

A Comprehensive Survey on AI-Generated Image and Deepfake Detection: From Generative Adversarial Networks to Diffusion Models

Student, **Mr. Shourya Singh Rajput**

School of sciences, KK Modi University, Durg

HOD, **Dr. Mohammed Bakhtawar Ahmed**

School of sciences, KK Modi University, Durg

Abstract

The unprecedented advancement in deep generative models has democratized the creation of hyper-realistic artificial media, fundamentally transforming the landscape of digital forensics. Artificial intelligence-synthesized content, widely recognized as "Deepfakes," poses significant threats to privacy, democracy, and national security by enabling the frictionless generation of disinformation, financial fraud, and non-consensual fabricated media. While early synthesis methods relied on traditional computer graphics and basic autoencoder architectures, contemporary methodologies leverage sophisticated Generative Adversarial Networks (GANs) and state-of-the-art Text-to-Image Diffusion Models, rendering human visual inspection inadequate for authentication. Consequently, an adversarial battleground has emerged between malicious actors creating forged media and forensic researchers developing automated detection algorithms. This paper presents a comprehensive review of the evolutionary trajectory of AI-generated images and video deepfakes, alongside the taxonomy of corresponding detection methodologies. We systematically analyze spatial-domain analysis, frequency-domain artifacts, and biological signal inconsistencies utilized by modern detectors. Furthermore, we critically examine the generalization crisis faced by current deep neural network (DNN) classifiers when confronted with unseen generative architectures, particularly the recent paradigm shift from GANs to Diffusion models. By evaluating benchmark datasets, evasion techniques, and state-of-the-art countermeasures, this review provides a cohesive roadmap for future research in multimedia forensics and universal fake image detection.

Keyword

Deepfakes, Media Forensics, Generative Adversarial Networks (GANs), Diffusion Models, Image Manipulation Detection, Artificial Intelligence Security, Deep Learning.

1. Introduction

The proliferation of digital visual media has historically been accompanied by the desire to manipulate it; however, traditional editing required extensive domain expertise, specialized software, and significant time investments. In recent years, the advent of deep learning has drastically lowered the barrier to entry for digital forgery, giving rise to artificial intelligence-synthesized content commonly referred to as "Deepfakes" (a portmanteau of "deep learning" and "fake"). Deepfakes enable the seamless creation or manipulation of facial appearances, attributes, and expressions with a high degree of photorealism. While the democratization of digital human synthesis harbors positive applications in virtual reality, accessibility, education, and the film industry, the malicious exploitation of this technology has largely overshadowed its benign uses. Bad actors have weaponized deepfakes to produce non-consensual pornography, perpetrate financial fraud (e.g., a voice deepfake used to scam a CEO out of \$243,000), and fabricate political disinformation. Notable instances, such as the AI-generated images of a fabricated Pentagon explosion and the controversial video of Gabon's president which catalyzed a failed military coup, underscore the profound fragility of digital trust and the severe real-world implications of synthetic media.

The technical taxonomy of deepfakes can generally be stratified into four distinct categories based on their functional manipulation level: entire face synthesis, attribute manipulation, identity swap, and expression swap (also known as facial reenactment). Entire face synthesis relies on mapping random distributions to human face distributions to create non-existent identities. Identity swapping replaces the hard biometrics of a source individual with a target, whereas expression swapping and attribute manipulation tamper with soft biometrics, altering facial expressions or specific features like hair color and age without changing the underlying identity. In contrast to manual, non-AI modifications (termed "cheapfakes"), these deep learning-driven manipulations rely on sophisticated mathematical models to hallucinate photorealistic pixels, necessitating equally sophisticated, automated countermeasures capable of analyzing multimedia at scale.

