RCMOS – Multidisciplinary Scientific Journal The Knowledge.
ISSN: 2675-9128. São Paulo-SP.
Year II, v.2 2022. | Submission: July 12, 2022 | Accepted: July 14, 2022 | Publication: July 16, 2022

Deep Learning Models for Deepfake Detection: A Comparative Analysis between CNN, GAN, and Transformer Architectures

Deep Learning Models for Deepfake Detection: A Comparative Analysis among CNN, GAN, and Transformer Architectures

Matheus de Oliveira Pereira Paula holds a Bachelor's degree in Information Systems from the Federal Institute of Education, Science and Technology Fluminense and an MSc in Data Science and Artificial Intelligence from the Université Côte d'Azur.

SUMMARY

The rapid evolution of **synthetic media manipulation** technologies , known as *deepfakes,* represents a growing threat to information reliability and digital security.
Such falsified videos, created primarily by Generative Adversarial Networks (GANs), make the distinction between real and forged content increasingly complex, requiring sophisticated countermeasures based on **Deep Learning.** This article proposes a rigorous comparative analysis of three fundamental architectures in the field of Computer Vision for deepfake detection : Convolutional Neural Networks **(CNNs),** Generative Adversarial Networks (in their role as detectors or in hybrid models that exploit their signatures), and **Transformers.**
(particularly Vision Transformers - ViTs). The evaluation focuses on critical metrics for application in real-world scenarios, including classification **accuracy , processing time** (inference latency), and the fundamental ability to **generalize** to different forgery techniques and unseen datasets *(cross-dataset evaluation).* The results of the literature review and theoretical analysis indicate that, while CNNs (such as XceptionNet) remain relevant due to their efficiency and ability to capture local artifacts, Transformer-based architectures demonstrate a superior ability to model global dependencies and, consequently, exhibit better generalization against constantly evolving *deepfake* methodologies .

Keywords: Deepfake; Deep Learning; CNN; GAN; Transformer; Generalization; Digital Forensics.

ABSTRACT

The rapid advancement of **synthetic media manipulation** technologies, commonly known as *deepfakes,* poses an increasing threat to information trustworthiness and digital security. These forged videos, primarily created by Generative Adversarial Networks (GANs), make the

distinction between real and fake content increasingly difficult, necessitating sophisticated **Deep Learning-based** countermeasures. This paper presents a rigorous comparative analysis of three fundamental architectures in Computer Vision for *deepfake* detection: Convolutional Neural Networks **(CNNs),** Generative Adversarial Networks (in their role as detectors or in hybrid models that exploit their signatures), and **Transformers** (particularly Vision Transformers - ViTs). The evaluation focuses on critical metrics for real-world application scenarios, including classification **accuracy, processing time** (inference latency), and the essential **generalization** capability to unseen forgery techniques and datasets *(cross-dataset evaluation).* The results from the bibliographic and theoretical analysis indicate that while CNNs (such as XceptionNet) maintain relevance due to their efficiency and ability to capture local artifacts, Transformer-based architectures demonstrate a superior capability to model global dependencies and, consequently, exhibit better generalization against the constantly evolving *deepfake* methodologies.

Keywords: Deepfake; Deep Learning; CNN; GAN; Transformer; Generalization; Digital Forensics.

1. INTRODUCTION AND DELIMITATION OF THE DEEPFAKE PROBLEM

The proliferation of synthetic media, notably *deepfakes,* has emerged as one of the most pressing challenges at the intersection of computer science and information security, driven by the development of **Deep Learning models.** The ability to convincingly manipulate videos and audios, replacing faces or altering expressions with frightening realism, has transcended the realm of academic research and infiltrated the social, political, and economic landscape, being responsible for disinformation crises and reputational damage.
The root of the problem lies in the very nature of the deepfake creation process , which uses neural network algorithms, such as Generative Adversarial Networks (GANs) and, more recently, *Diffusion models,* to generate visual artifacts that deceive the human eye and, crucially, first-generation detection systems.

