cryptography

MDPI

Review

From Black Boxes to Glass Boxes: Explainable AI for Trustworthy Deepfake Forensics

Hanwei Qian [1], Lingling Xia [1,*], Ruihao Ge [1], Yiming Fan [2], Qun Wang [1] and Zhengjun Jing [3]

[1] Department of Computer Information and Cybersecurity, Jiangsu Police Institute, Nanjing 210031, China
[2] College of Information Science and Technology & Artificial Intelligence, Nanjing Forestry University, Nanjing 210037, China
[3] School of Computer Engineering, Jiangsu University of Technology, Changzhou 213001, China
* Correspondence: xialingling@jspi.cn

Abstract

As deepfake technology matures, its risks in spreading false information and threatening personal and societal security are escalating. Despite significant accuracy improvements in existing detection models, their inherent opacity limits their practical application in high-risk areas such as forensic investigations and news verification. To address this gap in trust, explainability has become a key research focus. This paper provides a systematic review of explainable deepfake detection methods, categorizing them into three main approaches: forensic analysis, which identifies physical or algorithmic manipulation traces; model-centric methods, which enhance transparency through post hoc explanations or predesigned processes; and multimodal and natural language explanations, which translate results into human-understandable reports. The paper also examines evaluation frameworks, datasets, and current challenges, underscoring the necessity for trustworthy, reliable, and interpretable detection technologies in combating digital misinformation.

Keywords: Deepfake detection; explainable AI; forensic analysis; model interpretability; multimodal explanation

Academic Editor: Josef Pieprzyk

Received: 20 August 2025
Revised: 4 September 2025
Accepted: 15 September 2025
Published: 26 September 2025

Citation: Qian, H.; Xia, L.; Ge, R.; Fan, Y.; Wang, Q.; Jing, Z. From Black Boxes to Glass Boxes: Explainable AI for Trustworthy Deepfake Forensics. *Cryptography* **2025**, *9*, 61. https://doi.org/10.3390/cryptography9040061

1. Introduction

The rapid proliferation of Deepfake technology is fundamentally blurring the lines between reality and fiction. Driven by sophisticated generative models such as Generative Adversarial Networks (GANs), these techniques can create hyper-realistic face swaps, manipulate facial expressions, and synthesize entire audio-visual streams with a fidelity that can deceive the human senses. This technological diffusion presents a range of severe societal challenges, including the propagation of disinformation and defamation of individuals, as well as the erosion of public trust and threats to national security. In response, the research community has developed a vast array of detection models, achieving remarkable success in classification accuracy [1]. However, the vast majority of these high-performance detectors operate as black boxes, capable of rendering a real or fake verdict but unable to articulate the rationale behind their decisions.

This opacity constitutes a critical barrier to the adoption of Deepfake detection technologies in real-world applications. In high-stakes domains such as forensic investigation, journalistic verification, and content moderation, a model that cannot explain its reasoning is of limited practical value. A court of law, for instance, requires a verifiable evidence chain that explicitly identifies manipulated regions, not merely a probabilistic score. Consequently, a significant trust gap exists between the high accuracy of current detectors

and the level of credibility required for their operational deployment. Bridging this gap is the central challenge of explainability. Within the context of Deepfake detection, explainability transcends a simple classification label; It is the comprehensive ability to provide the reasoning for a decision, localize the evidence artifacts, and even shed light on the manipulation methodology [2]. An explainable model would not only classify a video as a forgery, but would also function as a digital forensic expert, highlighting the boundaries of a face swap, pointing out inconsistencies between facial features and head movements, or clarifying that its conclusion is based on anomalies in sensor pattern noise.

This paper provides a comprehensive and systematic survey of explainability methods in deepfake detection, offering researchers a panoramic overview of the field's technological trajectory. Existing approaches can be broadly grouped into several paradigms. One line of research focuses on methods based on forensic analysis, which provide direct evidence by identifying physical or algorithmic artifacts left during media generation and manipulation. Another direction emphasizes model-centric approaches, which enhance transparency either through post hoc interpretation of black-box models or by designing architectures with inherent interpretability. More recently, advances have emerged in multimodal and natural language explanations, where Large Multimodal Models (LMMs) are leveraged to generate rich, human-readable reports that articulate the reasoning behind detection outcomes [3]. In addition, this survey systematically reviews the specialized metrics and datasets developed to evaluate the quality of these explanations. By clarifying the underlying principles, advantages, and limitations of these methods, this paper aims to chart the evolutionary trajectory of explainability in Deepfake detection and to highlight promising avenues for building the next generation of trustworthy and reliable detection systems.

2. Backgroud

2.1. Overview of Deepfake Detection

The rapid proliferation of generative models has led to an unprecedented surge in fake content, posing profound challenges to the integrity of digital media and the social structures that rely on it. Early research in this domain primarily focused on a technical arms race aimed at maximizing the accuracy of binary detectors. However, the field has reached a critical juncture: Deepfakes are no longer merely a classification problem but a tool for deception, disinformation, and malicious influence, demanding a paradigm shift from detection to understanding and explanation.

Many high-performance deep learning models remain black boxes. For example, prominent Deepfake detection architectures include XceptionNet [4], which has achieved strong performance on benchmarks like FaceForensics++ [2], as well as various methods that analyze frequency-domain artifacts. Despite their high accuracy on benchmark datasets, these models often function as black boxes due to their complex, high-parameter nature. This inherent opacity limits their reliability and trustworthiness in high-stakes applications such as forensic analysis. Although effective in benchmark datasets, their opacity is a major limitation in high-stakes scenarios. Black-box models provide a probability score, such as suggesting that a video is highly likely to be fake, without revealing the reasoning behind the decision. This lack of transparency undermines trust in key applications, including legal proceedings, journalism, and national security, where a simple label is insufficient for evidence-based decision-making [5].

Opaque models also often suffer from poor generalization. Many detectors achieve high accuracy by learning artifacts specific to particular generative models or datasets, but their performance can collapse when confronted with novel manipulations. Explainable AI provides a diagnostic lens that reveals whether a model relies on fundamental generalizable manipulation traces, such as blending boundaries or frequency inconsistencies, or overfits

to superficial, generator-specific cues [6]. Understanding these mechanisms is crucial for building robust and widely applicable detection systems.

2.2. Interpreting Explainability

In the context of explainable AI (XAI), interpretability and explainability represent distinct yet complementary concepts relevant to Deepfake forensics.

Interpretability refers to the extent to which a model's internal mechanisms and decision-making processes are transparent and intelligible without auxiliary tools. An interpretable model allows direct inspection of its logic. For instance, a decision tree is inherently interpretable, as each decision can be traced through explicit logical splits [7]. In Deepfake detection, prototype-based networks exemplify this principle: the model compares new inputs against learned prototype examples, providing insights through its structure rather than requiring post hoc explanation.

Explainability denotes the capability of a system to produce human-understandable justifications for specific outputs, often applied to complex, otherwise opaque models. Post hoc techniques generate explanations such as visual heatmaps highlighting regions critical to a decision, or textual summaries describing detected manipulation artifacts. Even when the underlying model is too complex to be fully interpretable, these methods allow its outputs to be meaningfully analyzed.