The technological evolution of these generators has progressed at an astonishing pace. Initial deepfake techniques primarily utilized paired autoencoder-decoder structures, wherein a shared encoder captured latent facial features (such as lighting and pose) and subject-specific decoders reconstructed the swapped identities. Subsequently, Generative Adversarial Networks (GANs) established a new paradigm in machine learning research by pitting a generator against a discriminator in a minimax game, forcing the generation of highly convincing samples. Architectures such as ProGAN, StyleGAN, and BigGAN introduced innovations like progressive growing and style-based generator architectures, allowing for unsupervised separation of high-level attributes and the synthesis of megapixel-level, high-fidelity natural images. Despite their visual perfection, researchers discovered that the upsampling operations inherent to GAN decoders (e.g., transposed convolutions) leave behind systematic flaws and structural artifacts. These artifacts manifest as unnatural checkerboard patterns and distinct spectral peaks in the frequency domain, serving as a forensic fingerprint for detection.

However, the forensic landscape is currently undergoing a massive disruption due to a paradigm shift from GANs to Diffusion Models. Models such as DALL-E 2, Stable Diffusion, and GLIDE have achieved unprecedented photorealism and semantic control via text-based prompting. Diffusion models generate images through a fundamentally different mechanism—iteratively removing noise from a signal rather than mapping a low-dimensional latent vector to a high-dimensional space via upsampling. Consequently, contemporary deepfake detectors, which are predominantly trained to identify the specific high-frequency upsampling artifacts of GANs, suffer catastrophic performance degradation when tasked with identifying images generated by unseen diffusion architectures. While diffusion models still leave distinct traces in the Fourier spectrum, their artifacts do not share the same grid-like frequency layout as GANs, highlighting a critical vulnerability in existing forensic methodologies.

This continuous arms race between media generation and detection has forged a highly competitive "battleground" in computer vision. Early detection paradigms modeled the problem as a vanilla binary classification task, feeding images into standard Convolutional Neural Networks (CNNs) like Xception or ResNet. While these models achieved near-perfect accuracy on their specific training distributions, they exhibited severe overfitting to manipulation-specific artifacts, failing to generalize across different generative families or when subjected to real-world degradations like JPEG compression and aggressive resizing. To counter this, modern forensic research has pivoted towards learning more intrinsic and universally applicable features. For instance, the "Face X-ray" approach shifts the focus from finding generation artifacts to detecting the subtle blending boundaries universally introduced when a synthesized face is spliced back into a background video. Other state-of-the-art frameworks, such as F^3 -Net, leverage frequency-aware decomposition and local frequency statistics to expose subtle manipulation patterns that remain resilient even under heavy video compression. Furthermore, researchers are exploring universal feature spaces, such as those extracted by vision-language models (e.g., CLIP), combined with nearest-neighbor searches to detect fakes across completely unseen generative model families without explicit binary training.

In light of the escalating sophistication of generative models and the urgent necessity for robust forensic countermeasures, this paper provides a comprehensive, end-to-end review of the deepfake threat landscape. The remainder of this paper is structured as follows: Section 2 details the evolution of AI-generated media and deepfake

creation techniques, exploring the architectural mechanics from early Autoencoders to modern Diffusion models. Section 3 presents a detailed literature review of seminal deepfake detection frameworks. Section 4 provides a technical taxonomy of detection methodologies, categorized into spatial-domain, frequency-domain, and biological/temporal signals. Section 5 evaluates the progression of benchmark datasets and the quantitative metrics used to assess forensic models. Section 6 critically analyzes current challenges, focusing on cross-model generalization, image degradation, and evasion techniques employed by attackers. Finally, Section 7 explores future horizons in multi-modal detection, and Section 8 concludes the paper with a summary of findings.

2. Evolution of AI-Generated Media and Deepfake Creation

The generation of synthetic visual media has evolved rapidly from labor-intensive manual editing and traditional computer graphics to fully automated, deep learning-driven pipelines capable of unprecedented photorealism. In the context of facial manipulation, these technologies are fundamentally categorized into four distinct groups: entire face synthesis, identity swap (commonly referred to as face-swapping or deepfakes), attribute manipulation, and expression swap (facial reenactment). The architectural progression of these generation methods outlines a continuous trajectory of increasing complexity, shifting from early graphics-based composites and autoencoders to Generative Adversarial Networks (GANs) and, most recently, Text-to-Image Diffusion Models.

