Future-Proofing Deepfake Detection by Integrating Audio, Video, and Text

PASQUALE LISENA, EURECOM, France
KONG AIK LEE and YI WANG, The Hong Kong Polytechnic University, Hong Kong
MASSIMILIANO TODISCO, CHIARA GALDI, RAPHAEL TRONCY, and NICHOLAS EVANS, EURECOM, France
WEIWEI LIN, LAP-PUI CHAU, and MAN-WAI MAK, The Hong Kong Polytechnic University, Hong Kong

The rapid advancement of AI-generated content has made deepfakes increasingly realistic, posing serious risks to identity security, social trust, and public and democratic institutions. Existing detection systems, typically focused on single modalities such as video or audio, often fail to generalize to new manipulation techniques and cannot effectively detect hybrid or low-effort deepfakes. In this perspective letter, we advocate for a new paradigm in deepfake detection, that emphasizes the integration of audio, video, and textual content. We examine the limitations of current systems, including their over-reliance on outdated datasets and limited adversarial robustness. We outline the technical motivations for integrating these modalities, and highlight emerging research directions. By aligning detection strategies with the multimodal nature of AI-driven manipulation, we call for a new generation of systems that are more generalizable and trustworthy.

CCS Concepts: • **Security and privacy** → **Spoofing attacks**; • **Computing methodologies** → *Natural language processing*; • **Information systems** → *Multimedia and multimodal retrieval*.

Additional Key Words and Phrases: Deepfake detection, Multimodality, Identity spoofing, Generative AI

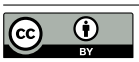
ACM Reference Format:
Pasquale Lisena, Kong Aik Lee, Yi Wang, Massimiliano Todisco, Chiara Galdi, Raphael Troncy, Nicholas Evans, Weiwei Lin, Lap-Pui Chau, and Man-Wai Mak. 2026. Future-Proofing Deepfake Detection by Integrating Audio, Video, and Text. *ACM AI Lett.* 1, 1 (January 2026), 7 pages. https://doi.org/10.1145/3797958

1 Introduction

In January 2025, thirty-one individuals were arrested by the Hong Kong Police in connection to a deepfake dating scam. The syndicate used photos of attractive people found online and AI-facilitated calls to deceive victims, swindling the equivalent of over 4 million dollars in Singapore, Malaysia, and Taiwan[1]. In the same month, in France, a 53-year-old woman lost €850,000 after being tricked into believing she was in a relationship with actor Brad Pitt, facilitated by AI-generated images[2]. The proliferation of AI-generated content (AIGC) [7] has made it easier for scammers to lure victims with highly realistic fake materials. As a matter of utmost urgency, the misuse of generative AI to

[1]South China Morning Post, January 5, 2025
[2]Newsweek, January 14, 2025

Authors' Contact Information: Pasquale Lisena, pasquale.lisena@eurecom.fr, EURECOM, Sophia Antipolis, France; Kong Aik Lee, kong-aik.lee@polyu.edu.hk; Yi Wang, The Hong Kong Polytechnic University, Hong Kong; Massimiliano Todisco; Chiara Galdi; Raphael Troncy; Nicholas Evans, EURECOM, Sophia Antipolis, France; Weiwei Lin; Lap-Pui Chau; Man-Wai Mak, The Hong Kong Polytechnic University, Hong Kong.



ACM 3068-8590/2026/1-ART
https://doi.org/10.1145/3797958

create deepfakes could lead to far-reaching consequences, far surpassing the severity of the cases we have witnessed so far.

Deepfakes are a type of AIGC created by manipulating existing media or synthesizing content from scratch to deceive humans. These include technologies such as speech synthesis, voice conversion, face swapping, lip-syncing [47], puppeteering, and natural language generation. The primary objective of such manipulations is to achieve convincing identity spoofing, making it nearly impossible for the naked senses (vision and hearing) to distinguish between fake and genuine media. Consequently, deepfake detection has gained significant attention, as demonstrated by several relevant initiatives [6, 43, 48].

