Proceedings of CONF-SEML 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA31921

From Deep Learning to Large Models: Detection Methods for AI-Generated Content and Deepfakes

Yifei Zeng

School of Mathematics, University of Edinburgh, Edinburgh, UK
yifei816@163.com

Abstract. The rapid evolution from traditional deep learning to large-scale foundation models has revolutionized content gen- eration while introducing unprecedented challenges for detecting AI-generated content (AIGC) and deepfakes. This paper comprehensively analyzes the detection technologies across the paradigm shift from GANs to modern diffusion models and large language models (LLMs). We systematically categorize detection methods across generation eras, modalities (text, image, video, audio), and technical approaches. Our analysis reveals that traditional methods designed for GAN-generated content exhibit catastrophic failure on modern diffusion and LLM outputs. We examine state- of-the-art techniques such as watermarking schemes, zero- shot detection methods like DetectGPT, reconstruction-based ap- proaches (DIRE), and multimodal verification. Furthermore, we analyze critical challenges related to cross-model generalization, adversarial robustness, and computational efficiency, providing practical guidance for method selection and deployment.

Keywords: AI-Generated Content, Deepfake Detection, Large Language Models, Diffusion Models, Zero-shot Detection

1. Introduction

The landscape of content generation has been fundamentally transformed by artificial intelligence, which is marked by a decisive transition from traditional deep learning models to large- scale foundation models, a process that has redefined the boundaries of creative and informational production. Early generative models based on Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) demonstrated impressive capabilities but faced limitations in training instability and mode collapse [1,2]. The emergence of transformer architectures catalyzed the development of a new era in AI, leading to the creation of Large Language Models (LLMs) such as GPT-4 and powerful diffusion models like Stable Dif- fusion, which can produce content that is virtually indistinguishable from authentic media [3-5].

While these advances enable exciting applications across creative industries, education, and research, they pose significant threats to information integrity and social trust. In politics, synthetic media has been weaponized to enable disinformation campaigns and election interference [6]. In finance, deep- fake videos and voice cloning have become tools for fraud activities [7]. Academia faces challenges as LLM-generated essays undermine academic integrity [8]. Thus, the proliferation

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.BA31921

of AIGC necessitates robust detection mechanisms. Traditional detection approaches for GAN-generated images face challenges when applied to modern diffusion models and LLMs, including issues of generalization, robustness, efficiency, and explainability. This paper offers a comprehensive review of detection methods across various modalities, a multi-dimensional taxonomy categorizing methods by era and approach, an analysis of differences between traditional and large-scale model detection, and practical insights for performance comparisons and deployment strategies.

2. Background and evolution

2.1. Traditional deep learning era (Pre-2022)

GANs. GANs revolutionized generative modeling a process of adversarial training between a generator G and a discriminator D:

$$\min_{G} \max_{D} \mathbb{E}_{x \sim p_{data}} \left[\log D(x) \right] + \mathbb{E}_{z \sim p_z} \left[\log \left(1 - D\left(G\left(z \right) \right) \right) \right] \tag{1}$$

Notable architectures include DCGAN , StyleGAN , and BigGAN, though they suffered from training insta- bility and mode collapse [9-11].

Early LLMs. Autoregressive models like GPT-2 demonstrated impressive text generation but were limited by scale and computational constraints [12].

2.2. Large model era (2022-present)

LLMs. The scaling of transformer - based models has led to significant breakthroughs. GPT-3 with 175B parameters demonstrated few-shot learning across diverse tasks [13]. ChatGPT and GPT-4 further improved instruction-following abilities, generating highly coherent text that is difficult to distinguish from human writing [4].

Diffusion Models. Diffusion models , emerged as the dominant paradigm for image generation [5,14]. The forward process gradually adds noise:

$$q\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I\right) \tag{2}$$

The reverse process learns to denoise, enabling photorealistic generation with fine-grained semantic control in models like Stable Diffusion, DALL-E 2, and Midjourney.

2.3. Task definition

AIGC detection is formulated as binary classification: given input (text, image, video, or audio), determine whether was AI-generated or human-created:

$$f : \mathcal{X} \rightarrow \{0,1\} \tag{3}$$

Detectors typically output continuous scores repre- senting probability of being AI-generated. For certain applica- tions, fine-grained detection identifies the specific generative model.