The complexity in detection lies in the fact that *deepfake* generation systems are in a constant "cat and mouse game" with detectors, in a cycle known as **Adversarial Warfare.** As detection models become more efficient at identifying specific artifacts (such as inconsistencies in blinking, or in lighting and skin texture), generative models rapidly evolve to eliminate these digital *fingerprints .* This dynamic imposes the need for an in-depth comparative study of detection architectures, not only in terms of performance on closed datasets, but, primarily, in their **robustness and generalization** to unknown manipulation techniques. *Deepfake* detection therefore requires a digital forensic focus that transcends simple binary classification and explores the high-level representations of neural networks.

This work aims to fill a critical gap in the literature by systematizing the comparison of the three most prominent classes of Deep Learning models in this domain: **Neural Networks**

Convolutional Networks (CNNs), which have long been the cornerstone of Computer Vision; **Generative Adversarial Networks (GANs),** in their role as base models for detecting *forgery signatures;* and the emerging **Transformers,** which have redefined the state of the art in modeling sequences and long dependencies in visual data. The choice of these architectures is not arbitrary, but rather representative of the main evolutionary approaches in the field: CNNs focus on **locality and translational invariance,** GANs exploit the **adversarial nature of the problem,** and Transformers prioritize the **modeling of global dependencies** through the attention mechanism.

The proposed analysis will not be limited to the simplistic metric of accuracy, but will include **computational efficiency,** measured by inference processing time, which is vital for real-time applications and resource-constrained environments, and the most decisive factor for practical applicability: **generalization capability.** A detection model is considered superior if it can maintain its high performance when tested on a dataset generated by a *deepfake* technique different from the one used for training. This is the true measure of its robustness against the continuous evolution of forgery techniques.

Thus, this article is structured to first present the theoretical foundation of each architecture in the context of *deepfake detection,* and then to critically and comparatively analyze their operational advantages and disadvantages in terms of accuracy, latency, and generalization. Finally, conclusions and future research directions will be outlined, with a strong emphasis on the need for hybrid architectures that combine the best of local and global feature extraction to address the challenge of digital media authenticity in the 21st century.

2. Convolutional Architectures (CNN) in Artifact Detection PLACES

Convolutional Neural Networks **(CNNs),** long the gold standard in Computer Vision, were the pioneering and most widely used architectures in *deepfake detection.* Their effectiveness lies intrinsically in the nature of their fundamental building blocks: the layers.

Convolutional techniques are designed to hierarchically extract local and invariant features with small translations in pixel space. For manipulation detection, this capability is exploited to identify **inconsistency artifacts** introduced in the deepfake generation process , often visible as high-frequency noise, discontinuities in skin texture, or subtle failures in preserving the spatial and temporal coherence of the video, such as blinking anomalies or lip closure during speech.

Models like **XceptionNet** and variations of **EfficientNet** have stood out in *deepfake* detection competitions for their ability to capture these low-level *fingerprints .* The *design*
Xception, for example, which uses depthwise *separable convolutions,* allows the model to learn channel mappings and correlations.

Spatials more efficiently and thoroughly, being remarkably effective in isolating residual signatures left by specific interpolation and facial *warping* algorithms. In terms of **accuracy on** viewed datasets (such as FaceForensics++ or CelebDF), CNNs consistently achieve robust performance, often exceeding 90%, which validates their suitability in detecting artifacts that are well represented in the training data.