While much of the past work has focused on audio and video, the detection across the three prominent deepfake media types – visual, speech, and textual – remains under-explored, and a comprehensive framework to address these forms holistically is missing. We believe that employing a system that integrates the three modalities can significantly enhance the detection of deepfakes by compensating for the inherent weaknesses of each individual modality and for adversarial attacks that target single modalities. While visual analysis may fail to detect sophisticated video manipulations—such as subtle alterations to facial expressions or lighting—audio analysis can identify inconsistencies in speech patterns or background noise. Textual analysis, in turn, may reveal anomalies in transcripts, such as unnatural language patterns or discrepancies with the speaker's prior statements. Additionally, desynchronization between speech and lip movements, or mismatches between vocal emotional tone and facial expressions, can serve as indicators of potential manipulation. Such a cross-verification process may enhance the overall accuracy of deepfake detection, making it more reliable.

In this perspective letter, we examine the potential limitations of current single-modality and multimodal detection methods for identifying deepfake content. Next, we advocate for future research in deepfake detection to increasingly integrate audio-visual and textual modalities.

2 Limitations of Current Paradigms

Despite significant progress in deepfake detection, current fraud detection systems still face limitations in generalization, robustness, and adaptability.

One major limitation is the **predominant focus on image-based detection**. Even when videos are considered, most methods focus solely on spatial inconsistencies, neglecting temporal dynamics that could offer crucial detection cues. These methods aim to expose imperceptible inconsistencies or artifacts, often extracting individual frames – i.e., treating them as still images – and then applying detection techniques frame by frame [8, 9, 17, 42]. As a result, very few methods explicitly leverage temporal information, such as motion dynamics or temporal inconsistencies, which have shown improved generalization to unseen data [51].

Spoofed speech detectors focus on identifying features and artifacts specific to synthetic speech, detecting signatures unique to individual speech synthesizers. However, as many speech synthesizers are built on similar underlying architectures and techniques (e.g., using the same vocoders), these detectors can generalize to some extent to unseen synthesizers. Nevertheless, they often face **substantial performance degradation when encountering out-of-distribution (OOD) test samples**. This phenomenon, referred to as the domain shift problem, remains a critical challenge for current anti-spoofing countermeasures.

Text authorship attribution, a field that has been studied extensively in literature [1], has regained importance in the GenAI era, where generative models can convincingly impersonate roles with minimal prompt modifications [41]. Both traditional [4, 5] and more recent works [30, 36] leverage stylometric, syntactic and lexical features as reliable indicators of an author's writing style. However, these approaches have two clear limitations: (1) they have largely been treated

as standalone methods, making direct comparisons challenging due to their application across different datasets; (2) they have been predominantly applied to written content, leaving substantial room for **much-needed research on spoken language**, such as transcripts derived automatically from speech recordings.

Single-modality methods often rely on supervised learning, analyzing low-level features and artifacts from imperfections in the generation process [37, 39]. While effective for detecting known manipulations, their performance **degrades substantially against unseen techniques** due to the rapid evolution in synthetic data generation technologies. **Retraining models** on expanding training sets, encompassing content generated with more contemporary techniques **is computationally impractical**, and fine-tuning risks catastrophic forgetting. Solutions like few-shot learning [10, 16], incremental learning [19, 29], and augmentation [38, 49] enhance generalization but rely on acquiring representative examples of new manipulations, which remains a major challenge.

