacks, reduces system complexity, and enables flexible deployment across diverse real-world scenarios.

Furthermore, the integration of lightweight models, explainable decision mechanisms, and confidence-aware outputs enhances the reliability and trustworthiness of the detection process, making the system suitable for cyber safety and digital forensic applications. Overall, the proposed framework contributes toward building practical, scalable, and interpretable multimedia forgery detection systems capable of addressing evolving threats in modern digital environments.

REFERENCES

[1] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.

[2] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Pro-tecting World Leaders Against Deepfakes," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019: IEEE, pp. 38–45.

[3] M. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Dis-tinguishing Computer Graphics from Natural Images Using Con-volutional Neural Networks," in IEEE Workshop on Information Forensics and Security (WIFS), 2017: IEEE, pp. 1–6.

[4] T. Dang, D. Nguyen, and H. Le, "A Deep Learning Approach for Image Forgery Detection," in International Conference on Computer Science and Information Technology, 2020: Springer, pp. 201–210.

[5] S. Tariq, S. Lee, H. Kim, and S. Woo, "Detecting Both Machine and Human Created Fake Face Images in the Wild," in ACM Workshop on Information Hiding and Multimedia Security (IH-MMSec), 2021: ACM, pp. 35–45.

[6] N. Dufour and O. Gloe, "Detection of Deepfake Videos: A Survey," in International Conference on Availability, Reliability and Security (ARES), 2020: IEEE, pp. 1–10.

[7] A. Trivedi, K. Jangal, and R. Gupta, "Identifying Deepfake Cyber Attacks: Challenges and Countermeasures," in International Jour-nal for Research in Applied Science and Engineering Technology (IJRASET), 2023: IJRASET, pp. 215–223.

[8] D. Vaishnavi, D. Mahalakshmi, and V. Siva Rao, "Visual Feature-Based Image Forgery Detection," in International Journal of En-gineering and Technology, 2018: Science Publishing Corporation, pp. 86–90.

[9] Gowsic K., Vinayaka Moorthi M., Siranjeevi S., and Viswa G., "Image Forgery Detection Using Convolutional Neural Network Algorithm," in ShodhKosh: Journal of Visual and Performing Arts, 2024: Granthaalayah Publications, pp. 1067–1073.

[10] M. Saha, P. Singh, and N. Rai, "Forensic Techniques for Forgery Detection and Localization in Digital Images," in IJRASET Jour-nal for Research in Applied Science and Engineering Technology, 2024: IJRASET, pp. 145–152.

[11] N. Kumar and A. Kundu, "SecureVision: Advanced Cybersecurity Deepfake Detection with Big Data Analytics," in Sensors, 2024: MDPI, vol. 24(19), article 6300.

[12] A. S. Bhardwaj, M. Kumar, and R. Bali, "Cyberbullying Detection Using Natural Language Processing Techniques," in International Conference on Smart Computing and Informatics, 2021: Springer, pp. 331–340.

[13] H. Zhou, L. Chen, and X. Li, "Fake Multimedia Detection Based on Deep Neural Networks," in IEEE International Conference on Multimedia and Expo (ICME), 2020: IEEE, pp. 185–190.

[14] M. Chawla, S. Verma, and P. Chauhan, "Real-Time Detection of Manipulated Images Using CNN," in International Journal of Computer Applications, 2021: IJCA, vol. 183, no. 15, pp. 25–30.

[15] A. Gupta and R. Sharma, "Deepfake Video Detection Using Capsule Networks," in International Conference on Computational Vision and Bio Inspired Computing, 2022: Springer, pp. 458–468.

[16] P. Kumar and M. Sharma, "A Review on Cybercrime Detection and Prevention Techniques," in International Conference on Com-munication and Cyber Security (ICCCS), 2021: IEEE, pp. 90–97.

[17] S. Patil and A. Deshmukh, "Deep Learning Based Framework for Multimedia

Authentication," in IEEE International Conference on Advances in Computing, Communication and Control (ICAC3), 2020: IEEE, pp. 305–312.

[18] R. Thomas and K. Menon, "An Integrated Approach for Detecting Cyberbullying in Online Social Media," in International Journal of Advanced Research in Computer Science, 2023: IJARCS, vol. 14, no. 2, pp. 45–50.

[19] Python Software Foundation, Python Lan-guage Reference, version 3.10, 2023. Available: https://www.python.org/

[20] Django Software Foundation, Django Documentation, 2023. Available:[https://www.djangoproject.com/] (https://www.djangoproject.com/)