In terms of **processing time** (inference latency), CNNs, especially variants optimized for *edge* or low-latency devices (such as MobileNetV2 or EfficientNet B0), offer a significant advantage. The nature of convolutional operations, which are highly parallelizable on graphics processing units (GPUs), and the hierarchical structure that progressively reduces spatial dimensionality, make them suitable for **real-time** detection in video *streams* . However, it is crucial to note that models with greater depth and a larger number of parameters (such as Xception or EfficientNet B7, which often achieve higher accuracy) demand substantial computational resources, which can be a bottleneck in restricted *deployment* environments , requiring compression or *quantization strategies.*

The major drawback of CNNs lies in their *cross-dataset generalization* **capability .** Because they are inherently designed to focus on local features and pixel neighborhoods, they tend to over-specialize in the **specific artifacts** of the *deepfake* method used to generate the training dataset. When confronted with faked videos created by a different technique (e.g., trained on FaceSwap and tested on StyleGAN2), the performance drop is often pronounced. This limitation is a direct reflection of their **inductive locality bias ,** which hinders the modeling of long-range inconsistencies or incoherencies that manifest across the entire face or in longer time sequences of the video.

Thus, the role of CNNs is evolving from monolithic detectors to **fundamental feature extractors** in hybrid models. Although their high accuracy in *intra-*

While *datasets* and their *runtime* efficiency are undeniable, the battle against generalization forces the research community to seek architectures that can capture global and temporal coherence, which traditional CNNs, on their own, struggle to model effectively. They established the starting point, but the increasing complexity of *deepfakes* demands tools that look beyond patch artifacts .

3. The Generative Adversarial Networks (GAN) Approach in Detection

Generative Adversarial Networks **(GANs),** while being the primary technology behind the creation of *deepfakes* (acting as the *Generator component),* also play a dual role in detection, either through their direct use as specialized discriminators or in inspiring methodologies that seek to capture the **unique signatures** that the process of

Adversarial training leaves synthetic media. The GAN concept, introduced by Goodfellow et al. (2014), is fundamentally a zero-sum game between a generator and a discriminator, and it is the discriminator's ability to learn to distinguish real samples from fake ones that is adapted for digital forensics.

The application of GANs in detection is often indirect, focusing on identifying the **"manufacturing trail"** or **"fingerprint"** of the generator that produced the *deepfake.* Such trails manifest as systematic and non-uniform artifacts in the frequency domain or in noise patterns that are learned and consistently replicated by the Generator. Detection models based on **frequency domain analysis** or **Noise Analysis Neural Networks**

(Noise Analysis Networks) utilize the adversarial principle by training a detector to be highly sensitive to these recurring noises. For example, exploratory techniques use the Discrete Cosine Transform (DCT) to analyze the frequency distribution, where *deepfakes* generated by GANs are expected to exhibit distinct statistical deviations compared to authentic videos.

In terms of **accuracy,** models that exploit the GAN signature achieve exceptional results when the tested *deepfake* was generated by a known algorithm and the detector was specifically trained for that signature. Accuracy can be extremely high (above 95%) for detecting *deepfakes* generated by architectures such as StyleGAN, PGAN, or CycleGAN, proving that the adversarial generation process does indeed leave a detectable fingerprint.

However, this specialization is the greatest source of weakness: the GAN signature detector is inherently **vulnerable to changes in the generating algorithm** or to post-processing techniques, such as compression, which can corrupt the noise pattern.

The computational efficiency and **generalizability** of these approaches are inversely proportional. GAN-based detection, often requiring the training of a specific adversarial model or the analysis of transformed domains (such as frequency), can be **computationally expensive and slow** in inference, especially if the analysis has to be done frame-by-frame and in real time. Furthermore, **generalization is severely limited;** a detector optimized for the fingerprint of a StyleGAN3 will fail when attempting to identify a *deepfake* created by a *Diffusion* model or a more rudimentary *FaceSwap* technique . GAN artifact detection is, in essence, a fight against the creator of the *deepfake,* not against the manipulation itself.

In short, the contribution of GANs to *deepfake* detection is more conceptual and methodological than architectural in the traditional sense. They highlight that digital forensics should focus on **flaws in the synthesis process** rather than just semantic characteristics. However, the extreme specialization and low generalization inherent in this approach relegate it to a complementary role or a reference technique, making it less suitable as a universal solution for the dynamic and diverse landscape of contemporary *deepfakes ,* which are constantly evolving in their generation techniques.