This distinction establishes two primary directions in the research of fake detection. One path develops inherently transparent architectures emphasizing interpretability, while the other applies post hoc techniques to high-performance black-box models to enhance explainability. Both approaches aim to move beyond binary detection and provide the contextual information necessary for informed assessment.

2.3. Motivation for Explainable Detection

High-stakes applications increase the need for explainable forensic systems. Deepfakes pose a cognitive threat by undermining trust in digital media, potentially creating a scenario in which citizens can no longer rely on visual or auditory information. Reliable detection systems serve as the primary defense against this threat, but their effectiveness depends on explainability.

In legal and forensic contexts, digital evidence must be verifiable and defensible. Expert testimony cannot rely solely on AI predictions; it requires the presentation of underlying evidence. Explainable systems provide this capability by highlighting manipulated pixels or identifying inconsistencies in sensor noise patterns, thereby transforming AI from a black box into a tool that strengthens human analytical expertise.

In journalism and fact-checking, detecting falsified content is insufficient without understanding the nature of the manipulation. Explainable models can classify manipulations, such as face swaps or lip-sync alterations, and highlight the relevant artifacts, offering actionable intelligence that supports accurate reporting and public media literacy [8].

For content moderation, platforms face unprecedented volumes of user-generated media. Automation is necessary, but human oversight remains critical. Explainable detection guides moderators to specific frames or regions deemed suspicious, improving review efficiency and decision accuracy.

Across all domains, explainability underpins accountability. It enables auditing, challenges algorithmic conclusions, and integrates AI outputs into structured human decision processes. Without it, even highly accurate detectors remain isolated tools; with it, they become trusted partners in preserving digital integrity.

3. Foundational and Forensics-Based Explainability Methods

This chapter examines detection methods grounded in the intrinsic physical properties or generative process artifacts of digital media. These approaches provide explanations by identifying forensic traces resulting from manipulation, with interpretability arising directly from the data itself rather than from the internal decision-making mechanisms of complex detection models.

3.1. Sensor Noise Pattern Analysis

One of the earliest directions in Deepfake detection focused on the physical source of digital imagery—the camera sensor. Methods based on sensor pattern noise analysis, particularly Photo Response Non-Uniformity (PRNU), form a fundamental component of digital image forensics. The underlying principle is that every digital camera sensor possesses unique microscopic manufacturing imperfections, which produce a stable and distinctive noise pattern across images. This PRNU pattern is inherent to the sensor and remains consistent for authentic images. In contrast, manipulated content often exhibits local inconsistencies or a complete absence of this pattern in altered regions, thereby providing direct, physically grounded evidence of tampering.

3.1.1. Principle of PRNU

The seminal work by Lukas et al. [9] systematically introduced the use of sensor pattern noise, particularly PRNU, as a unique identifier for the source camera. PRNU arises from minute manufacturing variations between sensor pixels, which cause slight differences in their sensitivity to incident light. This results in a stable, multiplicative noise pattern. By extracting and averaging the noise residuals from multiple images captured by the same camera, a reference PRNU pattern can be estimated. In forensic analysis, this reference serves as a camera fingerprint. By computing the correlation between the noise residual of a query image and the reference pattern, it is possible to determine whether the image originated from the same camera. The logic is intuitive, with regions lacking the expected PRNU fingerprint consistent with the rest of the image likely to have been altered or sourced from a different device.

3.1.2. Advancements in PRNU Techniques

Marra et al. [10] addressed a more challenging forensic setting involving large collections of images with unknown origins. They proposed a blind clustering technique that relies solely on the correlation between PRNU patterns, without prior knowledge of camera sources. Their multi-stage optimization strategy combines consensus clustering with a maximum-likelihood-based merging step, improving both clustering robustness and PRNU estimation accuracy. This enables investigators to identify common sources across extensive image datasets.

Saito et al. [11] focused on the statistical reliability of PRNU-based source attribution, introducing a theoretical framework for estimating the False Acceptance Rate (FAR). They recommended the use of Peak-to-Correlation Energy (PCE) over normalized correlation, as PCE offers a more stable decision threshold that is less sensitive to variations in fingerprint strength or structured noise such as linear patterns. This work enhanced the statistical rigor of PRNU-based analysis.

Anti-forensic research has also emerged as a countermeasure to PRNU-based detection. The DIPPAS method proposed by Picetti et al. [12] employs a Deep Image Prior (DIP) framework to attenuate PRNU traces in an image. A convolutional neural network is trained to generate an anonymized image that minimizes correlation with the original PRNU pattern

while preserving high visual quality. This demonstrates that PRNU-based evidence, while physically interpretable, remains susceptible to sophisticated anti-forensic techniques.

3.1.3. Strengths and Limitations of Sensor-Based Traces

PRNU constitutes one of the most direct and physically interpretable forms of evidence in media forensics, enabling conclusions such as the following: a given region was not captured by the same device as the rest of the image. Its strength lies in the clear causal link between the signal and a specific physical camera sensor [9]. Early work, such as that by Koopman et al., exploited this property by detecting disruptions in the spatial or temporal consistency of the PRNU pattern [13]. However, reliance on a physical imaging device becomes a limitation when addressing modern Deepfakes, particularly fully synthetic content. When manipulated content is generated entirely by an algorithm, it lacks any physical camera origin and therefore carries no PRNU signal. The detection problem then shifts from identifying inconsistent PRNU to detecting its absence, a more challenging task, since many factors, including compression, resizing, or noise, can attenuate or remove the PRNU pattern. Furthermore, anti-forensic techniques such as DIPPAS [12] can deliberately suppress PRNU traces while preserving high visual fidelity. As a result, PRNU-based methods remain effective for partially manipulated forgeries, such as face swaps, but their reliability decreases significantly for fully synthetic media or under advanced adversarial attacks.

3.2. Convolutional Traces and Residual Analysis

As sensor-based traces proved less reliable for synthetic content, research attention shifted from physical device fingerprints to algorithmic fingerprints, unique artifacts left by the image generation process itself. These methods treat the generative pipeline as a source of forensic evidence.

3.2.1. Noiseprint

Cozzolino et al. [14] introduced Noiseprint, which improves on traditional hand-crafted PRNU features by replacing the fixed noise model with a learned representation. A Siamese network is trained on pairs of patches from the same or different cameras, learning to suppress image content while amplifying subtle model-specific artifacts. This learned camera model fingerprint is more robust to post-processing and can be used for forgery localization by detecting spatial inconsistencies within the Noiseprint map. Cozzolino et al. [15] extended the concept to video, introducing a video noiseprint derived from temporal sequences of patches. This method enables both camera model identification and the localization of temporal manipulations.

3.2.2. Exposing the Generative Traces of Convolutional Networks

Guarnera et al. [16] targeted artifacts inherent to GAN-based image generation, particularly those introduced by transposed convolution layers during upsampling, termed convolutional traces. These traces manifest as distinctive local correlation patterns determined by the generator architecture. Using an Expectation–Maximization algorithm, they extracted feature vectors capturing these correlations, enabling the creation of architecture-specific fingerprints. This approach can determine authenticity and, in some cases, identify the specific GAN model responsible (e.g., StyleGAN vs. StarGAN). Later work demonstrated that convolutional traces are robust to operations such as compression and rotation and are independent of semantic image content, making the method applicable to both facial and nonfacial images [17].