2.1 Early Computer Graphics and Autoencoder-Based Deepfakes

Before deep neural networks dominated the forensic landscape, highly convincing facial manipulations were predominantly achieved through advanced computer graphics techniques, notably FaceSwap and Face2Face. FaceSwap operates as a graphics-based approach to transfer a face region from a source video to a target. It extracts the face region utilizing sparse detected facial landmarks, fits a 3D template model using blendshapes, and back-projects this model to the target image by minimizing the difference between the projected shape and localized landmarks. The final rendered model is blended seamlessly into the background using color correction. Conversely, Face2Face focuses on facial reenactment, transferring the expressions of a source actor to a target video while maintaining the target's underlying identity. The Face2Face system automatically selects keyframes to obtain a temporary 3D face identity, tracks the expression, rigid pose, and lighting parameters per frame, and generates the output by transferring 76 source Blendshape coefficients to the target video.

The phenomenon that popularized the term "Deepfake" was built upon a relatively simple deep autoencoder architecture designed specifically for identity swapping. The fundamental pipeline begins with face detection, landmark extraction, and alignment to a standard configuration. The core generation mechanism relies on the parallel training of two autoencoders. The architecture utilizes a single, shared encoder E that compresses input images (typically at a resolution of $64 \times 64 \times 3$) into a reduced latent space of 1024 variables. This shared encoder is forced to learn general contextual information such as illumination, facial positioning, and expression. However, the architecture employs two separate, dedicated decoders (D_A and D_B) tasked with reconstructing the constant, identity-specific morphological details of the respective subjects.

During the inference or swapping phase, the target face is passed through the shared encoder, but the resulting latent code is decoded by the source identity's dedicated decoder. This effectively synthesizes the source individual's morphological features while retaining the target's original expression and lighting. The final step of this pipeline warps the synthesized face back to the original target configuration, masks the boundaries using facial landmarks, and seamlessly merges the altered face into the original video frame using Poisson image blending.

2.2 The Generative Adversarial Network (GAN) Era

The introduction of Generative Adversarial Networks (GANs) marked a paradigm shift in media synthesis, enabling the generation of entire non-existent faces and the complex manipulation of attributes. A conventional GAN architecture comprises two competing neural networks: a generator G and a discriminator D . The generator aims to produce realistic images $G(z)$ from noise signals z to mimic the real data distribution $p_{data}(x)$, while the discriminator attempts to accurately distinguish between genuine images x and those fabricated by the generator. The entire system is optimized through backpropagation in a minimax game, described by the value function: $\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))]$.

Despite early successes, standard GANs struggled with training instability and were largely limited to generating low-resolution images. To overcome these limitations, Progressive Growing of GANs (ProGAN) was introduced. ProGAN progressively grows both the generator and the discriminator during training, starting from an extremely low spatial resolution (e.g., 4×4 pixels) and incrementally adding layers to increase the resolution up to 1024×1024 . This progressive structural expansion significantly speeds up and stabilizes the training process, allowing for the reliable synthesis of high-resolution facial images.

Building upon the progressive growing foundation, StyleGAN revolutionized facial synthesis by entirely altering the generator's input structure. Instead of feeding a latent code directly into the input layer, StyleGAN utilizes an 8-layer mapping network f to transform a latent code z into an intermediate latent space w . This intermediate representation is specialized into style vectors $y = (y_s, y_b)$ via affine transformations, which are then injected into the synthesis network using adaptive instance normalization (AdaIN) operations. The AdaIN operation normalizes each feature map separately and scales and shifts it according to the provided style. Furthermore, StyleGAN injects random Gaussian noise directly into the generator layers to introduce high-frequency stochastic variations, such as freckles and hair placement. This architecture leads to the automatic, unsupervised separation of high-level attributes (like pose and identity) and enables intuitive, scale-specific control of the image synthesis process.