4. The Role of Transformers and the Attention Mechanism

Transformers , originally proposed for Natural Language Processing (NLP), revolutionized Computer Vision with the introduction of **Vision Transformers (ViTs),** and quickly established themselves as a cutting-edge architecture in *deepfake detection,* challenging the dominance of CNNs. The fundamental strength of Transformers lies in their **self-attention** mechanism , which allows them to model **long-range and global dependencies.**
between different parts of an image or video, something that CNNs based on local convolutional windows struggle to do efficiently.

In deepfake detection , this global modeling capability is crucial. Instead of just focusing on texture artifacts in a pixel *patch ,* Transformer can analyze the **coherence of the entire scene,** the **spatial relationships** between the face and the environment, or **lighting inconsistencies .** that propagate throughout the image. Self-attention allows the model to weigh the importance of each "token" (pieces of the input image) in relation to all others, learning to identify anomalies that are not obvious locally, but manifest as a break in the overall visual logic, such as the misallocation of shadows or the difference in resolution between the falsified region and the rest of the frame.

In terms of **generalization** *(cross-dataset generalization),* Transformers (or hybrid architectures like Swin Transformer and GenConViT) demonstrate **superior performance** compared to traditional CNNs. Recent literature suggests that, instead of learning to detect compression artifacts or generator-specific *fingerprints* (as CNNs tend to do), ViTs learn **more semantic and robust representations** of "reality" or "authenticity." This more abstract representation, less sensitive to noise variations or post-processing, gives Transformers greater resilience when exposed to unseen spoofing methods, making them the preferred architecture in the search for a universal detector.

However, the primary disadvantage of Transformers lies in **processing time** and **computational requirements.** The calculation of the self-attention mechanism, which scales quadratically with the number of input *tokens ,* is substantially more costly than convolutional operations. This results in a **higher inference latency** for ViTs compared to optimized CNNs. Furthermore, Transformers require a significantly larger amount of training data to achieve peak performance and mitigate the risk of *overfitting,* since they have **weak inductive bias** (less prior knowledge about the spatial structure of the images) compared to the locality bias of CNNs.

The current trend, therefore, is the adoption of **hybrid models** that seek to mitigate the shortcomings of Transformers. Architectures such as the Convolutional Vision Transformer (ConViT) and the integration of attention modules into CNN *backbones* (as in ResNet models with attention blocks) seek to combine **efficiency and the ability to extract local features.**

CNNs with the **long-range modeling and generalization power of Transformers.** This architectural fusion represents the state of the art in deepfake detection , indicating that the future of digital forensics lies in the synergy of these deep learning paradigms.

5. COMPARATIVE ANALYSIS OF ACCURACY AND GENERALIZATION

Accuracy metrics alone are notoriously insufficient to assess the effectiveness of a *deepfake* detector in the real world, serving only as a starting point for comparative analysis. In *intra-dataset scenarios,* where the model is trained and tested on data from the same forgery distribution, the most advanced **CNN** architectures, such as **XceptionNet** or **EfficientNet B7,** frequently achieve the highest values, reaching over 95% in *benchmarks* like FaceForensics++. This is due to their excellence in detecting artifacts.

Specific, high-frequency fingerprints are characteristic of trained forgery. Specialization in extracting low-order *fingerprints* is key to the local success of these networks.

However, the true test for any *deepfake* detection model is its cross-dataset generalization **capability ,** that is, the model's performance when exposed to a new forgery technique or an unseen dataset (e.g., trained on FaceForensics++ and tested on CelebDF-V2). In this critical scenario, **Transformer-** based architectures and their hybrids demonstrate a clear and consistent advantage over CNNs. Comparative studies show that while the accuracy of CNNs can drop drastically (e.g., from 95% to 60%), Transformers (such as **ViT** or **DeiT models),** although they may have slightly lower accuracy in the *intra-dataset* scenario (88% to 92%), exhibit a **much less pronounced performance drop** in the *cross-dataset* scenario.

