Discovery of Deepfakes in Art

Alexander I. Iliev[1, 2][0000-0002-4220-3637]

[1] SRH University of Applied Sciences, Ernst-Reuter-Platz 10, 10587 Berlin, Germany
[2] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,
Acad. Georgi Bonchev Str., Block 8, 1113 Sofia, Bulgaria
ailiev@berkeley.edu

Abstract. The proliferation of deepfakes—AI-generated or manipulated media—has transformed the landscape of contemporary art. Deep generative models, including GANs, VAEs, diffusion models, and Transformers, have enabled artists to explore new creative realms while simultaneously raising critical questions around authenticity, ethics, and detection. This paper presents a comprehensive analysis of deepfake technologies across five key media modalities: image, video, text, speech, and music. We examine the architectures that enable content creation, and the state-of-the-art techniques used for detection. Further, we evaluate detection accuracy, robustness, and practical implementation, incorporating diagrams, comparative tables, and performance formulas. This work aims to provide a balanced perspective on the opportunities and challenges posed by synthetic media in the artistic domain.

Keywords: Deepfakes, Generative Models, Art Forensics, GANs, Transformers, Multimedia AI.

1 Introduction

In the evolving intersection of artificial intelligence and art, deepfakes have emerged as one of the most provocative forces reshaping creative practice. Initially synonymous with digital deception, deepfakes—AI-generated or manipulated images, sounds, videos, and texts—are now at the forefront of both artistic innovation and ethical debate. Their growing presence in contemporary art and media challenges long-standing notions of authorship, authenticity, and aesthetic value.

Recent breakthroughs in generative architectures such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Transformer-based models have enabled the creation of synthetic content with unprecedented realism. Artists, musicians, writers, and filmmakers are leveraging these tools not only to replicate existing styles or voices but to explore new artistic territories—reanimating lost voices, generating fictional dialogues, and blending genres in ways that were once inconceivable (Chakraborty et al., 2024; Jain & Aggarwal, 2024). In this context, deepfakes function

less as forgeries and more as collaborative instruments of expression, giving rise to a new kind of hybrid creativity.

This paper examines the transformative role of deepfakes in the creative arts by exploring both their technical foundations and cultural consequences. Focusing on image, video, sound, speech, and textual media, we analyze how AI-generated content is redefining the creative process while also prompting a reevaluation of authenticity and ethical responsibility.

2 Media Deepfake Generative Technique

Image and Visual Art Generation. Image-based deepfakes are most produced using Generative Adversarial Networks (GANs), with notable variants such as DCGAN, Progressive GAN, CycleGAN, and especially StyleGAN, including its newer versions (StyleGAN2 and StyleGAN3). These models can generate high-resolution, photorealistic synthetic faces that allow detailed manipulation of attributes such as age, emotion, or lighting while maintaining identity coherence. GANs function via a competitive framework between two neural networks: a generator "G" and a discriminator "D", as described in Equation 1.

$$\min_{G}\max_{D}\mathcal{V}(D,G) = \mathrm{E}_{x\sim pdata(x)}\left[\,LogD(x)\right] + E_{z\sim pz_{(z)}}[\log\left(1-D\big(G(z)\big)\right)] \qquad (1)$$

The generator learns to produce convincing images, while the discriminator attempts to distinguish them from real images (Babael et al. 2025). Over time, the generator improves its outputs to fool the discriminator more effectively. In addition to GANs, autoencoders and Vision Transformers (ViTs) are employed for facial reanimation and style transfer tasks. These work by encoding an image into a low-dimensional latent representation and reconstructing it back with modified features. These techniques are particularly valuable in artistic applications due to their ability to manipulate and reconstruct visual content with contextual awareness. More recently, diffusion models have emerged as a powerful alternative, capable of producing high-fidelity images by iteratively denoising a random noise signal

Video Art Generation. Video deepfakes involve the synthesis of moving visual content using artificial intelligence to manipulate faces, expressions, and entire identities across time. The most sophisticated generation pipelines for video rely on GAN-based frameworks, specifically Hybrid Attention GANs (HA-GAN), which employ both spatial and temporal attention modules to ensure frame-by-frame consistency (Chakraborty et al., 2024). These features are essential in artistic video productions where uniformity in lighting, emotion, and facial identity enhances immersion. They may also be used in studies related to preventing aggressive behaviors in the future (Ignatova, 2025). HA-GAN integrates facial detail preservation through spatial attention and applies a temporal regularization loss to maintain consistency across sequences. This is particularly valuable in performance-based deepfake applications—such as digital acting and virtual storytelling—where continuity errors can disrupt the viewer's experience.