Other notable GAN architectures include BigGAN, which advanced unconditional image generation by employing a monolithic, class-conditional structure. BigGAN is trained on exceptionally large batch sizes and integrates self-attention layers alongside orthogonal regularization to enforce fidelity, utilizing a "truncation trick" to control the trade-off between image variety and visual quality. For attribute manipulation, conditional frameworks like StarGAN allow for multi-domain image-to-image translation using a single generator. StarGAN leverages domain-specific style codes to manipulate targeted facial attributes, such as hair color, age, and gender, without requiring separate models for every attribute pair.

2.3 The Rise of Text-to-Image Diffusion Models

Most recently, the generative landscape has experienced a monumental disruption with the emergence of Diffusion Models (DMs). Architectures such as DALL-E 2, Stable Diffusion, and GLIDE have gained the spotlight for providing an impressive level of photorealism and enabling the creation of complex visual content guided entirely by textual descriptions (prompts) or pilot sketches.

Unlike GANs, which map a low-dimensional latent vector to a high-dimensional pixel space through a series of upsampling convolutions, diffusion models learn the data distribution by iteratively removing noise. The generative process begins with a signal of pure Gaussian noise, and a neural network (often a U-Net architecture) is trained to progressively reverse a defined forward diffusion (noising) process, step-by-step recovering a structured, photorealistic image.

This fundamental architectural divergence carries significant forensic implications. Traditional CNN-based generators and GANs rely heavily on upsampling operations (e.g., transposed convolutions, nearest-neighbor, or bilinear upsampling), which consistently leave behind systematic flaws and structural artifacts. These upsampling mechanisms produce periodic, grid-like checkerboard patterns that manifest as clear spectral peaks in the frequency domain. Because diffusion models bypass this traditional upsampling paradigm and lack explicit 3D modeling of objects and surfaces, they do not exhibit the same grid-like frequency artifacts. Instead, diffusion models generally exhibit a systematic mismatch towards higher frequencies when analyzed in the Fourier spectrum. Consequently, the distinct generative pipeline of Diffusion Models has rendered many existing, GAN-trained forensic classifiers ineffective, necessitating the development of novel detection paradigms capable of adapting to this new frontier of synthetic media.

3. Literature Review on Deepfake Detection

The evolution of deepfake detection has been defined by a continuous technological arms race between advanced media generation techniques and forensic countermeasures. As generative models transitioned from basic autoencoders and early graphics-based methods to sophisticated Generative Adversarial Networks (GANs) and modern Text-to-Image Diffusion Models, the detection paradigm underwent a profound transformation. Early detection models relied heavily on spatial-domain binary classification, struggling with generalization across unseen architectures and video compression. To map this trajectory, this section provides an in-depth literature review of five seminal research contributions that trace the evolution of deepfake detection: **FaceForensics++**, **Face X-ray**, **Thinking in Frequency (F^3 -Net)**, **Universal CNN-Artifact Detection**, and the **Detection of Synthetic Images Generated by Diffusion Models**.

3.1 Establishing the Supervised Spatial Baseline: FaceForensics++

Methodology & Framework: The foundation of modern supervised facial forgery detection was significantly standardized by Rössler et al. through the introduction of **FaceForensics++**. Prior to this work, forensic datasets were severely limited in scale, often containing fewer than a thousand videos with minimal variations in subject identity and lighting. FaceForensics++ established a large-scale dataset comprising 1,000 pristine YouTube video sequences manipulated using four prominent state-of-the-art face manipulation methods: two computer graphics-based approaches (**Face2Face** and **FaceSwap**) and two learning-based approaches (**DeepFakes** and **NeuralTextures**).

To evaluate data-driven forgery detectors under realistic operational conditions, the authors instituted a rigorous evaluation protocol that incorporated multiple video compression levels: RAW (lossless), High Quality (HQ, constant rate quantization parameter $QP = 23$), and Low Quality (LQ, $QP = 40$), effectively simulating social media distribution laundering. For automated detection, the authors evaluated hand-crafted steganalysis features alongside deep learning classifiers, demonstrating that a domain-specific pipeline—combining robust face tracking to crop facial regions (enlarged by a factor of 1.3) with a pre-trained **XceptionNet** architecture—achieved unprecedented classification accuracy.