(remaining around 75% to 85%).

The superiority in generalization of Transformer is attributed to its attention mechanism, which allows it to capture **global and semantic coherence anomalies** that are less dependent on the low-level artifacts of the generation algorithm. Instead of learning codec *or* interpolation noise, Transformer is more likely to learn consistency breaks in facial geometry or light physics throughout the video. This ability to model **high-level inconsistencies** gives the model a more **robust** representation and makes it less susceptible to variation from *forgery* techniques (such as *face swapping* versus *face reenactment).* Transformer's accuracy is therefore more "honest" and less inflated by specialization in a single artifact type.

Models that utilize **GAN** philosophy for detection, focusing on the generator's signature, exhibit the worst generalization performance. Because they are highly tuned to identify the noise or statistical *fingerprint* of a specific generator (e.g., StyleGAN), they **completely collapse** when confronted with a different generation technique. Their high accuracy in very specific scenarios does not translate into a useful defense system for the real-world environment, where new *deepfake* techniques are continuously being launched. Detection based on

GAN signature analysis is, at best, a diagnostic tool for specific classes of *deepfake,* and not a universal forensic solution.

In summary, the analysis shows a clear *trade-off :* **CNNs** are **quickly accurate** in observed scenarios, but **lack generalization; GAN** approaches are **extremely specialized.** and **not generalizable;** and **Transformers** offer consistently robust accuracy **and superior generalization,** making them the most promising architecture for developing sustainable and resilient *deepfake* detection systems that are resilient to evolving synthesis techniques.

6. COMPARISON OF PROCESSING TIME AND EFFICIENCY COMPUTATIONAL

The **processing time** for inference and the **computational efficiency** of Deep Learning models are crucial factors that define the viability of a *deepfake* detector.
In practical applications, especially those requiring **real-time detection,** such as social media platforms or live identity verification systems, inference latency is typically measured in frames per second (FPS) that the model can process. This metric reveals a notable *trade-off* between accuracy/generalization and efficiency.

CNN architectures set the standard in terms of computational efficiency. Optimized models, such as the lighter versions of **EfficientNet (B0 to B4)** or variants of **MobileNet,** can achieve processing rates of tens to hundreds of frames per second on consumer hardware (GPUs) or even CPUs, using *pruning* and *quantization techniques.* The local nature of convolutions and the efficient use of parameters allow CNNs to extract critical features with a smaller number of floating-point operations (FLOPs) compared to attention-based models. This **high frame rate** is why CNNs remain the primary choice for large-scale pre-screening or rapid screening of *deepfakes ,* even with their disadvantage in generalization.

In stark contrast, **Transformer** architectures , due to the inherent complexity of the **self-attention mechanism,** are significantly **slower and more resource-intensive.** The attention matrix calculation, which involves multiplying matrices that scale quadratically with the number of *tokens* (or image *patches),* imposes a heavy computational load. Even optimized **Vision Transformers (ViTs)** and hybrid models (such as the Swin Transformer, which adopts shifted attention windows to reduce quadratic to linear complexity relative to the number of pixels in a window) still exhibit **higher inference latency** and require **more GPU memory** than comparable CNNs in terms of accuracy. The processing time of Transformers is a significant obstacle to real-time detection and *deployment* on resource-constrained devices.

Regarding **GAN-** based approaches (focus on the generator signature), efficiency varies widely, but they often impose a high computational load. If detection involves the analysis of transformed domains (such as DCT) or the execution of a complete adversarial process, the processing time may be impractical for continuous *streaming .* Such methods are generally relegated to ***offline*** **forensic analyses ,** where time is not a critical constraint, but the depth of analysis is paramount. They are not viable candidates for a high-speed detection system.

