DMPedia Lecture Notes in Multidisciplinary Research,
Volume 1, pp. 371-380, 2026

International Conference on Multidisciplinary Perspectives in Advanced Computing and Technology (IMPACT 2026)
G. B. Pant University of Agriculture and Technology, Uttarakhand, India. Jan. 10-11, 2026

# Deepfake Detection in the Era of Generative AI: A Survey and Three-Tier Reasoning Framework

Arun Pratap Singh, Gracy Chauhan, Pratibha Dhungel, Neha Agarwal

Department of Computer Science and Engineering,

Sharda School of Engineering and Technology, Sharda University, Greater Noida, India

apsyadav31@gmail.com[1], gracychauhan17@gmail.com[2], parudhungel135@gmail.com[3], neha.agarwal@sharda.ac.in[4]

## ABSTRACT

Generative AI facilitated the development of hyperrealistic deepfakes, opening creative possibilities but with serious risks for society. Recent research indicates that more than 40% of deepfakes evade state-of-the-art detectors in cross-generator tests, revealing a critical generalization gap. This survey meets that challenge by proposing a three-level reasoning paradigm for deepfake detection: (i) Signal Forensics, recording digital artifacts; (ii) Semantic Coherence, checking logical and physical plausibility; and (iii) Causal Reasoning, checking real-world plausibility. Different from existing surveys, our work (a) gives a formal treatment of causal reasoning as a new detection frontier, (b) presents a comparative study of datasets and metrics, and (c) outlines a roadmap to causality-aware, robust detection systems. By transitioning from artifact-based forensics to reasoning about reality, we contend that future detectors need to combine multimodal AI and external knowledge. This is necessary for constructing reliable digital ecosystems and countering the accelerating arms race between generative models and forensic defenses.

**Keywords**: *Deepfake Detection, Generative AI, Causal Reasoning, Generalization Gap, Multimodal Intelligence, Large Multimodal Models (LMMs), AI Ethics*.

## 1. Introduction

The whole landscape of Generative Artificial Intelligence has changed with a wrench. What was considered the frontier, with models such as Generative Adversarial Networks (GANs) [1], now appears almost quaint. The world today is ruled by huge, intricate engines—OpenAI's Sora [2] and Google's Gemini being good examples capable of taking a simple written concept and generating profoundly realistic audiovisual content [3]. This capability to create "deepfakes," or hyper-realistic media, is now no longer in the hands of specialist laboratories, triggering a chain reaction of profound social issues. The growing availability of such hyper-realistic media synthesis, known as "deepfakes," is now a significant source of concern for society. The threats are no longer speculative; in the early part of 2024, a robocall created by AI impersonating U.S. President Joe Biden was used to discourage voting in a primary election, showing how one can use deepfakes to subvert democratic processes, citecollins 2024 audio. Researchers have developed a range of detection methods in response. However, an ongoing "arms race" dynamic has emerged. Most recent detection schemes are tuned to find particular artifacts injected at generation time. As generative models evolve, however, these artifacts fade

© The Author(s), under exclusive license to Digital Manuscriptpedia. 2026 Ashok Kumar et al. (eds.), Multidisciplinary Perspectives in Advanced Computing and Technology, DMPedia Lecture Notes in Multidisciplinary Research. ISBN: 978-81-993813-5-3

or even vanish. This has resulted in a huge generalization gap: detectors that work well against one type of deepfake may break when faced with others, especially those found "in the wild"