Current research in multimodal detection of AI-generated fake content has made significant strides, particularly in leveraging individual modalities such as audio [26, 31, 40], visual [14, 28, 46], and textual data [14, 15, 27] to identify manipulated content. Audio-video approaches mostly focus on detecting synchronization errors between audio and video streams [20, 21, 52]. However, advances in deepfake generation now produce highly realistic and seamlessly synchronized manipulations [35], reducing the efficiency of such methods. **The natural language is hardly a part of the combined modality**, despite some notable exceptions [45]. Moreover, current deepfake datasets reveal several limitations, such as the predomination of legacy deepfake generation techniques [25], lack of explicit annotations for audio forgeries [11], limited manipulation diversity [18], lack of adversarial attacks [2, 13].

3 Multimodality as a Strategic Imperative for Deepfake and Fraud Detection

The rise of highly convincing AI-generated content poses a significant challenge across security, media, and social domains. Traditional detection approaches, often limited to individual modalities, are insufficient to counter the sophistication of modern generative systems. For this reason, we argue that multimodal analysis – integrating video, audio, and text – is no longer a complementary direction but a strategic imperative.

Across modalities, each channel captures a distinct and complementary dimension of identity. **Video** encodes facial dynamics, micro-expressions, head movement, and other cues that reflect the physical embodiment of a speaker. **Audio** preserves vocal identity through prosody, rhythm, spectral structure, and breathing patterns—signals that are difficult to forge consistently. **Text** provides the cognitive and stylistic layer: lexical choices, syntactic preferences, discourse structure [4, 30], sentiment and topic patterns [32, 34], and background knowledge that form an individual's linguistic fingerprint [3, 22]. While each modality alone can be convincingly forged, **replicating the coherence of all three simultaneously remains substantially more challenging** and **cross-modal inconsistencies** are the strongest indicators of manipulation. A multimodal approach also enhances **robustness**: when one modality is noisy, deliberately perturbed, or expertly forged, the remaining channels can provide stable evidence.

To effectively leverage cross-modal correlations, a hybrid of early, late, and **semantic fusion strategies** can be employed. Ensemble architectures integrate the outputs of modality-specific detectors, while large language models are used to provide semantic-level alignment between modalities, capturing subtle inconsistencies that might be imperceptible otherwise. Such fusion frameworks may enable better generalization and detection of hybrid or low-effort deepfakes, particularly where existing methods [37, 39] fail against unseen manipulations. Fusion strategies

also build on prior work exploring audio-visual synchronization discrepancies [20, 21, 52], although it is crucial to overcome their limitations in the presence of high-quality, synchronized forgeries [35].

To enhance transparency, **hierarchical disentanglement techniques** can be adopted to isolate latent factors such as speaker identity, facial expression, and background context. These techniques address a known challenge in deepfake detection, namely, the difficulty of interpreting learned representations from black-box models, by structuring the representation space. In this context, a proper evaluation framework should be developed, tailored to partial deepfakes, where only segments of a video or utterance are manipulated. Hierarchical disentanglement should also be coupled with one-class learning strategies [50], focusing solely on modeling bonafide samples. During training, meta-learning techniques simulate domain shift scenarios, teaching the model to learn generalized and transferable decision boundaries. This approach avoids retraining on growing datasets and reduces the risk of catastrophic forgetting.

Deepfake detection systems are only as good as the data on which they are trained and evaluated. Unfortunately, most existing benchmarks fall short, relying on outdated generation methods, simplistic annotations, and an absence of adversarial scenarios. **New datasets are required using the latest generation pipelines**, e.g., voice conversion systems [24], lip synchronization [23], and diffusion-based text-to-video generators. To support localized detection, each sample should be annotated with metadata including forgery type, manipulation location, and speaker or character attributes. These annotations facilitate research in manipulation localization techniques and facilitate interpretable benchmarking, going beyond the global binary classification common in prior studies. Adversarial examples are crucial to develop robust detection systems, as in recent examples targets separately audio [33] and video [12], with some efforts also present in adversarial attacks in text [44].