DOI: 10.54254/2755-2721/2026.BA31921

3. Datasets and evaluation

High-quality benchmarks are crucial for developing detection methods, as summarized in various datasets across modalities. In text, HC3 compares human and ChatGPT outputs, while M4 provides multilingual samples from different LLMs, and DetectRL simulates real-world scenarios with adversarial attacks. For images, GenImage features 1M images from text-to-image models, AIGCDetect offers comprehensive benchmarks, and Chameleon contains curated images with minimal artifacts. In video, FaceForensics++ includes 1.8M frames with manipulation methods, while DFDC presents 128K diverse videos, and WildDeepfake gathers real-world deepfakes. Audio benchmarks include the ASVspoof series for voice synthesis detection and WaveFake for neural vocoder detection. Evaluation metrics focus on accuracy, AUC-ROC, false positive/negative rates, and robustness under perturbations, alongside generalization across datasets and models.

4. Detection methods

4.1. Traditional era methods

Before the era of large models, detection efforts focused on GAN-generated images and early deepfakes.

CNN-based Detection. Convolutional networks learned discriminative features from labeled datasets. XceptionNet [15] achieved ¿95% accuracy on FaceForensics++. However, these methods demonstrate poor generalization to unseen generators.

Frequency Analysis. GANs present distinct frequency- domain patterns [16]. By applying Fourier or DCT transforms, detectors identified high-frequency artifacts. While effective for cross-GAN generalization, these signals diminish with mod- ern diffusion models.

Biological Priors. Early deepfake detection leveraged biological signals such as eye blinking patterns and remote photo- plethysmography (rPPG) for cardiac signals . Face swap- ping often disrupts these natural patterns, offering detection cues.

4.2. Large model era methods

Modern detection requires fundamentally new approaches for high-quality LLM and diffusion model outputs.

4.2.1. LLM-generated text detection

Watermarking. Proac- tive approaches embed signatures during the generation process. KGW watermark partitions the vocabulary into "green" and "red" tokens, boosting the logits of green tokens by a constant [17]:

$$ {}_{\mathrm{t}}^{\prime}\left(\mathrm{w}\right) = \begin{cases} {}_{\mathrm{t}}(w) + \delta & \text{if } w \in \mathscr{G}_{\mathrm{t}} \\ {}_{\mathrm{t}}(w) \text{ otherwise} \end{cases} \tag{4} $$

Detection involves computing the proportion of green tokens. While highly reliable, watermarking requires generator side cooperation and may affect generation quality.

Statistical Methods. Training-free approaches analyze dis- tributional properties. Perplexity-based detection observes that LLM-generated text exhibits lower perplexity when evaluated by the same

DOI: 10.54254/2755-2721/2026.BA31921

model family. LLMDet analyzes token proba- bility distributions across n-grams, achieving 98.5% accuracy with fine-grained attribution.

Zero-shot Detection. DetectGPT revolutionized the field by requiring no training [18]. It exploits the observation that LLM-generated text occupies negative curvature regions in log probability space:

$$\mathrm{d}(\mathrm{x}) = \log \mathrm{p}(\mathrm{x}) - \mathbb{E}_{\tilde{\mathrm{x}} \sim \mathrm{P}_{\mathrm{perturb}}(\mathrm{x})}\left[\log \mathrm{p}(\tilde{\mathrm{x}})\right] \tag{5}$$

Fast-DetectGPT reduces computational cost by 75% while maintaining accuracy, making zero-shot detection practical.

Table 1. Representative datasets for aigc and deepfake detection across different modalities. Scale indicates approximate number of samples

Modality	Dataset	Year	Scale	Generation Method	Key Features
Text	HC3	2023	87K	ChatGPT	Multi-domain, human comparison
	M4	2023	150K	GPT-3.5/4, LLaMA	Multilingual, diverse LLMs
	DetectRL	2024	60K	Multiple LLMs	Adversarial attacks, real-world
	RAID	2024	50K	Multiple LLMs	Academic writing focus
Image	FaceForensics++	2019	1.8M frames	GAN-based	Four manipulation methods
	GenImage	2023	1M	Diffusion models	Eight generators, diverse
	AIGCDetect	2023	200K	Diffusion models	Comprehensive benchmark
	Synthbuster	2023	9K	Nine diffusion models	Photorealistic focus
	Chameleon	2024	10K	Advanced diffusion	Minimal artifacts, challenging
Video	DFDC	2020	128K	Various methods	Largest scale, diverse
	Celeb-DF	2020	6K	Deepfake	High-quality synthesis
	DeeperForensics	2020	60K	Deepfake	Controlled variations
	WildDeepfake	2021	7K	Real-world collected	In-the-wild scenarios
Audio	ASVspoof 2019	2019	100K	TTS, VC	Logical access attacks
	ASVspoof 2021	2021	600K	Various methods	Physical access extended
	WaveFake	2021	100K	Neural vocoders	Vocoder-specific
Multimodal	FakeAVCeleb	2021	20K	Audio-visual	Synchronized A-V manipulation
	DF-Platter	2023	77K	Heterogeneous	Multiple techniques combined