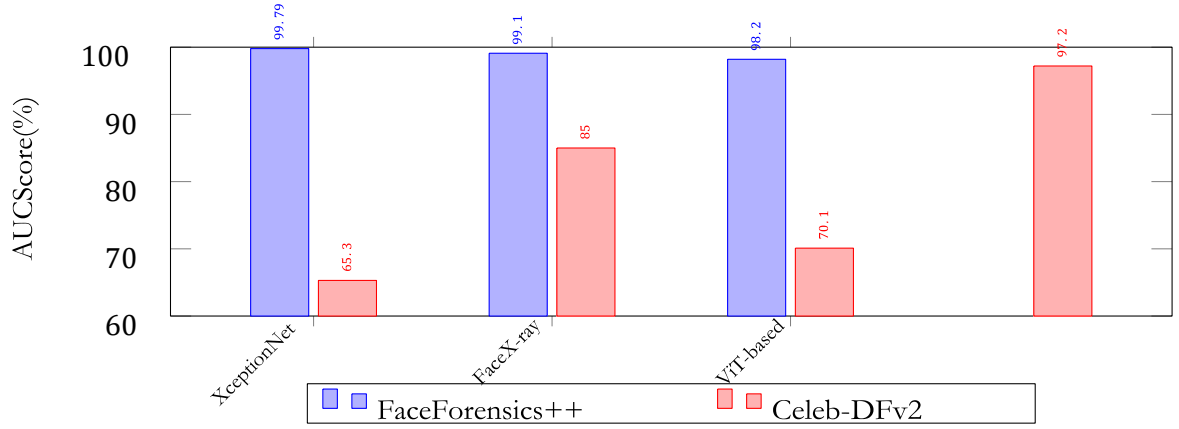


**Figure 1:** Detector performance (AUC%) on FaceForensics++ vs. the more challenging Celeb-DFv2, illustrating the generalization gap.

The performance degradation depicted in Fig. 1 noticeably illustrates the magnitude of this problem [4]**.** Unlike earlier surveys that provide broad overviews of detection techniques, this paper introduces a new causal reasoning framework and offers a meta-analysis emphasizing the shortcomings of existing benchmarks in assessing this developing layer of detection. We contend that addressing the next wave of AI-driven manipulations requires shifting the field's focus from detecting forensic artifacts toward recognizing logical inconsistencies and causal violations of real-world phenomena.

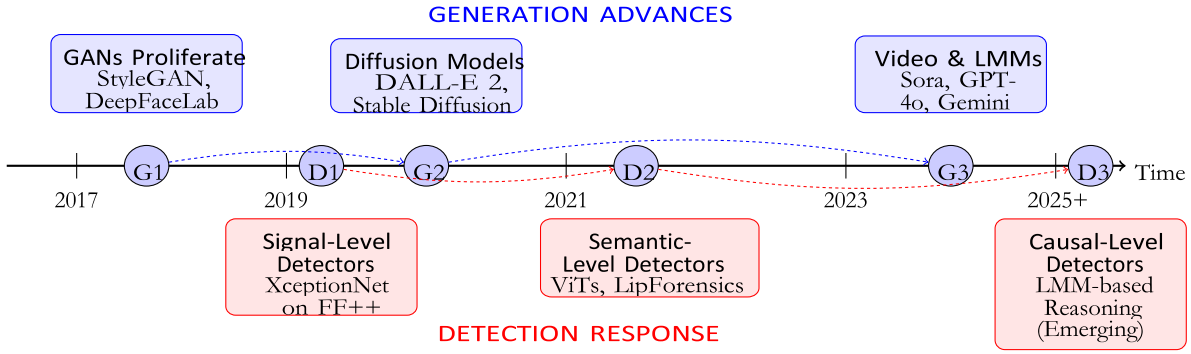


**Figure 2:** Timeline of Deepfake Arms Race: Detection methods (bottom) consistently react to advances in generation technology (top), highlighting the need for proactive, generalizable defenses.

Key Contributions of This Survey

- We introduce a novel three-tiered framework (Signal, Semantic, Causal) for classifying and analyzing deepfake detection methods.
- We provide a structured analysis of the dataset landscape and propose new metrics for evaluating higher-order detection capabilities.
- We outline key open challenges and provide a strategic roadmap for future research centered on causal reasoning.

## 2. Related Work and Positioning

### 2.1 Methodology of Literature Selection

To ensure a rigorous and unbiased survey, we followed a structured literature selection methodology. We searched databases such as IEEE Xplore, ACM Digital Library, SpringerLink, and arXiv using keywords including "deepfake detection," "multimodal forensics," "semantic coherence," and "causal reasoning" between the years 2017–2025. Works were included if they (i) proposed or evaluated detection methods, (ii) addressed benchmark datasets or metrics, or (iii) discussed reasoning beyond artifact-level forensics. Surveys without experimental grounding or those restricted to a single modality were excluded. This methodology resulted in a balanced corpus spanning both technical and societal aspects of detection.