4 Conclusions

Deepfakes present growing risks to identity security through the manipulation or fabrication of media. As generative techniques become more sophisticated, the limitations of single-modality deepfake detectors, particularly their vulnerability to unseen generation methods, underscore the need for more resilient and adaptive solutions. Multimodal approaches offer a promising direction by jointly analyzing audio and visual cues, especially their cross-modal inconsistencies. However, the current landscape is hindered by outdated and low-quality datasets that fail to capture the complexity of modern deepfakes.

To advance the field, we advocate for a comprehensive strategy built on six key priorities, as a roadmap for the next five years of research in deepfake detection:

(1) include spoken-text authorship features in detection;
(2) build modern multimodal datasets with realistic impersonation cases;
(3) establish metrics capturing cross-modal inconsistencies;
(4) design explainable fusion models with uncertainty quantification;
(5) develop adversarially robust architectures;
(6) encourage interdisciplinary collaboration across NLP, speech, vision, and cybersecurity.

These priorities should be supported by progress in adversarial robustness, meta-learning frameworks, and more fine-grained evaluation metrics. Together, they will lay the groundwork for the next generation of reliable and generalizable multimodal deepfake detection systems.

Acknowledgments

This work has been partially supported by the Innovation and Technology Fund, Hong Kong SAR (Grant n°MHP/048/24) and the French National Research Agency (ANR) within the COMPROMIS project (Grant n°ANR22-PECY-0011).

References

[1] Ahmed Abbasi, Abdul Rehman Javed, Farkhund Iqbal, Zunera Jalil, Thippa Reddy Gadekallu, and Natalia Kryvinska. 2022. Authorship identification using ensemble learning. *Scientific Reports* 12, 1 (2022), 9537. doi:10.1038/s41598-022-13690-4

[2] Sarah Barrington, Matyas Bohacek, and Hany Farid. 2025. The DeepSpeak Dataset. arXiv:2408.05366 [cs.CV] https://arxiv.org/abs/2408.05366

[3] Samy Benslimane, Jérome Azé, Sandra Bringay, Maximilien Servajean, and Caroline Mollevi. 2021. Controversy Detection: A Text and Graph Neural Network Based Approach. In *Web Information Systems Engineering - WISE 2021 - Part I* (Melbourne, VIC, Australia). Springer-Verlag, Melbourne, Australia, 339–354. doi:10.1007/978-3-030-90888-1_26

[4] Shane Bergsma, Matt Post, and David Yarowsky. 2012. Stylometric Analysis of Scientific Articles. In *2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, Montréal, Canada, 327–337. https://aclanthology.org/N12-1033/

[5] Michael Brennan and Rachel Greenstadt. 2009. Practical Attacks Against Authorship Recognition Techniques. In *Conference on Innovative Applications of Artificial Intelligence*. AAAI, Pasadena, California, USA, 6 pages. https://aaai.org/papers/257-3903-1-pb-iaai-09/

[6] Zhixi Cai, Abhinav Dhall, Shreya Ghosh, Munawar Hayat, Dimitrios Kollias, Kalin Stefanov, and Usman Tariq. 2024. 1M-Deepfakes Detection Challenge. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. Association for Computing Machinery, New York, NY, USA, 11355–11359. doi:10.1145/3664647.3689145

[7] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip Yu, and Lichao Sun. 2025. A Survey of AI-Generated Content (AIGC). *ACM Comput. Surv.* 57, 5, Article 125 (Jan. 2025), 38 pages.