The cost-benefit *trade-off ,* therefore, leans in different directions depending on the application. If the priority is **speed and low cost** in high volume (e.g., a social media *feed),* **CNNs** are the most pragmatic choice, accepting the risk of lower generalization. If the priority is **maximum robustness and generalization** to protect high-value assets, the **Transformer,** despite its slowness and high cost, is the safest option. Current research focuses on **knowledge distillation** and **optimization of hybrid models** to transfer the generalization capabilities of large Transformers to more efficient CNNs, seeking the ideal balance between accuracy, robustness, and processing speed.

7. Hybrid Models and Future Direction of Research

A comparative analysis of CNN, GAN, and Transformer architectures reveals that no single approach is the definitive solution to the constantly evolving problem of *deepfake* detection . CNNs are fast and detect local artifacts, but do not generalize well. Transformers generalize superiorly by modeling global coherence, but are computationally expensive. GAN signature-based techniques are very specific and not generalizable. This scenario drives the research field towards the development of **Hybrid Models,** which seek to leverage the strengths of each paradigm while minimizing their weaknesses.

The main line of hybrid research focuses on the combination of **CNNs and Transformers.** Models such as **GenConViT** or variations of the **ConvNeXt-Swin Transformer** utilize a CNN architecture (such as EfficientNet or ResNet) as a *backbone* or low-level feature extractor, responsible for the efficient analysis of local *patches* and the initial capture of high-frequency artifacts. The extracted features are then passed to a **Transformer** component. (like a Swim Transformer or a Vision Transformer) that applies the self-awareness mechanism. This division of tasks allows the hybrid model to benefit from the **efficiency of convolution in extracting *local features*** and the **generalization power of the Transformer in modeling global and temporal dependencies** (across video frames).

The future of *deepfake* detection also lies in the integration of **multimodal and multispectral** information . Video *deepfakes* can be detected more robustly if the model analyzes not only the visual content (pixels) but also the **audio** (inconsistencies in synchronization).

Lip alteration or synthetic voice artifacts) and **physiological signals (such as** heart rate measurement via *Remote Photoplethysmography* - rPPG, which is a notoriously difficult signal to fake). Hybrid multimodal models that combine CNNs for extracting spatial visual *features ,* LSTMs or *Transformers* for temporal modeling of rPPG, and *Attention Mechanisms* for audio-visual synchronization promise significantly greater robustness against complex manipulations.

Another key direction is focusing on **robustness against post-processing** and **adversarial attacks.** In practice, *deepfakes* are often compressed by video *codecs* (such as H.264 or VP9) and distributed in noisy environments (social networks), which degrades the subtle artifacts that CNNs look for. Future research should focus on training models (primarily Transformers, which are already more robust) with **synthetically degraded data .**
(with compression, noise and blurring) or using **adversarial training** techniques to ensure that the detector does not specialize in artifacts that are easily destroyed in distribution.
The goal is to develop a detector that learns the **invariants of authenticity,** not the weaknesses of a specific *forgery* method or *codec .*

In conclusion, the research goes beyond simple architectural comparison, embracing a philosophy of **synergy and functional specialization.** The ideal *deepfake* detection model will likely be a **hybrid CNN-Transformer-Multimodal** architecture that utilizes the local efficiency of convolution, the generalization power of the attention mechanism, and the robustness of merging multiple forensic signals, ensuring that detection capabilities can keep pace with the accelerated rate of innovation in synthetic media generation.

8. CONCLUSION AND FUTURE IMPLICATIONS

This detailed comparative study of **Deep Learning** architectures for *deepfake* detection – **CNNs, GANs, and Transformers** – confirmed that the choice of the ideal model is intrinsically linked to the ***deployment* scenario** and the **requirements for robustness and generalization.**
The analysis demonstrated that **Convolutional Neural Networks** (CNNs) are undeniably efficient in terms of **processing time** and achieve high **accuracy** in *intra-dataset scenarios,* where manipulation artifacts are known. This performance makes them the preferred solution for video screening in high-throughput, resource-constrained environments. However, their locality bias and tendency to *overfit* to training set-specific artifacts condemn them to **poor generalization** when confronted with unseen falsification methods or post-processing typical of the real world, limiting their use as a last resort.