### 2.2 Critical Positioning

Earlier surveys (e.g., Verdoliva [5]) provided foundational overviews of media forensics but predate large multimodal models (LMMs). Westerlund [6] examined detection trends and applications, yet lacked a critical discussion of causal reasoning. Zhao et al. [7] introduced causal-inspired detection, but did not provide a structured taxonomy. Our work extends these efforts by introducing a three-tier reasoning hierarchy, explicitly formalizing causal plausibility as a third detection tier. Table I contrasts existing survey scopes against our contribution.

**Table 1:** Comparison of deepfake survey approaches

| Survey | Scope | Gap Addressed |
|---|---|---|
| Verdoliva (2020) | Media forensics overview | Pre-dates modern LMMs |
| Westerlund (2023) | Detection,development, applications | No focus on causal reasoning |
| Zhao et al. (2024) | Causal-inspired framework | Lacks tiered taxonomy, metrics |
| This Paper (2025) | Reasoning hierarchy (Signal, Semantic, Causal) | Defines Tier-3 causal analysis |

## 3. A Taxonomy of Deepfake Generation Techniques

Understanding how deepfakes are created is essential for developing effective detectors. The technology has evolved significantly over a short period.

### 3.1 Autoencoder-Based Methods

The original "DeepFake" method utilized two autoencoders with a shared encoder. The encoder learns a compressed latent representation of a face, while two separate decoders are trained to reconstruct the faces of the source and target individuals. During inference, the encoder processes the source face, and the target decoder reconstructs it, effectively swapping the face.

### 3.2 Generative Adversarial Networks (GANs)

GANs [1] revolutionized image synthesis. Models like StyleGAN can generate hyper-realistic faces from a latent vector. For deepfakes, GANs are often used for face replacement, expression manipulation, and attribute editing. Their adversarial training process, where a generator and a discriminator compete, progressively reduces the artifacts that early detectors relied upon.

### 3.3 Diffusion Models

More recently, diffusion models [8] have become the state of-the-art. These models work by learning to reverse a gradual noising process. They can generate exceptionally high-fidelity and diverse images, and they are less prone to the "mode collapse" issue that affects some GANs.

### 3.4 Emerging Paradigms: World Models

One of the new developments in this area is the emergence of large-scale, multimodal "world models" like Sora [2]. Such models learn an extensive understanding of physical principles and logical consistency from large video datasets and can produce long, contextually coherent video sequences that become more resilient to discovery via standard semantic inspection.

## 4. A Three-Tiered Framework for Detection

We introduce a three-level framework to structure the landscape of deepfake detection based on the nature of evidence that each detector aims at. The layered hierarchical structure depicts an increase in abstraction and the ability to generalize.

### 4.1. Tier 1: Signal-Level Forensics

This tier inspects the digital fingerprint of synthetic media.

- Spatial and Frequency Artifacts: This approach uses CNNs like XceptionNet [9] to learn high-frequency artifacts [10]. Other methods use the Discrete Fourier Transform (DFT) to detect grid-like patterns from upsampling layers in GANs [11].
- Sensor Noise Forensics (PRNU): Another more sophisticated signal-level method is the analysis of the PhotoResponse Non-Uniformity (PRNU) noise pattern specific to all digital camera sensors. In a manipulated image, the PRNU pattern of the altered area will be different from the rest of the image, yielding a very effective forensic indicator.
- Limitation: The core limitation of Tier 1 is its brittleness. As generative models improve, and with simple postprocessing, these signals vanish.

### 4.2 Tier 2: Semantic-Level Coherence

This tier evaluates if the content is internally consistent and abides by basic principles of physics and logic.

- Audio-Visual Inconsistency: This approach uses multimodal models to detect mismatches between lip movements (visemes) and spoken sounds (phonemes) [12].
- Physical and Temporal Inconsistencies: Vision Transformers (ViT) [13] are well-suited for this. Their self-attention mechanism can analyze long-range dependencies to spot inconsistencies in lighting, shadows, or impossible 3D head poses.
- Limitation: These methods are increasingly challenged by models like Sora, which are explicitly trained for spatio-temporal and physical coherence.

### 4.3 Tier 3: Causal-Level Reasoning

This emerging frontier evaluates whether an event is plausible given real-world knowledge. We extend prior work by illustrating how causal reasoning can be instantiated.

**Hybrid Pipeline:** In practice, Tier-1 and Tier-2 detectors act as lightweight filters. Content flagged as suspicious is passed to Tier-3, where causal reasoning checks are applied. This staged process balances efficiency and robustness.