4.2.2. Diffusion model image detection

Reconstruction- based. DIRE utilizes pre-trained diffusion models to reconstruct input images [19]. Diffusion-generated images display significantly lower reconstruction error:

$$\mathrm{DIRE}(\mathrm{x}) = \|\mathrm{z}_0 - \hat{\mathrm{z}}_0\|_2 \tag{6}$$

where is the original latent and is reconstructed. DIRE demonstrates strong generalization across unseen diffusion models and maintains robustness to perturbations.

Frequency Analysis. Despite enhanced spatial realism, diffusion models present frequency-domain anomalies. DEFEND examines deviations across frequency bands, highlighting discriminative components.

Deep Learning. CLIP-based detection harness powerful pre-trained representations for semantic understanding. Vision Transformers capture global context through self- attention, outperforming

DOI: 10.54254/2755-2721/2026.BA31921

CNNs in the task of diffusion-generated image detection.

4.2.3. Video and audio detection

Video. Spatiotemporal transformers jointly model spatial appearance and temporal dynamics via 3D attention over temporal windows. Audio- visual synchronization analysis identifies lip-sync inconsisten- cies in talking face videos.

Audio. Self-supervised methods like Wav2Vec undergo fine-tuning to yield robust features [20]. AASIST achieves state-of-the- art performance by employing graph attention networks for spectro-temporal analysis.

5. Challenges and analysis

Current AI-generated content detection confronts several challenges: the generalization gap, where detectors trained on specific models show reduced accuracy (from 99% to 55%) when facing new generators; adversarial robustness issues that allow minimal perturbations to mislead detectors; and computational efficiency hurdles, as current methods are too resource-intensive for real-time use. Ethical considerations, such as privacy, fairness, transparency, and potential misuse of detection tools, also play a vital role. Additionally, there is an ongoing arms race between detection and generation technologies, necessitating strategies like watermarking and provenance tracking to maintain effectiveness (Follow Table 2).

Table 2. Comprehensive comparison of aigc and deepfake detection methods. ratings: +++ (excellent), ++ (good), + (fair), - (poor). Accuracy reported on primary benchmarks

Method	Representative Work	Year	Accuracy	General-ization	Robust-ness	Efficiency	Key Characteristics
Text Detection (LLM-Generated)							
Watermarking	KGW [17]	2023	99%+	+++	++	+++	Proactive; requires generation-side access
	Unigram [?]	2023	98%	+++	+++	+++	Robust to editing; prov- able guarantees
Statistical	Perplexity-based [?]	2019	75-85%	++	+	+++	Training-free; high false positives
	Entropy Analysis [?]	2008	70-80%	+	+	+++	Fast but unreliable on formal text
	LLMDet [?]	2023	98.5%	+++	++	++	Fine-grained attribution; efficient
Neural-based	RoBERTa Fine- tuned [?]	2019	95%	+	+	++	Strong in-domain; poor generalization
	DetectGPT [18]	2023	90-95%	+++	++	+	Zero-shot; expensive in- ference
	Fast-DetectGPT [?]	2023	92%	+++	++	++	4 × faster than Detect-GPT
	Intrinsic Dim. [?]	2024	93%	+++	+++	++	Best cross-domain; ro- bust to attacks
Image Detection (Diffusion Model-Generated)							
Reconstruction	DIRE [19]	2023	92-96%	+++	+++	+	Training-free; generalizes well
	Latent Analysis [?]	2024	94%	++	++	++	Faster than full recon- struction

DOI: 10.54254/2755-2721/2026.BA31921

Table 2. (continued)