[8] Hong-Shuo Chen, Mozhdeh Rouhsedaghat, Hamza Ghani, Shuowen Hu, Suya You, and C.-C. Jay Kuo. 2021. DefakeHop: A Light-Weight High-Performance Deepfake Detector. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, Online, 1–6. doi:10.1109/ICME51207.2021.9428361

[9] Davide Alessandro Coccomini, Roberto Caldelli, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. 2022. Cross-Forgery Analysis of Vision Transformers and CNNs for Deepfake Image Detection. In *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation* (Newark, NJ, USA) *(MAD '22)*. Association for Computing Machinery, New York, NY, USA, 52–58. doi:10.1145/3512732.3533582

[10] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. 2019. ForensicTransfer: Weakly-supervised Domain Adaptation for Forgery Detection. arXiv:1812.02510 [cs.CV]

[11] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The DeepFake Detection Challenge (DFDC) Dataset. arXiv:2006.07397 [cs.CV]

[12] Chiara Galdi, Michele Panariello, Massimiliano Todisco, and Nicholas Evans. 2024. 2D-Malafide: Adversarial Attacks Against Face Deepfake Detection Systems. In *2024 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, Darmstadt, Germany, 1–7. doi:10.1109/BIOSIG61931.2024.10786754

[13] Yang Hou, Haitao Fu, Chunkai Chen, Zida Li, Haoyu Zhang, and Jianjun Zhao. 2024. PolyGlotFake: A Novel Multilingual and Multimodal DeepFake Dataset. In *Pattern Recognition: 27th International Conference (ICPR), Proceedings, Part XIV* (Kolkata, India). Springer-Verlag, Berlin, Heidelberg, 180–193. doi:10.1007/978-3-031-78341-8_12

[14] Viktor Dénes Huszár and Vamsi Kiran Adhikarla. 2024. Securing Phygital Gameplay: Strategies for Video-Replay Spoofing Detection. *IEEE Access* 12 (2024), 52282–52301. doi:10.1109/ACCESS.2024.3385373

[15] Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan. 2022. Automatic Detection of Entity-Manipulated Text using Factual Knowledge. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Dublin, Ireland, 86–93. doi:10.18653/v1/2022.acl-short.10

[16] Hyeonseong Jeon, Youngoh Bang, Junyaup Kim, and Simon S. Woo. 2020. T-GD: transferable GAN-generated images detection framework. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. JMLR.org, Online, Article 441, 16 pages.

[17] Achhardeep Kaur, Azadeh Noori Hoshyar, Vidya Saikrishna, Selena Firmin, and Feng Xia. 2024. Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review* 57, 6 (2024), 159.

[18] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. 2021. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*

(Round 2). NeurIPS, Online, 14 pages.

[19] Sohail Ahmed Khan and Hang Dai. 2021. Video Transformer for Deepfake Detection with Incremental Learning. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 1821–1828. doi:10.1145/3474085.3475332

[20] Pavel Korshunov, Michael Halstead, Diego Castan, Martin Graciarena, Mitchell McLaren, Brian Burns, Aaron Lawson, and Sébastien Marcel. 2019. Tampered Speaker Inconsistency Detection with Phonetically Aware Audio-visual Features. In *International Conference on Machine Learning (Synthetic Realities: Deep Learning for Detecting AudioVisual Fakes)*. IEEE, Greater Noida, India, 1094–1098. doi:10.1109/IC3ECSBHI63591.2025.10991244

[21] Pavel Korshunov and Sébastien Marcel. 2018. Speaker Inconsistency Detection in Tampered Video. In *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, Rome, Italy, 2375–2379. doi:10.23919/EUSIPCO.2018.8553270

[22] Yasemin Lheureux. 2024. Predictive insights: leveraging Twitter sentiments and machine learning for environmental, social and governance controversy prediction. *Journal of Computational Social Science* 7, 1 (2024), 23–44. doi:10.1007/s42001-023-00228-5

[23] Chunyu Li, Chao Zhang, Weikai Xu, Jingyu Lin, Jinghui Xie, Weiguo Feng, Bingyue Peng, Cunjian Chen, and Weiwei Xing. 2025. LatentSync: Taming Audio-Conditioned Latent Diffusion Models for Lip Sync with SyncNet Supervision. arXiv:2412.09262 [cs.CV]

[24] Jingyi Li, Weiping Tu, and Li Xiao. 2023. Freevc: Towards high-quality text-free one-shot voice conversion. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Rhodes island, Greece, 1–5. doi:10.1109/ICASSP49357.2023.10095191

[25] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *IEEE/CVF conference on computer vision and pattern recognition*. IEEE, Denver, CO, USA, 3207–3216.