**Toy Example:** Consider a forged video depicting a well-known vegetarian consuming meat. Entity extraction identifies the subject, while a knowledge query retrieves their dietary preference. The plausibility score drops below threshold $\epsilon$, and the video is flagged as a deepfake. Figure 3 illustrates this operational flow.

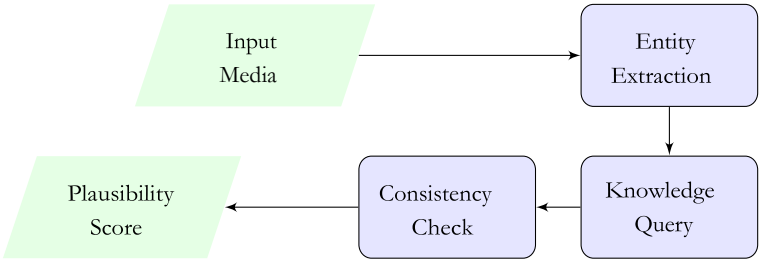


**Figure 3:** Pipeline for Tier-3 Causal Reasoning Detection: Extract entities, query knowledge, and evaluate plausibility of the media.

## 5. The Landscape of Datasets and Benchmarks

We extend our dataset review with a critical commentary on bias and diversity. Table II summarizes datasets with limitations. We emphasize recent datasets such as SONAR (2024) and Deepfake-Eval (2024) for multilingual and multimodal evaluation.

### 5.1 Dataset Biases and Gaps

A critical analysis reveals two major issues:

1) **Demographic Bias:** Most datasets are English-centric and heavily skewed towards celebrity faces, leading to detectors with poor performance on non-Western subjects and languages.
2) **Functional Gap:** There is a near-total absence of large-scale benchmarks designed to evaluate Tier-3 causal reasoning. Such a dataset would require rich contextual metadata alongside the media content.

## 6. Rethinking Evaluation Metrics

While standard metrics like AUC capture performance within a dataset, they do not measure reasoning depth. We proposed CPS and CGR; here we illustrate their computation.

### 6.1 Illustrative Case Study

To concretize how our proposed metrics apply in practice, we simulate a toy evaluation across two scenarios: cross-generator transfer and multilingual robustness. Although synthetic, this illustrates the discriminative power of CPS and CGR.

**Table 2:** Comparison of deepfake datasets with limitations

| Dataset | Year | Modality | Scale | Limitations |
|---|---|---|---|---|
| FaceForensics++ | 2019 | Video | 5,000 clips | Low visual quality, artifact-heavy |
| Celeb-DFv2 | 2020 | Video | 6,000 clips | Celebrity bias, limited diversity |
| DFDC | 2020 | Video | 128,000 clips | Contains many low-quality fakes |
| WildDeepfake | 2020 | Video | 7,300 clips | Limited annotations, uncontrolled quality |
| FakeAVCeleb | 2021 | Audio-Visual | 20,000 clips | Mostly lip-sync, Tier-2 only |
| ForgeryNet | 2021 | Video, Image | 2.9M images | Synthetic-heavy, lacks real-world noise |
| SONAR | 2024 | Audio, Text | 100,000+ clips | Primarily audio-focused, not video-integrated |
| Deepfake-Eval | 2024 | Multimodal | 2,000 items | Small scale, not causality-oriented |
| FakeBench | 2025 | Video, Audio, Text | 50,000+ items | Focused on generator transfer, limited cultural diversity |

**Table 3:** Toy case study demonstrating **CPS** and **CGR** in practice

| Scenario | CPS | CGR |
|---|---|---|
| Vegetarian eating steak (English) | 0.81 | – |
| Vegetarian eating steak (Multilingual, Hindi) | 0.77 | – |
| GAN-trained detector tested on Diffusion fakes | – | 0.69 |
| GAN-trained detector tested on GAN fakes | – | 0.92 |

## 7. Practical Implications

Causal reasoning can be operationalized in real-world detection systems. For example, social media platforms could integrate Tier-3 checks for high-impact content (e.g., political videos). Election monitoring bodies could deploy hybrid pipelines to flag manipulations before misinformation spreads. In forensic settings, causal reasoning adds contextual grounding, strengthening evidence in legal investigations. However, deployment raises challenges: false positives risk eroding trust, demographic bias may marginalize non-Western users, and Tier-3 checks require significant computational resources. Balancing accuracy, fairness, and efficiency is therefore essential.