3.2.3. Separating Tampering Traces via Residual Learning

Residual learning techniques aim to separate the natural image signal from manipulation artifacts. Zhang et al. [18] pioneered this approach with DnCNN, a CNN designed to predict the residual (noise) rather than the clean image. This implicitly isolates clean content in the intermediate features of the network.

Motivated by this concept, Guo et al. [19] introduced GRnet, featuring a Manipulation Trace Extractor (MTE) that leverages guided filtering to retain genuine content while extracting detailed residual information. By fusing residual domain and spatial domain features through an attention-fusion mechanism, GRnet achieves enhanced resilience to common degradations, including low resolution.

Chen et al. [20] further formalized this approach in the SNIS network, explicitly framing post-processed forgery detection as a signal–noise separation problem. Their method isolates manipulated regions from background noise, improving resilience to compression and blurring.

3.2.4. From Physical to Algorithmic Fingerprints

This research trajectory reflects a conceptual shift in forensic explainability: from physical, device-specific artifacts (e.g., PRNU) to statistical, algorithm-specific artifacts (e.g., convolutional traces, learned residuals). *Noiseprint* represents a transitional stage, learning camera model fingerprints that are more abstract than PRNU but still tied to device classes [14]. Guarnera et al. [17] advance this further by directly modeling the fingerprints of the GAN architecture, changing the explanatory statement from "captured by camera X" to "generated by algorithm Y". Residual-based methods such as GRnet and SNIS [19,20] generalize this approach, identifying manipulation artifacts regardless of the generating algorithm. Here, the explanation becomes the following: manipulated regions contain statistical inconsistencies separable from natural image content. This evolution demonstrates increasing abstraction in explainability, from concrete physical traces to generalized statistical anomalies introduced by digital synthesis or editing.

3.3. Frequency-Domain and Re-Synthesis Methods

These methods examine representations beyond the spatial domain, such as frequency spectra, or actively probe the image through further transformations to reveal inconsistencies.

3.3.1. Detecting Artifacts Beyond the Visual Spectrum

Early generative models, particularly those employing transposed convolutions, often produced periodic artifacts in the frequency domain. Early detection strategies leveraged transformations such as the Discrete Fourier Transform (DFT) to reveal these spectral patterns, laying the foundation for later, more robust approaches.

3.3.2. Probing Forgeries via Re-Synthesis and Contextual Discrepancies

He et al. [21] argued that reliance on static frequency artifacts is unsustainable as generative quality improves. This concern is valid, as many early frequency-based methods focused on specific, tell-tale signs like upsampling artifacts introduced by transposed convolutions, which newer generator architectures may avoid. They proposed a re-synthesis framework, passing the test image through a model trained exclusively on authentic data (e.g., super-resolution or denoising). Real images yield low reconstruction error, whereas fake images, drawn from a different distribution, produce higher and more structured residuals. Detection is based on these residual patterns, providing a distributional inconsistency explanation.

Nirkin et al. [22] targeted semantic inconsistencies in face-swap forgeries. They trained separate recognition networks for the inner facial region and the surrounding context. Discrepancies between the two identity predictions strongly indicate a swap. The explanation is intuitive: the identity of the face does not match the identity cues from the context.

3.3.3. The Role of Transformed Domains and Generative Probing

These methods signal a shift from passive observation of artifacts to active interrogation of an image. Re-synthesis [21] evaluates an image's compatibility with models trained on genuine data, while contextual identity checks [22] test for semantic contradictions. As synthetic media becomes increasingly artifact-free, active probing strategies are likely to become central to detection, focusing on logical and statistical inconsistencies rather than fixed forensic markers.

3.4. Feature Decoupling and Contrastive Learning

Advanced representation learning methods aim to construct feature spaces that naturally separate real and manipulated content, yielding explanations derived from the learned embedding structure.

3.4.1. Learning Generalizable Forgery Features with Contrastive Learning

Xu et al. [23] applied Supervised Contrastive Learning (SupCon) to enhance generalization in Deepfake detection. The SupCon loss encourages intra-class compactness and inter-class separation in the learned feature space. By contrasting authentic images with a diverse set of forgeries, the model learns a robust representation of authenticity. Heatmap visualizations reveal that the network often attends to facial boundaries where manipulations are introduced, offering spatial interpretability. The Dual Contrastive Learning (DCL) framework [24] modifies the standard classification objective by operating at two granularities. First, Inter-Instance Contrastive Learning (Inter-ICL) is designed to learn a globally discriminative feature space by increasing the similarity between representations of authentic images while decreasing their similarity to those of forged images. Second, Intra-Instance Contrastive Learning (Intra-ICL) addresses local inconsistencies within a single forged image by contrasting features from manipulated regions against those from original regions. This dual mechanism is intended to produce a more generalizable representation by training the model to be sensitive to both global authenticity and local artifacts.

3.4.2. Explaining Detection via Source–Target Image Matching

Dong et al. [25] analyzed Deepfake detectors through the lens of source–target forgery relationships, constructing fake-source-target (FST) image triplets to study model responses under different matching conditions (FST-Matching). They proposed that detectors implicitly learn artifact-related visual concepts via this relational structure and designed a model that explicitly incorporates such matching. This improvement is particularly noticeable for heavily compressed videos.

3.4.3. Separating the Forgery Signal

Unlike methods that search for predefined artifacts, contrastive learning approaches define the objective in terms of feature-space geometry: maximize the distance between real and fake representations. Interpretability arises from post hoc analysis of model attention and activation patterns, revealing which features drive this separation. This reflects a higher level of abstraction in forensic reasoning, focusing on the learned discriminative features rather than fixed, physically motivated traces, allowing adaptation to novel manipulation techniques.

The effectiveness of such feature space separation is often visually explained using dimensionality reduction techniques like t-Distributed Stochastic Neighbor Embedding (t-SNE). By projecting the high-dimensional feature embeddings of real and fake images into a two- or three-dimensional map, t-SNE provides an intuitive visualization of whether a model has successfully learned to organize authentic and manipulated content into distinct clusters. This serves as an effective diagnostic tool for interpreting the structure of the model's learned representation.

4. Model-Centric and Interpretable-by-Design Methods

This chapter shifts the focus from artifacts embedded in the data to the decision-making process of the detection model itself. It reviews two main paradigms, namely the use of post hoc techniques to explain existing black-box models and the design of inherently transparent model architectures.

4.1. Visualization and Post Hoc Explanations

These approaches typically begin with a pre-trained, high-performance deep learning model and apply external tools to generate explanations for its predictions.

4.1.1. Application of Gradient and Perturbation Methods

A significant body of work has adapted general-purpose XAI techniques for Deepfake detection. Among the most popular of these are gradient-based attribution methods such as Gradient-weighted class activation mapping (Grad-CAM) [26]. This technique produces a visual explanation in the form of a heatmap, highlighting the input regions that are most influential for a specific prediction. In the context of deepfake forensics, Grad-CAM is widely used to reveal which facial regions a detector focuses on when classifying an image as fake.