Text Generation. The emergence of Transformer-based language models has fundamentally transformed how textual deepfakes are generated. Unlike traditional rule-based or statistical methods, these models rely on self-attention mechanisms and massive pretraining corpora to produce coherent, contextually appropriate, and often indistinguishably human-like text. Models like GPT-2, GPT-3, and newer variants underlie the current surge in high-quality AI-generated writing. In the creative arts, these models are used to simulate fictional interviews, generate poetry, replicate literary voices, and even simulate text messages or dialogues in interactive media environments. For instance, Vasudeva et al. (2024) describe how current models can emulate historical or fictional styles, allowing for the simulation of Shakespearean prose, presidential speeches, or narrative monologues that fit seamlessly into games or digital films.

Audio and Speech Art Generation. Audio and speech deepfakes involve the synthetic generation of human-like speech, voices, or soundscapes using AI. With major advancements in text-to-speech (TTS) and voice cloning technologies, it is now possible to replicate not only the words a person might say but also their accent, intonation, and emotional nuance. One of the most advanced speech synthesis systems described is UC-VITS (Universal Cross-lingual Variational Inference TTS), a neural architecture capable of generating speech in multiple languages while preserving the speaker's unique vocal identity (Zhou et al., 2025). This model combines variational inference and Transformer-based TTS layers to generate natural-sounding, emotionally modulated speech from a small amount of reference data.

According to Zhou et al. (2025), UC-VITS achieved a MOS score of 4.51, outperforming traditional vocoders and even some state-of-the-art neural TTS systems in blind listening tests. Another application mentioned in the study is multilingual speech synthesis for voiceovers, where UC-VITS is used to render a single actor's voice across multiple languages with seamless fidelity (Zhou et al., 2025).

3 Media Deepfake Detection Techniques

Image Detection Techniques. To combat the increasing realism of AI-generated images, researchers have developed a range of sophisticated detection models. The most common starting point is the convolutional neural network (CNN), which detects local inconsistencies in textures, edges, and illumination that often arise from synthetic generation. Models such as XceptionNet, ResNet, and EfficientNet have shown impressive accuracy on public datasets like FaceForensics++.

However, CNNs alone sometimes struggle to generalize to unseen generative methods. Therefore, researchers incorporate frequency domain analysis to catch hidden patterns invisible in pixel space. Notably, Wavelet-Packet Decomposition (WPT) and Discrete Fourier Transform (DFT). These methods identify unnatural high-frequency artifacts introduced during GAN upsampling, such as checkerboard patterns and subtle color distortions. Figure 1 would depict how WPT identifies discrepancies in image textures that CNNs might overlook.

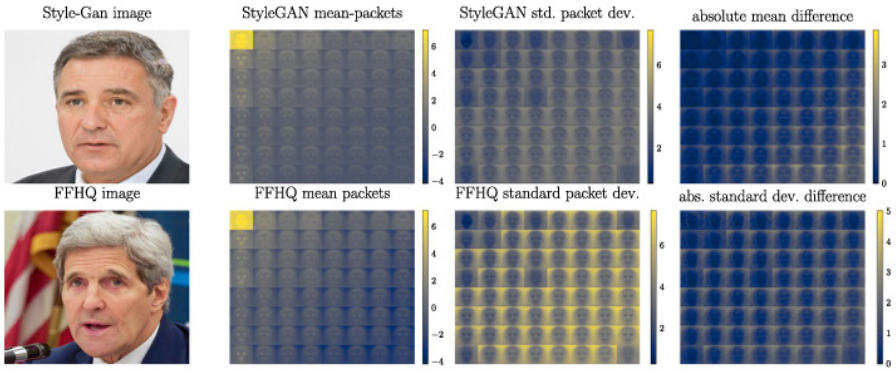

Fig. 1. Differences in wavelet-packet means between real and fake images (Wolter et al., 2022).

Hybrid architectures further improve detection robustness by combining spatial and frequency-based analysis. For instance, CIFAKE uses saliency maps to visualize the areas contributing most to the "fake" classification. Meanwhile, DIRE employs contrastive learning to distinguish images from real and diffusion-generated sources, achieving better generalization. One of such hybrid novel architecture combines a CNN with a spectral analysis channel, feeding both raw images and their Fourier-transformed versions into the network. This dual-stream approach significantly improves generalization across different manipulation types.