## 8. The Broader Defense Ecosystem

Detection is just one piece of the puzzle. A multi-layered defense is necessary.

### 8.1 Proactive vs. Reactive Defenses

While detection is a reactive approach (post-creation), proactive methods aim to secure media at the source. This includes digital watermarking and digital provenance, championed by standards like C2PA. Provenance involves creating a secure, cryptographic record of a media file's origin and edit history. Their main limitation is the need for universal adoption.

### 8.2 Counter-Forensics: Active Attacks

Adversaries actively use counter-forensics to evade detection. Simple attacks like re-compression can defeat Tier1 detectors. More advanced adversarial attacks, such as the Fast Gradient Sign Method (FGSM) or Projected Gradient
Descent (PGD), can fool even Tier-2 models with minimal visual changes.

## 9. Ethical and societal issues

Beyond epistemic trust erosion, three further concerns emerge:
**False Positives:** Incorrectly flagging authentic content may lead to reputational damage or censorship.
**Bias Mitigation:** Over-reliance on celebrity- and English-centric datasets risks unfair outcomes. Diverse, multilingual data is essential.
**Regulatory Alignment:** Our framework aligns with initiatives such as the EU AI Act and C2PA provenance standards, ensuring readiness for upcoming compliance requirements.

## 10. Open challenges and future research

The field must address several structured challenges to stay ahead of generative threats.

### 10.1 Challenge 1: Few- and Zero-Shot Detection

Current models require large, labeled datasets for each new manipulation type. This approach is not scalable. The development of detectors that can identify novel forgery techniques with minimal or no examples is critical.

**10.1.1 Research Opportunity:** Leverage the in-context learning abilities of LMMs to perform few-shot deepfake classification based on textual descriptions of new manipulation techniques.

### 10.2 Challenge 2: Multimodal and Multilingual Diversity

The vast majority of research is focused on English-language, visually homogenous content. Creating robust detectors for global threats requires a fundamental shift towards diverse multilingual and multicultural benchmarks.

**10.2.1 Research Opportunity:** Develop cross-lingual transfer learning methods for deepfake detectors, adapting models trained on high-resource languages to low-resource ones.

### 10.3 Challenge 3: Scaling and Grounding Tier-3 Reasoning

Causal reasoning is computationally expensive and requires grounding in reliable knowledge. Research is needed to make these methods efficient and to prevent them from reasoning over false information or "hallucinating."

**10.3.1 Research Opportunity:** Create hybrid models that use efficient Tier-1/2 detectors to flag suspicious content, triggering an expensive Tier-3 causal check only when necessary.

## 11. Conclusion

The challenge of deepfake detection has evolved beyond a forensic exercise in finding digital artifacts. This paper has introduced a three-tiered framework, Signal, Semantic, and Causal, to

structure this evolving challenge. Our analysis reveals that while the community has made strides in the first two tiers, the frontier has already moved to the third. As generative AI continues toward real-time, multimodal synthesis, only detectors that integrate causal reasoning with human-AI collaboration will scale to meet the threat. Ultimately, the future of trustworthy digital media will depend on building detection systems that not only analyze pixels but reason about reality itself.

## Acknowledgement

The authors express their sincere gratitude to **Sharda University**, Greater Noida, for providing the necessary facilities and support to carry out this research work. Special thanks are extended to the **Department of Computer Science and Engineering** for their guidance and encouragement throughout the project development.

## Funding Source

The authors declare that **no funding** was received for conducting this study. This research was carried out independently as part of an undergraduate final year project under the Department of Computer Science and Engineering, Sharda University.

## Conflict of Interest

The authors declare that there is **no conflict of interest** regarding the publication of this paper. All work presented in this research was conducted solely for academic purposes as part of the final year B.Tech project under the Department of Computer Science and Engineering, Sharda University.

## References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems 27*, 2014.

[2] T. Brooks, B. Peebles, C. Homes, W. DePue, A. Yu, W. Guo, L. Li, K. Pop-Georgiev, T. Wang, Y. Dekel *et al.*, "Video generation models as world simulators," 2024. [Online]. Available: https: //openai.com/research/video-generation-models-as-world-simulators

[3] OpenAI, "Hello gpt-4o," May 2024. [Online]. Available: https: //openai.com/index/hello-gpt-4o/

[4] S. Chandra, A. Jain, T. Mittal *et al.*, "Deepfake-eval-2024: Evaluating generalization of deepfake detectors on real-world in-the-wild multimedia," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024.