Malolan et al. [27] proposed a framework utilizing visual interpretability methods such as saliency maps and guided backpropagation to reveal which facial regions a CNN relies on in its decision-making process. They further compared white-box, black-box, and model-specific techniques, and adapted SHAP and Grad-CAM to explain an EfficientNet-based detector. Ge et al. [28] demonstrated the utility of SHAP for analyzing both spoofing and Deepfake detectors. Their study showed how SHAP can expose unexpected model behaviors, such as reliance on non-speech intervals in audio, identify the most contributory artifacts, and highlight behavioral differences between competing classifiers. Parvez [29] proposed combining a CNN with a CapsuleNet, integrated with Grad-CAM to visualize the input regions most important for prediction, thereby enhancing model transparency and accountability.

4.1.2. The Utility and Pitfalls of Post Hoc Explanations

The primary value of post hoc explanations lies not in providing definitive forensic evidence but in serving as diagnostic tools for model developers and as confidence-building mechanisms for end-users. These methods can expose whether a model is focusing on plausible cues, but their fidelity is limited, as the explanation itself is only an approximation of the model's internal processes. The opacity of deep learning detectors creates a trust gap, which post hoc tools such as Grad-CAM and SHAP attempt to mitigate by providing heatmaps or feature importance scores that suggest why a model reached a given decision [27,30]. Ge et al. [28] highlighted the critical role of this approach in discovering when a model exploits spurious correlations, such as irrelevant audio features.

However, explanations derived from gradient-based methods remain indirect. As Gowrisankar et al. [31] emphasized, standard XAI evaluation frameworks may not be suitable for Deepfake detection, and the validity of explanations must themselves be

scrutinized. Overall, post hoc explanations represent an important first step toward transparency, but their role is primarily to verify that the model attends to meaningful input regions rather than to provide direct forensic proof. This limitation motivates the design of interpretable-by-architecture models.

4.2. Attention Mechanisms and Transformers

Attention-based methods incorporate explainability directly into the model architecture. By design, the model outputs attention weights that indicate which parts of the input most influenced its decision.

4.2.1. Self-Attention and Cross-Modal Attention for Forgery Localization

Asha et al. [32] introduced a defensive attention mechanism for detecting multimodal Deepfakes involving both audio and video. Their system employs a self-attentive VGG16 for visual features and a self-attentive RNN for audio features. The core innovation is a cross-modal attention mechanism that quantifies discrepancies between audio and video streams, simultaneously supporting detection and interoperability. Another line of research shifts focus from low-level visual artifacts to higher-level semantic inconsistencies. The Voice-Face Homogeneity approach [33], for example, operates on the premise that an individual's voice and face share a biometric link. It uses a pre-trained speaker-face matching model to determine if the audio speaker is the same person as the individual depicted visually. By targeting a potential identity mismatch in face-swap forgeries, rather than specific visual artifacts, this method is designed to generalize to novel manipulation techniques.

4.2.2. Case Study

Qi et al. [34] proposed DeepRhythm, which detects Deepfakes by exploiting disruptions in remote photoplethysmography (rPPG) signals, corresponding to subtle heartbeat rhythms observable in facial videos. To achieve robust detection, they designed a dual spatio-temporal attention network that learns both spatial regions and temporal segments most reliable for rPPG extraction. The resulting attention maps not only drive the detection process but also provide intuitive explanations of where and when the model relies on physiological cues.

4.2.3. Building Intrinsic Focus into Detection Models

Unlike post hoc explanations, attention mechanisms provide explanations that are integral to the model's computations. Post hoc tools approximate importance retrospectively, whereas attention-based models explicitly assign weights to input regions during training. In DeepRhythm [34], for example, the model learns which facial areas are most informative for rPPG extraction, and the attention map functions both as a signal-processing step and as an explanation. Similarly, in Asha et al. [32], cross-modal attention weights directly quantify consistency between audio and video streams, making them both computationally essential and interpretable. This tight coupling of explanation and decision processes makes attention-based methods more faithful to the model's reasoning.

4.3. Prototype-Based and Case-Based Reasoning

Prototype-based methods aim to provide explanations in terms of similarity to representative cases, mirroring forms of human reasoning that rely on precedents.

4.3.1. Explaining Decisions with Dynamic and Learned Prototypes

Trinh et al. [7] introduced the Dynamic Prototype Network (DPNet), which employs learnable prototypes to capture temporal artifacts in Deepfake videos. Instead of returning only a binary classification, the model explains a decision by comparing the input to a

set of class-specific prototypes. For example, an unnatural head movement is detected as fake because it closely resembles a learned prototype of Deepfake head motion. PUDD (Prototype-based Unified Framework for Deepfake Detection) [35] presents a similarity-driven approach to deepfake detection, where input data is compared to known prototypes for classification. The system identifies potential deepfakes and previously unseen classes by analyzing similarity drops. PUDD integrates image classification as an upstream task during training, allowing it to perform well in both image classification and deepfake detection. This approach offers notable efficiency and environmental benefits, as it requires minimal retraining time and has a significantly lower carbon footprint compared to existing models.

4.3.2. Human-in-the-Loop Refinement with Visual Analytics

den Bouter et al. [36] addressed challenges in prototype interpretability, such as redundancy and poor human comprehensibility, by developing ProtoExplorer, a visual analytics tool for forensic experts. The system enables exploration and refinement of prototype sets, including visualization of spatio-temporal prototypes, interactive filtering of predictions, and removal of uninformative prototypes. This human-in-the-loop process improves interpretability while maintaining detection accuracy.

4.3.3. Towards Case-Based Explainability

Prototype methods shift the explanation from abstract features to concrete archetypal cases. Trinh et al. [7] demonstrated that decisions can be explained by reference to representative examples, offering a precedent-based rationale closer to human reasoning. Whereas saliency maps highlight important regions, prototype methods categorize anomalies by comparison to known cases. The quality of these prototypes is therefore crucial. ProtoExplorer [36] exemplifies how expert-guided curation of prototype sets can ensure interpretability and non-redundancy. Such case-based reasoning offers a more tangible form of evidence for forensic analysts, who can inspect the referenced prototype to understand the detected artifact.

4.4. Other Interpretable Architectures

Shang et al. [37] introduced PRRNet, which captures inconsistencies between local facial regions and global context. By explicitly modeling the relationships between pixel-level and region-level representations, the model highlights disharmonies characteristic of forgeries. The interpretability arises from its architectural design, which attends to these discrepancies. Soltandoost et al. [38] proposed a method to extract local explanations from global representations. This architecture enables decomposition of holistic feature vectors into localized evidence, bridging the gap between strong classification performance and interpretable outputs. Recent work has also explored architectures that address considerations such as the computational cost and parameter efficiency of deploying large models. The MoE-FFD framework [39], for instance, adapts a pre-trained Vision Transformer (ViT) for face forgery detection using a Mixture of Experts (MoE) approach combined with Parameter-Efficient Fine-Tuning (PEFT) techniques like Low-Rank Adaptation (LoRA). In this architecture, a gating network dynamically selects a subset of 'expert' models to process an input image, while the ViT backbone remains frozen. While the primary goal is efficiency, this modular design provides a mechanism for potential interpretability; for example, analyzing which experts are activated for different forgery types may offer insight into model specialization.

5. Multimodal and Language-Based Explanations

This chapter discusses a recent and significant development in the field: the transition from purely visual or feature-based explanations to natural language explanations enabled by LMMs.