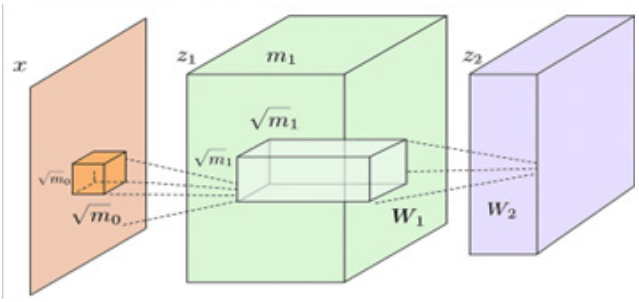

Fig. 2. A simplified Hybrid CNN Architecture example.

This detection system uses a binary cross-entropy loss, optimized with the Adam optimizer. Equation 2 shows the loss function for a batch of 'n' samples with expected probability and true labels and the Adam optimizer (learning rate = 1e-4, beta_1 = 0.9, beta_2 = 0.999) used to optimize the parameters in this model.

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}(Y_i \,.\, logY_i + (1 - Y_i).\log(1 - Y_i)) \qquad (2)$$

To increase robustness of this model, the training incorporates JPEG compression simulation, flips, and minor rotations. Performance is measured using accuracy, precision, recall, and AUC (Area Under the Curve). Table 1 shows a performance comparison of this architecture on a mixed deepfake image dataset.

Table 1. Hybrid CNN Architecture example with Frequency Pathway

Method	Accuracy	Precision	Recall	AUC
Proposed CNN Architecture	94.0%	0.98	0.96	0.99
ResNet-50 (baseline)	88.5%	0.90	0.88	0.94
XceptionNet (Thies et al.)	90.2%	0.92	0.90	0.95
Hybrid Ensemble (Patel et al.)	93.2%	0.95	0.94	0.9744

Video Detection Technique. Visual-based detection targets anomalies in video frames that are difficult for humans to spot. Early deepfakes often involved simple identity-swapping algorithms such as Face Swap, Face swap-GAN. These models replaced one person's face with another's, often emitting subtle visual errors, abnormal lighting, and mismatched skin textures.

To tackle these issues, researchers developed convolutional neural network (CNN)-based detectors. For instance, mesoscale CNNs were skilled at recognizing coarse artifacts in manipulated videos. The Extreme Inception Network, in contrast, focused on preserving channel-wise and geometric details, though it sometimes failed to catch some irregulapities.

As the generation of synthetic videos becomes more refined, so too must the methods for detection. Various studies outline an array of technical strategies that detect video deepfakes by exploiting both spatial anomalies (within a single frame) and temporal inconsistencies (across sequential frames). Figure 3 is a procedure for detecting face manipulation that combines CNN and RNN (Jbara, Hussein, 2024).

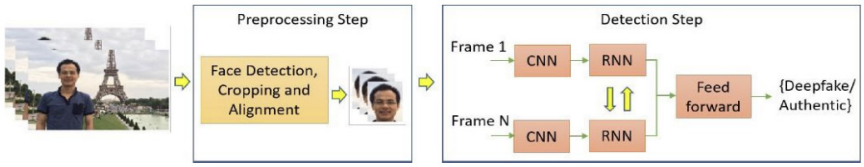

Fig. 3. Detection system that combines CNN and RNN (Nguyen et al., 2019).

Recently, transformer-based architectures have pushed detection further (Alrashoud, 2025). One of the most powerful approaches is the Spatiotemporal Dropout Transformer (STD-T). This model introduces random dropout at both the frame and patch level during training, which forces the model to generalize better and focus on deepfake-specific temporal patterns rather than overfitting on superficial cues. In addition to transformers, EfficientNet-V2 has been implemented as a lightweight and fast solution for frame-by-frame detection. When combined with temporal aggregation layers

(such as GRUs or LSTMs), this pipeline can track facial dynamics across video with lower latency, making it suitable for deployment in live settings or mobile platforms. Performance of video deepfake detection systems is rigorously evaluated using well-known datasets:

Table 2. Detection Performance on Video Deepfake Datasets

Model	Dataset	Frame Accuracy (%)	AUC	Temporal Method
Spatiotemporal Transformer	DFDC	98.0	0.97	Dropout Attention
XceptionNet + LSTM	Celeb-DF	93.5	0.94	Bidirectional LSTM
EfficientNet-V2 + GRU	DFDC	95.2	0.96	GRU

Text Detection Techniques. One of the leading detection frameworks is DetectGPT, as cited by Ranjan et al. (2023). This method evaluates the curvature of log-probabilities over the model's own output space. In simple terms, human-authored sentences tend to reside in high-entropy regions of a language model's distribution, whereas machine-generated sentences often cluster in low-curvature areas that reflect "too likely" or overly optimized phrasing. Another technique mentioned is GLTR (Giant Language Model Test Room), which uses token-level probability histograms to analyze how likely each word in a passage is, according to a baseline model. Human writing tends to have a mix of common and rare word choices, while AI-generated text often skews toward medium- to high-probability vocabulary. DeepTextMark, a third approach introduced by Vasudeva et al. (2024), functions like a digital watermark for generated text. It modifies sentence structure or introduces stylized synonym replacements that serve as invisible signatures embedded in output by compliant language models.