Frequency	Fourier Analysis [16] DEFEND [?] Wavelet-based [?]	2020 2024 2020	88-92% 95% 85-90%	++ +++ ++	++ +++ +	+++ ++ ++	Universal artifacts; lightweight Multi-band analysis; ro- bust Multi-resolution features
Deep Learning	CLIP-based [?] Vision Trans. [?] CNN-based [?]	2023 2023 2020	97% 96% 99%	+++ ++ +	++ + -	++ + +++	Semantic features; mini- mal training Global context; compu- tationally heavy Strong in-domain; poor generalization
Hybrid	Multi-scale Fusion [?]	2024	98%	+++	+++	+	Combines spatial, freq., semantic
			Video Detection (Deepfake)				
Traditional CNN	XceptionNet [?] MesoNet [?]	2019 2018	95% 92%	+ +	+ -	++ +++	Spatial artifacts; Face- Forensics++ Lightweight; mesoscopic features
Spatiotemporal	3D CNN [?] Spatio-Temp. Trans. [?]	2018 2024	93% 97%	+ ++	+ ++	+ +	Temporal consistency Joint spatial-temporal modeling
Multimodal	Audio-Visual Sync [?] Emotion Consist. [?]	2020 2020	94% 91%	++ ++	++ +	++ ++	Lip-sync analysis Cross-modal verification
			Audio Detection (Voice Synthesis)				
Spectral	WaveFake [?] Mel-spectrogram CNN [?]	2021 2021	90-95% 93%	++ +	+ +	++ ++	Neural vocoder detection End-to-end training
Self-supervised	Wav2Vec Fine- tuned [20] AASIST [?]	2020 2022	96% 98%	++ +++	++ ++	+ ++	Robust representations ASVspoof champion; graph attention
			Traditional Era (For Comparison)				
GAN Detection	CNN Freq. Features [?] Frequency Analysis [16]	2020 2020	99% 97%	- ++	- +	+++ +++	Excellent for GANs; fails on diffusion Cross-GAN generaliza- tion

6. Conclusion

The evolution from traditional deep learning to large foun- dation models has reshaped both generative capabilities and detection challenges. This paper presents a comprehensive analysis, highlighting the paradigm shift from GAN-era to LLM/diffusion-era detection. Key findings include: Traditional detectors experience catastrophic failure when applied to modern large - model outputs, necessitating the development of architecture - specific solutions. No single method dominates—practitioners must navigate trade - offs among accuracy, generalization, robustness, and effi- ciency. Detection systems remain susceptible to adversarial attacks and perturbations. The interplay between generative and discriminative models constitutes a perpetual adversarial cycle, demanding continuous innovation to maintain diagnostic validity.

Success hinges on robust methods generalizing across models and modalities, deployment strategies integrating detection into content platforms, and ethical frameworks ensuring tech- nologies serve societal interests. The ultimate goal is not to eradicate AI - generated content but to establish trustworthy information ecosystems where synthetic and authentic content can be reliably

DOI: 10.54254/2755-2721/2026.BA31921

distinguished. Continued innovation, ethical consider- ation, and collective commitment to responsible AI develop- ment will be indispensable for nurturing credible digital environments.

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, 2014, pp. 2672– 2680.

[2] D. P. Kingma and M. Welling, "Auto-encoding variational bayes, " arXiv preprint arXiv: 1312.6114, 2013.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł . Kaiser, and I. Polosukhin, "Attention is all you need, " Advances in Neural Information Processing Systems, vol. 30, 2017.

[4] OpenAI, "Gpt-4 technical report, " arXiv preprint arXiv: 2303.08774, 2023.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High- resolution image synthesis with latent diffusion models, " in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni- tion, 2022, pp. 10 684–10 695.

[6] R. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security, " California Law Review, vol. 107, p. 1753, 2019.

[7] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey, " ACM Computing Surveys, vol. 54, no. 1, pp. 1–41, 2021.

[8] D. R. Cotton, P. A. Cotton, and J. R. Shipway, "Chatting and cheating: Ensuring academic integrity in the era of chatgpt, " Innovations in Education and Teaching International, pp. 1–12, 2023.

[9] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks, " arXiv preprint arXiv: 1511.06434, 2015.

[10] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks, " in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

[11] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis, " arXiv preprint arXiv: 1809.11096, 2018.

[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners, " OpenAI blog, vol. 1, no. 8, p. 9, 2019.

[13] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language models are few-shot learners, " Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.

[14] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models, " Advances in Neural Information Processing Systems, vol. 33, pp. 6840– 6851, 2020.

[15] A. Ro… ssler, 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.

[16] J. Frank, T. Eisenhofer, L. Scho… nherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recog- nition, " pp. 3247–3258, 2020.

[17] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A watermark for large language models, " pp. 17 061–17 084, 2023.

[18] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "Detectgpt: Zero-shot machine-generated text detection using probability curvature, " pp. 24 950–24 962, 2023.

[19] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "Dire for diffusion-generated image detection, " pp. 22 445–22 455, 2023.

[20] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A frame- work for self-supervised learning of speech representations, " Advances in Neural Information Processing Systems, vol. 33, pp. 12 449–12 460, 2020.