The **GAN** signature-based detection approach proved to be the weakest and most specialized. While it can identify the *fingerprints* of a specific generator with very high accuracy, its **generalization is virtually** nil outside that restricted domain. This methodology serves more as a **forensic attribution** tool (identifying the software or algorithm used).

(in counterfeiting) rather than as a robust and universal defense system against the constantly evolving threat. Excessive reliance on manufacturing noise makes them easily susceptible to adversarial cleanup attacks or simple variations in the generation pipeline.

The study unequivocally highlighted **Transformer architectures** as offering the **best generalization capabilities.** By employing the self-attention mechanism to model global and long-range dependencies, Transformers are able to learn high-level **authenticity invariants** and semantic inconsistencies, rather than low-order pixel artifacts.
This capability gives them **significantly greater robustness** against new *deepfake* techniques and variations in data distribution. However, the computational cost and **low inference rate** associated with the quadratic complexity of attention are the main impediment to their large-scale adoption in real time.

The main future implication of this work is the validation of the **CNN-Transformer hybrid architecture** as the most promising approach. Research should focus on designing models that utilize CNN for the **efficient extraction of local *features*** (where it excels) and Transformer for the **integration and validation of the global and temporal coherence of these** *features.*
(where it is unsurpassed in generalization). This architectural synergy promises an ideal balance between **cutting-edge accuracy, robustness against generalization, and *deployment feasibility.*** Furthermore, the incorporation of **multimodal information,** such as physiological signal analysis (rPPG) and audio-visual coherence, should be the next frontier to raise the difficulty for deepfake creators .

Ultimately, the war against synthetic media is a continuous cycle of adversarial innovation. The research community must move away from the pursuit of a detector that achieves 100% accuracy on a static *benchmark* and focus on building **resilient and adaptive models.** The *deepfake* detector of the future will not be an unbreakable black box, but rather a system that **learns and continuously updates itself** with new degradation and generation techniques, prioritizing generalization over over specialization. The transition from CNNs to Transformers, and now to hybrid models, is not just an architectural evolution, but a paradigm shift in the **philosophy of digital forensics:** from searching for **implementation flaws .**

searching for **breaks in the logic of reality**

REFERENCES

Books and Articles.

1. GOODFELLOW, Ian et al. Generative Adversarial Networks. In: **Advances in Neural Information Processing Systems** (NIPS), 2014.
2. HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long Short-Term Memory. **Neural Computing,** vol. 9, no. 8, p. 1735–1780, 1997.

3. KINGMA, Diederik P.; BA, Jimmy. Adam: A Method for Stochastic Optimization. In: **International Conference on Learning Representations** (ICLR), 2015.
4. LECUN, Yann et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE,** vol. 86, no. 11, p. 2278–2324, 1998.
5. ROSSLER, Andreas et al. Face Forensics++: Learning to Detect Manipulated Facial Images. In: **International Conference on Computer Vision** (ICCV), 2019.
6. VASWANI, Ashish et al. Attention Is All Needed. In: **Advances in Neural Information Processing Systems** (NIPS), 2017.
7. XU, Haonan et al. Positional Encoding for Deepfake Detection. In: **IEEE International Conference on Image Processing** (ICIP), 2021.
8. ZHOU, Peng et al. Two-Stream Neural Networks for Tampered Face Detection. In: **IEEE Conference on Computer Vision and Pattern Recognition Workshops** (CVPRW), 2017.
9. AFCHAR, Dariush et al. Mesop Net: A Compact Deepfake Detection Network. In: **IEEE International Conference on Image Processing** (ICIP), 2020.
10. COZZOLINO, Davide; VERDOLIVA, Luisa. **Forensic analysis of Neural Networks for generative model attribution.** International Workshop on Digital Watermarking (IWDW), 2018.