Table 3. Comparison of Text Deepfake Detection Models

Model	**Detection Mechanism**	**Accuracy (%)**	**Key Limitation**	**Source**
DetectGPT	Curvature in log-probability space	92.0	Model-specific sensitivity	Ranjan et al. (2023)
GLTR	Token likelihood histograms	86.3	Manual review may be needed	Jain & Aggarwal (2024)
DeepTextMark	Stylized synonym watermarking	90.5	Vulnerable to rephrasing	Vasudeva et al. (2024)

Audio and Speech Detection Techniques. State-of-the-art systems use Capsule Networks (CapsNets), particularly the ABC-CapsNet architecture. These models analyze

mel-spectrograms—a time-frequency representation of sound—to uncover fine-grained details in pitch, energy, and voiceprint (Jain & Aggarwal, 2024). CapsNets are especially useful for speech detection because they preserve spatial hierarchies and can capture subtle differences in how features like pitch or timbre evolve over time.In tests cited by Zhou et al. (2025), ABC-CapsNet achieved a 97.5% detection accuracy on the ASVspoof2019 dataset, outperforming traditional CNN and RNN-based classifiers.

In addition to speech and audio, Spotting fake music created by AI has become reliable, with some techniques reaching almost 99.8% accuracy. The trick is to find tiny hints left by AI-generated audio instead of just listening to the music. Scientists begin with a big set of real songs, like the FMA database that has thousands of tracks from many genres. They then make fake versions using special AI tools to rebuild the audio. These pairs of real and fake sounds help teach systems to spot the differences. Most detection methods use convolutional neural networks, and how they prepare the audio (such as turning it into spectrogram pictures) plays a big role in how well it works. Experts have started to look at deeper musical elements like melody, rhythm, and lyrics such as things AI still can't copy. Table 4 shows the comparison between Audio Deepfake Detection Models and Performance:

Table 4. Audio Deepfake Detection Models and Performance

Model	Technique	Dataset	Accuracy (%)	Reference
ABC-CapsNet	Mel-Spectrogram + Capsules	ASVspoof 2019	97.5	Zhou et al. (2025)
ResNet Audio-Net	Time-domain CNN	Fake-VoiceDB	93.8	Jain & Aggarwal (2024)
LSTM-Audio	Temporal Spectral Analysis	LibriFake	91.2	Vasudeva et al. (2024)

4 Results and Discussion

Detection systems across all media types have advanced considerably, often achieving over 90% accuracy on benchmark datasets. In image and video, models like XceptionNet, STD-Transformer, and wavelet-enhanced CNNs demonstrate near-human performance on datasets such as FaceForensics++ and DFDC (Chakraborty et al., 2024; Jain & Aggarwal, 2024). In speech and music, capsule networks and spectrogram-based classifiers like ABC-CapsNet and FIONA deliver strong results against synthetic audio (Zhou et al., 2025; Gao et al., 2024).

Text detection lags slightly behind due to the subtlety of language and the high quality of current generative models. Techniques like DetectGPT and GLTR perform well under controlled conditions but remain vulnerable to paraphrasing and fine-tuned outputs (Ranjan et al., 2023; Vasudeva et al., 2024).

Despite strong lab performance, real-world robustness is still a concern. Many detectors are vulnerable to:

- Post-processing obfuscation (e.g., compression or noise),

- Model evolution, where newer architectures evade older detection patterns,
- Multimodal manipulation, where deepfakes combine text, audio, and visual inputs.
- A critical gap remains in multimodal deepfake detection, where integrating signals across text, speech, and visuals in a single coherent framework remains technically and computationally demanding.

5 Conclusions

As deepfake technologies become increasingly sophisticated across text, image, video, speech, and music, the need for robust, reliable, and scalable detection systems has never been more critical. While generative AI continues to unlock creative potential in digital art, film, audio production, and storytelling, it simultaneously erodes long-standing boundaries between real and synthetic media. This duality positions detection not as an auxiliary function, but as a core requirement for preserving authenticity, authorship, and trust.