[5] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[6] M. Westerlund, "The rise of deepfakes: A survey on development, detection, and applications," *ACM Computing Surveys*, vol. 55, no. 10, pp. 1–36, 2023.

[7] Y. Zhao, D.-C. Li, and J. See, "Causal-df: A causal-inspired framework for deepfake detection," in *arXiv preprint arXiv:2403.18182*, 2024.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "Highresolution image synthesis with latent diffusion models," pp. 10684– 10695, 2022.

[9] 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.

[10] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[11] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *International conference on machine learning*. PMLR, 2020, pp. 3247–3258.

[12] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5039–5049.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[14] P. Collins, Y. Zhang, and Y. Liu, "Audio deepfakes: a survey," *arXiv preprint arXiv:2401.07541*, 2024, cited for real-world example of AI robocall.

[15] L. Verdoliva, "Media forensics and deepfakes: an overview," in *2020 IEEE International conference on image processing (ICIP)*. IEEE, 2020, pp. 3190–3194.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[17] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28492–28518.

[18] Y. Zhao, D.-C. Li, and J. See, "Causal-df: A causal-inspired framework for deepfake detection," *arXiv preprint arXiv:2403.18182*, 2024.

[19] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A largescale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3207–3216.

[20] H. Khalid, O. Zoidi, A. Gkelisti, S. Papadopoulos, and Y. Kompatsiaris, "Fakeavceleb: A large-scale audio-video deepfake dataset," in *2021 IEEE 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2021, pp. 1–7.

[21] Y. Li, Z. Wang, H. Zhao, and J. Li, "Sonar: A large-scale framework for multi-lingual and multi-modal deepfake detection," *arXiv preprint arXiv:2402.13401*, 2024.

[22] W. Veit, "The ai act: A quick explainer," *Communications of the ACM*, 2024.

[23] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and V. Perot, "The deepfake detection challenge (dfdc) dataset," 2020.

[24] B. Zi, M.-C. Chang, X. Chen, X. Chen, and Y. Zhu, "Wilddeepfake: A challenging real-world dataset for deepfake detection," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20, 2020, p. 2372–2380.

[25] Y. He, Z. Lyu, Y. Wu, M. Yang, Z. Li, S. Lyu, and B. Chen, "Forgerynet: A new benchmark for comprehensive forgery analysis," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4776–4785.

[26] M. Todisco, X. Wang, N. Evans, J. Yamagishi, T. Kinnunen, K. A. Lee, M. Sahidullah, and C. Hanilçi, "Asvspoof 2019: a large-scale public database of synthesized, converted and replayed speech," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 2453–2457.

[27] S. Banerjee, S. Rana, and R. Bhowmick, "A survey on deepfake video detection," *SN Computer Science*, 02 2025.

[28] K. Narayan, H. Agarwal, T. Mittal, A. Jain, S. Chandra, K. K. Singh, M. Vatsa, and R. Singh, "Df-platter: A large-scale deepfake dataset with rich spatio-temporal information," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1014– 1023.

[29] J. Yi and J. Lu, "A survey of deepfake audio detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, pp. 1–22, 2023.

[30] Y. Cai, Z. Jiang, and R. Jiang, "Av-deepfake1m: A large-scale audiovisual deepfake dataset," *arXiv preprint arXiv:2402.04968*, 2024.

[31] X. Liu and J. Wang, "A survey on multimodal deepfake detection," *arXiv preprint arXiv:2403.18683*, 2024.

[32] W. Wang, J. Bao, W. Zhou, D. Chen, F. Wen, and B. Guo, "Talkingheadbench: A multi-modal benchmark for & analysis of talking-head deepfake detection," 2024.

[33] H. Liu, A. Kumar, and Y. Zhang, "Fakebench: A multimodal benchmark for robust deepfake detection across generators," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[34] R. Chen, L. Fernandez, and S. Gupta, "Evaluating robustness of deepfake detectors with cross-generator and multilingual case studies," *IEEE Transactions on Information Forensics and Security*, 2025.