5.1. Textual Explanations and Visual Question Answering (VQA) Frameworks

To enhance both interpretability and robustness in Deepfake detection, recent works have increasingly adopted VQA-style formulations that jointly address classification and explanation. Kundu et al. [40] proposed TruthLens, a framework that formulates Deepfake detection as a VQA task. The model provides not only binary classification but also detailed textual reasoning for its predictions, and it is capable of addressing fine-grained queries such as whether a specific facial attribute appears authentic. The architecture is hybrid, combining a vision-only model for local feature extraction with a Large Multimodal Model (LMM) for global context understanding and text generation. This design enables the system to detect subtle artifacts and explain them in natural language. Guo et al. [41] introduced M2F2-Det, a multimodal detector that jointly generates detection scores and textual explanations. The model leverages the visual representation capabilities of a pre-trained CLIP model together with the generative ability of a Large Language Model (LLM), thereby linking subtle forgery-related visual cues with natural language descriptions. Jia et al. [42] emphasized the role of common-sense reasoning in detection. They argued that many Deepfakes violate basic perceptual rules, such as inconsistent hairlines or unnatural skin shading, which are readily identified by humans but often overlooked by conventional CNN-based approaches. To address this, they introduced the DD-VQA dataset, which pairs images with questions and detailed, commonsense-based textual explanations. Their framework employs a vision-and-language transformer to execute the VQA task. Yan et al. [43] proposed X^2-DFD, a framework for systematically assessing and enhancing the Deepfake detection capabilities of LMMs. The framework comprises three modules: Model Feature Assessment (MFA), which evaluates an LMM's ability to capture different forgery features; Strong Feature Strengthening (SFS), which fine-tunes the model on features where it already demonstrates competence; and Weak Feature Supplementation (WFS), which incorporates external detectors for features where the LMM performs poorly. This structured approach yields a more robust and interpretable hybrid system.

Huang et al. [44] developed SIDA, a framework tailored to social media platforms that performs detection, localization, and explanation simultaneously. Leveraging an LMM, SIDA produces not only a classification output but also a segmentation mask of manipulated regions together with a textual justification. They also introduced SID-Set, a large-scale dataset designed specifically for this multi-task setting. Following this direction, recent forensics-driven frameworks such as Propose and Rectify [45] utilize the capabilities of Multimodal Large Language Models (MLLMs). These models are designed to first propose potential forgery regions and then rectify their localization, demonstrating a multi-step process that aims to improve the precision of manipulation detection. LayLens [46] is a user-centric tool for deepfake forensics that combines explainable forgery localization, natural language simplification, and visual reconstruction. It translates complex model reasoning into accessible explanations for non-experts while maintaining technical depth, and presents side-by-side comparisons of original and reconstructed images. User studies demonstrate that LayLens improves clarity, reduces cognitive load, and enhances trust in deepfake detection.

5.2. The Paradigm Shift to Natural Language Explanations

The introduction of Large Vision-Language Models (LVLMs), also referred to as LMMs, marks a methodological shift in deepfake detection. While earlier methods often treated

explainability as a separate, post hoc analysis, LVLM-based frameworks integrate explanation directly into their primary output. The task is consequently reoriented from a focus on binary classification toward a process of generating reasoned explanations for why an image is considered authentic or manipulated. This aligns the model's output more closely with human-centric, evidence-based analysis. Traditional forensic and model-centric approaches, such as those relying on PRNU correlation plots, heatmaps, or prototype analysis, generate explanations that typically require expert interpretation. In contrast, LMMs can integrate visual and linguistic reasoning to produce directly human-readable explanations.

Jia et al. [42] highlight the discrepancy between machine and human reasoning: while models excel at detecting low-level statistical artifacts, humans rely heavily on high-level common-sense reasoning. The outputs of these LVLM-based systems differ from those of traditional methods. For example, an explanation from a PRNU analysis may be a correlation plot, while a Grad-CAM output is a heatmap. Such visualizations typically require domain expertise for interpretation. In contrast, LVLMs can generate explanations in natural language. For instance, they may describe an inconsistency such as "the texture of the skin in the cheek area appears artificially smooth." This form of output is more directly accessible to non-expert users such as journalists or content moderators. LMMs represent the generation of models capable of bridging this gap by articulating observations such as unrealistic texture or improper alignment of facial features in natural language. This reframes the detection task. Instead of training a classifier and then constructing an auxiliary explanation mechanism, models such as TruthLens and M2F2-Det perform detection through the process of explanation itself [40,41].

This shift has substantial implications for the field. It suggests that the most effective and trustworthy detection systems will not be simple classifiers but reasoning frameworks that articulate their findings in a transparent and accessible manner. The focus of research is therefore moving from accuracy alone to the quality, coherence, and faithfulness of generated explanations.

6. Evaluation Metrics and Datasets for Explainability

As explainability becomes a core objective, the development of systematic evaluation methods has become an important area of research. This section reviews work on metrics and datasets designed specifically to support explainability in Deepfake detection.

6.1. Metrics and Evaluation Frameworks

Standard metrics for evaluating deepfake detection performance typically include Accuracy, Area Under the Curve (AUC), and Equal Error Rate (EER) [6]. While these metrics effectively measure a model's classification accuracy, they do not address the interpretability of its decisions. This limitation motivates the development of the explainability-focused evaluation frameworks that this section details. Baldassarre et al. [47] noted that the evaluation of explanations in Deepfake detection has often relied on subjective visual inspection. To address this limitation, they proposed a set of quantitative, human-centric metrics tailored for video-based explanation heatmaps. These metrics capture two complementary aspects:

- Visual Quality: Heatmap interpretability is assessed through its structural properties, namely smoothness, spatial locality, and sparsity. Smoothness is evaluated using total variation, spatial locality is characterized by the volume of the covariance matrix, and sparsity is quantified via the Gini index. These metrics collectively reflect the clarity and coherence of the heatmap from a human perceptual perspective.
- Informativeness: The explanatory power of the heatmap is measured by its ability to accurately highlight manipulated regions, typically quantified by the overlap with

ground-truth manipulation masks. This overlap is commonly measured using standard segmentation metrics such as the Intersection over Union (IoU), which calculates the ratio of the intersection to the union of the predicted and ground-truth masks. This metric captures both the precision and relevance of the explanation.

Gowrisankar et al. [31] focused on the robustness of explanations. They argued that generic evaluation methods, such as random pixel removal, are not well-suited for Deepfake detectors. Instead, they proposed using adversarial attack methods to evaluate explanations in a more task-specific manner. Their approach tests whether explanations correctly identify features that are most vulnerable to adversarial perturbations capable of altering the model's decision, thereby providing an assessment of the faithfulness of the explanation to the underlying decision process.

The introduction of dedicated metrics and evaluation frameworks has advanced the evaluation of explainability in Deepfake detection from qualitative observation to systematic assessment. In earlier work, explanations were often presented through visual examples, without standardized measures for comparison. Baldassarre et al. [47] formalized evaluation by proposing metrics such as sparsity and spatial locality, enabling quantitative comparisons of the quality of visual explanations. Gowrisankar et al. [31] extended this line of work by emphasizing faithfulness and robustness, introducing adversarial evaluation methods that examine whether explanations remain consistent under targeted perturbations. Together, these contributions represent a shift toward reproducible and comparable evaluation of explainability methods in this domain.