[26] Guoyuan Lin, Weiqi Luo, Da Luo, and Jiwu Huang. 2024. One-Class Neural Network With Directed Statistics Pooling for Spoofing Speech Detection. *IEEE Transactions on Information Forensics and Security* 19 (2024), 2581–2593. doi:10.1109/TIFS.2024.3352429

[27] Dongliang Luo, Yuliang Liu, Rui Yang, Xianjin Liu, Jishen Zeng, Yu Zhou, and Xiang Bai. 2025. Toward real text manipulation detection: New dataset and new solution. *Pattern Recognition* 157 (2025), 110828. doi:10.1016/j.patcog.2024.110828

[28] Asad Malik, Minoru Kuribayashi, Sani M. Abdullahi, and Ahmad Neyaz Khan. 2022. DeepFake Detection for Human Face Images and Videos: A Survey. *IEEE Access* 10 (2022), 18757–18775. doi:10.1109/ACCESS.2022.3151186

[29] Francesco Marra, Cristiano Saltori, Giulia Boato, and Luisa Verdoliva. 2019. Incremental learning for the detection and classification of GAN-generated images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, Delft, Netherlands, 1–6. doi:10.1109/WIFS47025.2019.9035099

[30] Ahmed M. Mohsen, Nagwa M. El-Makky, and Nagia Ghanem. 2016. Author Identification Using Deep Learning. In *15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, Anaheim, California, USA, 898–903. doi:10.1109/ICMLA.2016.0161

[31] Nicolas Müller, Pavel Czempin, Franziska Diekmann, Adam Froghyar, and Konstantin Böttinger. 2022. Does Audio Deepfake Detection Generalize?. In *Interspeech 2022*. ISCA, Incheon, Korea, 2783–2787. doi:10.21437/Interspeech.2022-108

[32] Mina Narayanan, Joshua Gaston, Gerry Dozier, Lisa Cothran, Clarissa Arms-Chavez, Marcia Rossi, Michael C. King, and Kelvin Bryant. 2018. Adversarial Authorship, Sentiment Analysis, and the AuthorWeb Zoo. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, Bengaluru, India, 928–932. doi:10.1109/SSCI.2018.8628806

[33] Michele Panariello, Wanying Ge, Hemlata Tak, Massimiliano Todisco, and Nicholas Evans. 2023. Malafide: a novel adversarial convolutive noise attack against deepfake and spoofing detection systems. In *Interspeech 2023*. ISCA, Dublin, Ireland, 2868–2872. doi:10.21437/Interspeech.2023-703

[34] Nektaria Potha and Efstathios Stamatatos. 2019. Improving author verification based on topic modeling. *Journal of the Association for Information Science & Technology* 70, 10 (2019), 1074–1088. doi:10.1002/asi.24183

[35] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In *28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 484–492.

[36] Cosmina-Mihaela Rosca, Adrian Stancu, and Emilian Marian Iovanovici. 2025. The New Paradigm of Deepfake Detection at the Text Level. *Applied Sciences* 15, 5 (2025), 28 pages. doi:10.3390/app15052560

[37] Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Andreas Nautsch, Xin Wang, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. 2023. *Introduction to Voice Presentation Attack Detection and Recent Advances*. Springer Nature, Singapore, 339–385.

[38] Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans. 2022. Rawboost: A Raw Data Boosting and Augmentation Method Applied to Automatic Speaker Verification Anti-Spoofing. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Singapore, 6382–6386.

[39] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. 2020. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion* 64 (2020), 131–148.