In the visual domain, detection tools have kept pace with increasingly realistic outputs generated by GANs and diffusion models. CNN-based systems like XceptionNet and frequency-aware models leveraging wavelet decomposition have achieved detection accuracies above 99% on datasets such as FaceForensics++ and Celeb-DF. For video, spatiotemporal transformer architectures provide a deeper understanding of motion continuity and frame-level coherence, offering resilience even against high-quality face reenactment and lip-syncing deepfakes.

Text detection, while inherently more abstract, has made strides through methods like DetectGPT, GLTR, and synonym-based watermarking. These systems analyze token-level probability distributions and curvature in language model output space to identify statistically "too perfect" text. However, the line between machine and human authorship in text remains blurred, making linguistic deepfakes one of the most difficult to detect reliably—especially in paraphrased or fine-tuned outputs.

In speech and audio, tools like ABC-CapsNet and mel-spectrogram classifiers have demonstrated remarkable effectiveness, with detection accuracies reaching 97.5% on ASVspoof datasets. These models focus on uncovering subtle acoustic artifacts left behind by neural vocoders, particularly in pitch, harmonic structure, and spectral transitions. Nevertheless, the rise of high-fidelity models like UC-VITS and real-time voice cloning introduces new threats that require continuous adaptation of detection pipelines.

In conclusion, detection is not merely a technical hurdle—it is a cultural safeguard. As synthetic media becomes inseparable from the creative process, the integrity of that process will depend on how effectively we can distinguish invention from deception. The art world, legal systems, and digital platforms must treat detection as a foundational component of the creative ecosystem, essential not only for security but for preserving the meaning and value of art in the AI era.

References

Alrashoud, M. (2025). Ethical boundaries of AI-generated content in digital culture. *Journal of Emerging Media and Society*, *14*(2), 133–150.

Babael, F., Amini, S., & Chen, Y. (2025). Multimodal deepfake detection in cross-media environments. In *Proceedings of the International Conference on Multimedia Security* (pp. 45–59).

Chakraborty, A., Nair, V., Singh, R., & Joshi, P. (2024). Hybrid attention GAN for consistent face swapping in artistic performances. In *Proceedings of the International Conference on Creative AI* (pp. 22–34).

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The DeepFake Detection Challenge (DFDC) dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020*. https://ieeexplore.ieee.org/document/9156360

Gao, Y., Lin, Q., & Wu, T. (2024). StyleMusic: A cross-modal generative framework for stylized AI music composition. *Journal of Neural Audio Processing*, *17*(2), 110–125.

Ignatova, D. (2025). Application of wellness program for the prevention of aggressive behavior. *Scientific Journal "Kindergarten School"*, *6*(1), 60–116.

Jain, M., & Aggarwal, R. (2024). Understanding and detecting deepfakes across visual and audio modalities. *International Journal of Artificial Intelligence Research*, *31*(1), 78–99.

Mlynar, V., Polák, M., & Kruliš, M. (2023). DeepTextMark: A deep learning-driven text watermarking approach. *Applied Sciences*, *13*(17), 9875. https://www.mdpi.com/2076-3417/13/17/9875

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., & Nguyen C. M. (2019) '*Deep Learning for Deepfakes Creation and Detection: A Survey*'. Computer Vision and Image Understanding, *223*, Article 103525. https://doi.org/10.1016/j.cviu.2022.103525.

Ranjan, S., Li, H., & Das, T. (2023). DetectGPT and the future of synthetic text forensics. In *Proceedings of the Computational Linguistics and AI Ethics Symposium* (pp. 87–101).

Sabatelli, M., Sutherland, M., Cakmakci, O., & Reinecke, K. (2023). Organic or diffused: Can we distinguish human art from AI-generated images? In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. https://dl.acm.org/doi/abs/10.1145/3576915.3623189

Solaiman, I., Clark, J., & Zhang, J. (2024). Human vs. machine: A comparative study on the detection of AI-generated content. *arXiv Preprint*, arXiv:2402.03214. https://arxiv.org/abs/2402.03214

Vasudeva, N., Kapoor, A., & Iyer, S. (2024). DeepTextMark: Watermarking-based authorship verification for AI-generated text. In *Advances in AI for Digital Humanities* (pp. 54–69). Springer.

Wolter, M., Blanke, F., Heese, R. & Garcke, J. (2022) Wavelet-packets for deepfake image analysis and detection. *Machine Learning, 111*, 4295–4327. https://doi.org/10.1007/s10994-022-06225-5

Zhou, L., He, K., & Yuan, M. (2025). UC-VITS: Cross-lingual voice cloning and emotional speech synthesis via variational inference. *Transactions on Speech Technologies*, *29*(1), 15–37.

Received: April 15, 2025
Reviewed: May 15, 2025
Finally Accepted: June 01, 2025