6.2. Specialized Datasets with Annotations

Foundational benchmarks have been instrumental in advancing the accuracy of deepfake detectors. Key resources include FaceForensics++, a widely adopted dataset for benchmarking, and Celeb-DF [48], known for its high-realism forgeries. Additionally, large-scale efforts such as the Deepfake Detection Challenge (DFDC) [49] and the WildDeepfake [50] dataset have driven the development of more generalizable models. A key limitation of these benchmarks, however, is their general lack of the fine-grained annotations needed for interoperability.

To address this gap and support explainability research, a new generation of specialized datasets has been introduced. Miao et al. [51] introduced DDL, a large-scale dataset comprising 1.8 million samples, specifically designed for interpretable detection and localization. The dataset encompasses diverse scenarios (single- and multi-face, audio–visual content), a broad range of manipulation techniques (75 types), and fine-grained annotations, including spatial masks and temporal segments, which support both the training and evaluation of explainable models. Hondru et al. [52] released ExDDV, the first dataset specifically designed for explainable video Deepfake detection. In addition to localization masks, ExDDV provides human annotations in the form of clicks marking artifacts and textual descriptions explaining them. These annotations enable the development and benchmarking of models that produce multimodal explanations, including natural language. The DD-VQA dataset, introduced by Jia et al. [42], provides question–answer pairs designed to support textual explanations. The questions focus on reasoning about why a given image appears authentic or manipulated, linking visual cues to semantic justifications.

The evolution of datasets has closely followed and enabled methodological advances in explainable Deepfake detection. Early datasets typically contained only binary real/fake labels, supporting classification tasks but not interpretability. Later datasets introduced manipulation masks, allowing quantitative evaluation of localization but with limited support for explanation [47]. Recent datasets such as DDL [51] and ExDDV [52] are explicitly designed with explainability in mind, combining binary labels, pixel-level masks,

and human-generated textual descriptions. These resources provide the ground truth needed for training and evaluating models that aim not only to detect manipulations but also to localize and explain them. In particular, textual annotations form the basis for recent approaches that employ large multimodal models to generate natural language explanations. The availability of such datasets has therefore been a key factor in enabling the transition from classification-focused detection methods to approaches that emphasize reasoning and explanation.

7. Discussion and Future Directions

7.1. Evolution of Explainability Methods

Research on explainability in Deepfake detection has progressed along several stages. Early work concentrated on identifying physical or algorithmic artifacts of manipulation. Explanations at this stage were grounded in measurable irregularities, such as the PRNU patterns of camera sensors, or algorithmic traces introduced by generative architectures. These approaches connected the detection process to observable properties of either acquisition devices or synthesis methods.

As detection models became more complex, attention shifted from analyzing data artifacts to interpreting the models themselves. Post hoc methods such as Grad-CAM and SHAP were applied to visualize decision-making in trained models. In parallel, explainability was embedded directly into model architectures through mechanisms like attention, which indicates focus regions, or prototype-based learning, which relates predictions to known forgery cases. This development reframed explanations from pointing to artifacts in the data toward revealing the internal reasoning of the models.

Recent work has introduced a multimodal and language-based perspective. LMMs are increasingly used to generate natural language descriptions that articulate the reasoning behind classification outcomes. Frameworks such as TruthLens and DD-VQA treat the detection task not only as binary classification but as a process of producing textual explanations aligned with human understanding. This shift expands the role of explainability from providing technical cues for experts toward communicating results in a form accessible to broader audiences.

7.2. Challenges

A key difficulty concerns the balance between accuracy, interpretability, and computational efficiency. Simpler architectures tend to be more transparent, yet often lack the precision required to handle advanced forgeries. In contrast, highly accurate models usually remain opaque, which complicates the effort to provide explanations that are both faithful and efficient. Methods that can reconcile these competing objectives remain an open area of research.

Fidelity of explanations is also a persistent concern. Explanations should reflect the true reasoning process of the model rather than provide a post hoc narrative that appears plausible but is disconnected from the underlying decision. The reliability of methods such as Grad-CAM has been questioned in this respect. For example, Gowrisankar et al. [31] introduced adversarial evaluation protocols to test whether highlighted regions truly align with the model's reasoning. Robustness is equally important, since adversarial or anti-forensic manipulations can obscure forensic traces, such as deliberately suppressing PRNU signals.

Generalization across unseen forgery techniques presents another obstacle. Approaches that rely on specific artifacts, including PRNU patterns or convolutional traces, may lose effectiveness when confronted with manipulations produced by new generative architectures. Recent work has attempted to address this by focusing on universal forgery

representations or by framing unknown manipulations as out-of-distribution instances, though this remains an open problem.

Evaluation and usability form an additional dimension. A technically valid explanation may not necessarily assist human decision-making. Early studies frequently relied on subjective inspection without standardized benchmarks. Prototype-based methods, while offering case-based reasoning, sometimes generate redundant or unintuitive examples that require expert curation. Developing systematic evaluation protocols and presentation strategies that adapt to different user needs is essential for broader adoption.

Furthermore, the deployment of these detection systems introduces significant regulatory and ethical considerations. Emerging legal frameworks for artificial intelligence, for instance, are beginning to mandate transparency for high-risk systems, which would include tools used for verifying forensic evidence or news content. The admissibility of AI-generated explanations in a court of law, the ethical responsibility of news organizations to audit their verification algorithms, and the potential for algorithmic bias in detection models are all critical challenges that require interdisciplinary attention beyond the purely technical domain.

7.3. Opportunities

Hybrid and Multimodal Explainability. Future research can integrate multiple layers of evidence into unified frameworks. LMMs can be used to interpret low-level forensic artifacts such as PRNU inconsistencies or convolutional traces, while simultaneously generating high-level semantic explanations. This type of hybrid system retains the rigor of forensic evidence while presenting results in a human-readable form. The X^2-DFD framework, which incorporates external detectors to complement large models, illustrates an initial step in this direction.

Causal Reasoning and Robustness of Explanations. Moving beyond correlation-based attribution, causal inference provides a way to relate model outputs to underlying physical or semantic inconsistencies, such as deviations from lighting laws or natural dynamics. Explanation methods built on causal principles have the potential to improve reliability by clarifying why a decision was made rather than only where a model focused. At the same time, anti-forensic techniques highlight the need for explanations that are robust by design against adversarial manipulations intended to obscure forensic traces.

Continual Learning for Open-World Scenarios. Static detectors often fail when new forgery techniques emerge. Explainable models that incorporate continual, zero-shot, or few-shot learning could adapt dynamically to novel manipulations while preserving interpretability. Such approaches would allow systems to evolve in parallel with generative models rather than requiring retraining from scratch.

Interactive and User-Centered Explainable Systems. Explanation can extend beyond static outputs to interactive systems that allow users to query and refine results. Forensic experts may require the ability to trace specific artifacts, while general users may prefer concise textual descriptions. Systems such as ProtoExplorer demonstrate how prototypes can be curated and refined interactively, and VQA-based frameworks like TruthLens show the potential of dialog-style querying. These directions point toward explainable systems that adapt to the expertise and needs of different users.