[40] Hoan My Tran, David Guennec, Philippe Martin, Aghilas Sini, Damien Lolive, Arnaud Delhay, and Pierre-François Marteau. 2024. Spoofed Speech Detection with a Focus on Speaker Embedding. In *Interspeech 2024*. ISCA, Kos, Greece, 2080–2084. doi:10.21437/Interspeech.2024-481

[41] Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 2001–2016. doi:10.18653/v1/2021.findings-emnlp.172

[42] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. 2022. M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection. In *Proceedings of the 2022 International Conference on Multimedia Retrieval* (Newark, NJ, USA) *(ICMR '22)*. Association for Computing Machinery, New York, NY, USA, 615–623. doi:10.1145/3512527.3531415

[43] Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, Nicholas Evans, Kong Aik Lee, Junichi Yamagishi, Myeonghun Jeong, Ge Zhu, Yongyi Zang, You Zhang, Soumi Maiti, Florian Lux, Nicolas Müller, Wangyou Zhang, Chengzhe Sun, Shuwei Hou, Siwei Lyu, Sébastien Le Maguer, Cheng Gong, Hanjie Guo, Liping Chen, and Vishwanath Singh. 2025. ASVspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech. *Comput. Speech Lang.* 95, C (Dec. 2025), 27 pages. doi:10.1016/j.csl.2025.101825

[44] Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. Measure and Improve Robustness in NLP Models: A Survey. In *2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, Seattle, United States, 4569–4586. doi:10.18653/v1/2022.naacl-main.339

[45] R.L.M.A.P.C. Wijethunga, D.M.K. Matheesha, Abdullah Al Noman, K.H.V.T.A. De Silva, Muditha Tissera, and Lakmal Rupasinghe. 2020. Deepfake Audio Detection: A Deep Learning Based Solution for Group Conversations. In *2nd International Conference on Advancements in Computing (ICAC)*, Vol. 1. IEEE, Malabe, Sri Lanka, 192–197. doi:10.1109/ICAC51239.2020.9357161

[46] Junfeng Xu, Weiguo Lin, Wenqing Fan, Jia Chen, Keqiu Li, Xiulong Liu, Guangquan Xu, Shengwei Yi, and Jie Gan. 2024. A Graph Neural Network Model for Live Face Anti-Spoofing Detection Camera Systems. *IEEE Internet of Things Journal* 11, 15 (2024), 25720–25730. doi:10.1109/JIOT.2024.3383673

[47] Zhiyuan Yang, Lap-Pui Chau, and Bihan Wen. 2023. No Matter Small or Big Lip Motion: DeepFake Detection with Regularized Feature Learning on Semantic Information. In *International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, Singapore, 1–6. doi:10.1109/MIPR59079.2023.00034

[48] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, Shan Liang, Shiming Wang, Shuai Zhang, Xinrui Yan, Le Xu, Zhengqi Wen, and Haizhou Li. 2022. ADD 2022: the first Audio Deep Synthesis Detection Challenge. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Singapore, 9216–9220. doi:10.1109/ICASSP43922.2022.9746939

[49] Linjuan Zhang, Kong Aik Lee, Lin Zhang, Longbiao Wang, and Baoning Niu. 2024. CPAUG: Refining Copy-Paste Augmentation for Speech Anti-Spoofing. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Seoul, Korea, 10996–11000.

[50] You Zhang, Fei Jiang, and Zhiyao Duan. 2021. One-Class Learning Towards Synthetic Voice Spoofing Detection. *IEEE Signal Processing Letters* 28 (2021), 937–941.

[51] Cairong Zhao, Chutian Wang, Guosheng Hu, Haonan Chen, Chun Liu, and Jinhui Tang. 2023. ISTVT: interpretable spatial-temporal video transformer for deepfake detection. *IEEE Transactions on Information Forensics and Security* 18 (2023), 1335–1348.

[52] Yipin Zhou and Ser-Nam Lim. 2021. Joint Audio-Visual Deepfake Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Online, 14780–14789.