Proactive Forensics and Content Provenance. Beyond passive detection approaches, researchers have investigated proactive authentication strategies. The FractalForensics framework [53] is an example of this approach, which involves embedding an imperceptible, semi-fragile watermark into an image before its distribution. The watermark is designed to be robust to common operations like compression while being fragile to manipulations such as face swapping. In this framework, the integrity of the watermark serves as evidence

of authenticity. This method also provides a form of direct explanation, as damaged or missing sections of the watermark serve to localize the manipulated regions.

Bridging Research and Practice in Real-World Scenarios. A gap remains between the evaluation of explainable methods on benchmark datasets and their practical implementation in high-stakes environments, such as forensic investigations or newsrooms. Future work should prioritize closing this gap by conducting case studies on real-world legal or journalistic corpora. Furthermore, there is a need to develop evaluation metrics that extend beyond algorithmic performance to measure practical utility, such as the time required for a human expert to reach a conclusion with AI assistance, or the overall impact of the explanation on the decision-making process.

8. Conclusions

The research paradigm in Deepfake detection has fundamentally shifted, as classification accuracy alone no longer satisfies real-world demands for credibility. This paper has systematically traced the evolution of explainability, from methods grounded in forensic data artifacts to the interpretation of model behavior, culminating in the current paradigm of semantic explanation via large multimodal models. The future trajectory of the field, therefore, will be defined not just by enhancing detection performance, but by developing transparent and reliable explanation mechanisms that foster human–machine synergy in safeguarding information integrity. This necessitates that future research afford equal weight to a model's predictive power and the fidelity and usability of its explanations, a foundational strategy for confronting the challenges of digital disinformation.

Author Contributions: Conceptualization, H.Q. and L.X.; methodology, H.Q.; investigation, R.G., Y.F. and Q.W.; data curation, R.G. and Y.F.; writing, R.G., Y.F., H.Q. and L.X.; supervision, L.X. and Z.J.; funding acquisition, L.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the MOE Industry-University Collaborative Education Project: "AI Security Testing: Risk Prevention and Control in Deep Learning Models", the Philosophy and Social Sciences Research Project of Universities in Jiangsu Province (2024SJYB0345, 2023SJYB0464, 2023SJYB0468), and the "Cyberspace Security" construction project of key disciplines in Jiangsu Province during the "14th Five-Year Plan".

Acknowledgments: The authors sincerely thank the anonymous reviewers and the editor for their constructive feedback and insightful suggestions, which significantly enhanced the clarity and quality of this work.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GAN	Generative Adversarial Networks
LLMs	Large Multimodal Models
explainable AI	XAI
PRNU	Photo Response Non-Uniformity
FAR	False Acceptance Rate
PCE	Peak-to-Correlation Energy
DIP	Deep Image Prior
MTE	Manipulation Trace Extractor
DFT	Discrete Fourier Transform

SupCon	Supervised Contrastive Learning
rPPG	remote photoplethysmography
DPNet	Dynamic Prototype Network
VQA	Visual Question Answering
LLM	Large Language Model
MFA	Model Feature Assessment
SFS	Strong Feature Strengthening
WFS	Weak Feature Supplementation
FST	Fake-Source-Target

References

1. Tolosana, R.; Vera-Rodríguez, R.; Fiérrez, J.; Morales, A.; Ortega-Garcia, J. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Inf. Fusion* **2020**, *64*, 131–148. [CrossRef]
2. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–11. [CrossRef]
3. Li, Y.; Lyu, S. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, 16–20 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 46–52.
4. Ganguly, S.; Ganguly, A.; Mohiuddin, S.; Malakar, S.; Sarkar, R. ViXNet: Vision Transformer with Xception Network for deepfakes based video and image forgery detection. *Expert Syst. Appl.* **2022**, *210*, 118423. [CrossRef]
5. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R., Eds.; ACM: New York, NY, USA, 2016; pp. 1135–1144. [CrossRef]
6. Wang, T.; Liao, X.; Chow, K.; Lin, X.; Wang, Y. Deepfake Detection: A Comprehensive Survey from the Reliability Perspective. *ACM Comput. Surv.* **2025**, *57*, 58:1–58:35. [CrossRef]
7. Trinh, L.; Tsang, M.; Rambhatla, S.; Liu, Y. Interpretable and Trustworthy Deepfake Detection via Dynamic Prototypes. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, 3–8 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1972–1982. [CrossRef]
8. Tariq, S.; Woo, S.S.; Singh, P.; Irmalasari, I.; Gupta, S.; Gupta, D. From Prediction to Explanation: Multimodal, Explainable, and Interactive Deepfake Detection Framework for Non-Expert Users. *arXiv* **2025**, arXiv:2508.07596.
9. Lukas, J.; Fridrich, J.; Goljan, M. Digital camera identification from sensor pattern noise. *IEEE Trans. Inf. Forensics Secur.* **2006**, *1*, 205–214. [CrossRef]
10. Marra, F.; Poggi, G.; Sansone, C.; Verdoliva, L. Blind PRNU-based image clustering for source identification. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 2197–2210. [CrossRef]
11. Saito, S.; Tomioka, Y.; Kitazawa, H. A theoretical framework for estimating false acceptance rate of PRNU-based camera identification. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 2026–2035. [CrossRef]
12. Picetti, F.; Mandelli, S.; Bestagini, P.; Lipari, V.; Tubaro, S. DIPPAS: A deep image prior PRNU anonymization scheme. *EURASIP J. Inf. Secur.* **2022**, *2022*, 1–21. [CrossRef]
13. Koopman, M.; Rodriguez, A.M.; Geradts, Z. Detection of deepfake video manipulation. In Proceedings of the 20th Irish Machine Vision and Image Processing Conference (IMVIP), Belfast, UK, 29–31 August 2018; pp. 133–136.
14. Cozzolino, D.; Verdoliva, L. Noiseprint: A CNN-Based Camera Model Fingerprint. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 144–159. [CrossRef]
15. Cozzolino, D.; Poggi, G.; Verdoliva, L. Extracting camera-based fingerprints for video forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–20 June 2019; pp. 941–949.
16. Guarnera, L.; Giudice, O.; Battiato, S. DeepFake detection by analyzing convolutional traces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 666–667. [CrossRef]
17. Guarnera, L.; Giudice, O.; Battiato, S. Fighting deepfake by exposing the convolutional traces on images. *IEEE Access* **2020**, *8*, 165085–165098. [CrossRef]
18. Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; Zhang, L. Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Trans. Image Process.* **2017**, *26*, 3142–3155. [CrossRef]

19. Guo, Z.; Yang, G.; Chen, J.; Sun, X. Exposing Deepfake Face Forgeries with Guided Residuals. *IEEE Trans. Multimed.* **2022**, *25*, 3336–3348. [CrossRef]
20. Chen, J.; Liao, X.; Wang, W.; Qian, Z.; Qin, Z.; Wang, Y. SNIS: A signal noise separation-based network for post-processed image forgery detection. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 935–951. [CrossRef]
21. He, Y.; Yu, N.; Keuper, M.; Fritz, M. Beyond the spectrum: Detecting deepfakes via re-synthesis. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Montreal, QC, Canada, 19–27 August 2021; pp. 768–775. [CrossRef]
22. Nirkin, Y.; Wolf, L.; Keller, Y.; Hassner, T. DeepFake Detection Based on Discrepancies Between Faces and Their Context. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *45*, 1138–1151. [CrossRef]
23. Xu, Y.; Raja, K.; Pedersen, M. Supervised Contrastive Learning for Generalizable and Explainable DeepFakes Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 4–8 January 2022; pp. 379–389. [CrossRef]
24. Sun, K.; Yao, T.; Chen, S.; Ding, S.; Li, J.; Ji, R. Dual Contrastive Learning for General Face Forgery Detection. In Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; AAAI Press: Washington, DC, USA, 2022; pp. 2316–2324. [CrossRef]
25. Dong, S.; Wang, J.; Liang, J.; Fan, H.; Ji, R. Explaining deepfake detection by analysing image matching. *arXiv* **2022**, arXiv:2207.09679.
26. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [CrossRef]
27. Malolan, B.; Parekh, A.; Kazi, F. Explainable Deep-Fake Detection Using Visual Interpretability Methods. In Proceedings of the International Congress on Information and Communication Technology, London, UK, 9–12 February 2020; pp. 337–346. [CrossRef]
28. Ge, W.; Patino, J.; Todisco, M.; Evans, N. Explaining deep learning models for spoofing and deepfake detection with SHapley Additive exPlanations. In Proceedings of the Odyssey 2022: The Speaker and Language Recognition Workshop, Beijing, China, 28 June–1 July 2022; pp. 26–33.
29. Parvez, M.A.I. Explainable Deepfake Video Detection using Convolutional Neural Network and CapsuleNet. *arXiv* **2024**, arXiv:2404.12841.
30. Khan, S. Explainable Deepfake Detection Using Deep Learning. Master's Thesis, NED University of Engineering and Technology, Karachi, Pakistan, 2022. [CrossRef]
31. Gowrisankar, B.; Thing, V.L. An adversarial attack approach for eXplainable AI evaluation on deepfake detection models. *arXiv* **2023**, arXiv:2303.11993.
32. Asha, S.; Vinod, P.; Menon, V.G. A defensive attention mechanism to detect deepfake content across multiple modalities. *Multimed. Syst.* **2024**, *30*, 56. [CrossRef]
33. Cheng, H.; Guo, Y.; Wang, T.; Li, Q.; Chang, X.; Nie, L. Voice-Face Homogeneity Tells Deepfake. *ACM Trans. Multim. Comput. Commun. Appl.* **2024**, *20*, 76:1–76:22. [CrossRef]
34. Qi, H.; Guo, Q.; Juefei-Xu, F.; Li, X.; Yang, Y.; Zuo, M.; Sun, X.; Liu, J.; Feng, J.; Pu, G.; et al. DeepRhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 4317–4326. [CrossRef]
35. Pellcier, A.L.; Li, Y.; Angelov, P. PUDD: Prototype-based Unified Framework for Deepfake Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2023; pp. 3345–3354. [CrossRef]
36. den Bouter, M.d.L.; Pardo, J.L.; Geradts, Z.; Worring, M. ProtoExplorer: Interpretable Forensic Analysis of Deepfake Videos using Prototype Exploration and Refinement. *Inf. Visual.* **2024**, *23*, 239–257. [CrossRef]
37. Shang, Z.; Xie, H.; Zha, Z.; Yu, L.; Li, Y.; Zhang, Y. PRRNet: Pixel-region relation network for face forgery detection. *Pattern Recognit.* **2021**, *116*, 107950. [CrossRef]
38. Soltandoost, E.; Plesh, R.; Schuckers, S.; Peer, P.; Štruc, V. Extracting Local Information from Global Representations for Interpretable Deepfake Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Tucson, AZ, USA, 28 February–4 March 2025. [CrossRef]
39. Kong, C.; Luo, A.; Xia, S.; Yu, Y.; Li, H.; Kot, A.C. MoE-FFD: Mixture of Experts for Generalized and Parameter-Efficient Face Forgery Detection. In Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 5010–5014. [CrossRef]
40. Kundu, R.; Jia, S.; Mohanty, V.; Balachandran, A.; Roy-Chowdhury, A.K. TruthLens: Explainable DeepFake Detection for Face Manipulated and Fully Synthetic Data. *arXiv* **2024**, arXiv:2403.15867.
41. Guo, X.; Song, X.; Zhang, Y.; Liu, X.; Liu, X. Rethinking Vision-Language Model in Face Forensics: Multi-Modal Interpretable Forged Face Detector. *arXiv* **2024**, arXiv:2404.14811.

42. Jia, S.; Kundu, R.; Balachandran, A.; Moreira, D.; Roy-Chowdhury, A.K. Common Sense Reasoning for Deepfake Detection. *arXiv* **2024**, arXiv:2406.03425.
43. Yan, H.; Han, J.; Jiang, Y.; Yan, B.; Liu, J.; Wang, X.; Lyu, S. X^2-DFD: A framework for eXplainable and eXtendable Deepfake Detection. *arXiv* **2024**, arXiv:2402.06126.
44. Huang, C.; Huang, Z.; Chen, Z.; Huang, J.; Liu, Z.; Tao, D. SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model. *arXiv* **2024**, arXiv:2412.04292.
45. Zhang, K.; Kong, C.; Liu, H.; Ding, B.; Jiang, X.; Li, H. Propose and Rectify: A Forensics-Driven MLLM Framework for Image Manipulation Localization. *arXiv* **2025**, arXiv:2508.17976.
46. Narang, A.; Gupta, P.; Su, L.; Dhall, A. LayLens: Improving Deepfake Understanding through Simplified Explanations. *arXiv* **2025**, arXiv:2507.10066.
47. Baldassarre, F.; Debard, Q.; Pontiveros, G.F.; Wijaya, T.K. Quantitative metrics for evaluating explanations of video deepfake detectors. In Proceedings of the European Conference on Computer Vision Workshops (ECCVW), Tel Aviv, Israel, 23–27 October 2022; pp. 401–417.
48. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; IEEE: Piscataway, NJ, USA; pp. 3204–3213. [CrossRef]
49. Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; Canton-Ferrer, C. The DeepFake Detection Challenge (DFDC) Dataset. *arXiv* **2020**, arXiv:2006.07397.
50. Zi, B.; Chang, M.; Chen, J.; Ma, X.; Jiang, Y. WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2382–2390. [CrossRef]
51. Miao, C.; Chen, Y.; Zhang, Z.; Liu, Z.; Wang, Z.; Wang, S.; Feng, J. DDL: A Dataset for Interpretable Deepfake Detection and Localization in Real-World Scenarios. *arXiv* **2024**, arXiv:2406.13292.
52. Hondru, V.; Hogea, E.; Onchis, D.M.; Ionescu, R.T. ExDDV: A New Dataset for Explainable Deepfake Detection in Video. *arXiv* **2025**, arXiv:2503.14421.
53. Wang, T.; Cheng, H.; Liu, M.; Kankanhalli, M.S. FractalForensics: Proactive Deepfake Detection and Localization via Fractal Watermarks. *arXiv* **2025**, arXiv:2504.09451.