Strengths & Limitations:

- *Strengths:* FaceForensics++ provided an unprecedented scale of data (over 1.8 million manipulated images), enabling data-driven models to learn robust spatial features in a supervised fashion. The XceptionNet-based detector drastically outperformed human observers (who averaged roughly 58% to 68% accuracy across qualities) and proved highly effective in in-dataset scenarios.
- *Limitations:* Despite its high accuracy, FaceForensics++ revealed a critical vulnerability inherent to supervised spatial classifiers: catastrophic overfitting. When detectors are trained on specific manipulation

artifacts (e.g., DeepFakes or Face2Face), they struggle significantly when tested on unseen manipulation techniques or heavily compressed media. Furthermore, relying entirely on spatial pixel patterns leaves these models vulnerable to compression laundering and adversarial noise.

3.2 Bridging the Generalization Gap via Universal CNN Fingerprints

Methodology & Framework: Recognizing that technique-specific detectors quickly become obsolete as new generation methods emerge, Sheng-Yu Wang et al. investigated whether a classifier trained on a single CNN generator could exhibit universal generalization across entirely different architectures, datasets, and tasks. Their work, **CNN-Generated Images Are Surprisingly Easy to Spot... For Now**, introduced a large-scale dataset (ForenSynths) spanning 11 diverse CNN-based image generators, including unconditional GANs (ProGAN, StyleGAN, BigGAN), conditional GANs (CycleGAN, StarGAN, GauGAN), perceptual loss models (CRN, IMLE), low-level vision tasks (SITD, Super-Resolution SAN), and deepfakes.

The core methodology challenged the prevailing consensus that forensic classifiers fail to transfer across models. The authors trained a standard ResNet-50 binary classifier exclusively on images generated by a single unconditional model (**ProGAN**) using 20 distinct LSUN object categories. Crucially, they discovered that **aggressive data augmentation**—specifically combining Gaussian blurring and JPEG compression simulations during training—was the decisive factor enabling cross-model generalization, even when the test images themselves were uncompressed.

Strengths & Limitations:

- *Strengths:* This work demonstrated that CNN-generated images share common systematic flaws (such as upsampling and checkerboard artifacts) that allow a classifier trained on one architecture (ProGAN) to generalize surprisingly well out-of-the-box to unseen architectures, including the newly released StyleGAN2. It shifted the paradigm from single-model overfitting toward universal CNN artifact detection.
- *Limitations:* While the classifier generalized across various GAN architectures, its performance degraded when evaluated on models operating outside standard upsampling paradigms (such as certain super-resolution or heavily post-processed deepfake face-swapping pipelines). Furthermore, it remained fundamentally rooted in identifying spatial-frequency regularities native to CNN upsampling.

3.3 Moving Beyond Manipulation Artifacts to Blending Boundaries: Face X-ray

Methodology & Framework: To resolve the fundamental limitation of supervised classifiers—their susceptibility to overfitting to specific generator artifacts—Lingzhi Li et al. proposed a paradigm-shifting representation called **Face X-ray**. The authors observed that almost all facial manipulation methods (DeepFakes, Face2Face, FaceSwap, NeuralTextures) share a universal physical pipeline stage: **blending a synthesized foreground face into an existing background image**.

Instead of training a model to recognize the specific textural artifacts of a generator, Face X-ray trains a fully convolutional network (leveraging HRNet as a backbone) to predict a greyscale "X-ray" image that reveals whether an input image can be decomposed into two distinct sources. The brilliant innovation of Face X-ray lies in its training data generation: the model requires **zero fake images generated by state-of-the-art manipulation algorithms** during

training. Instead, the training set is constructed dynamically using only real images from FaceForensics++. By taking a real background image (I_B), searching for a visually similar real foreground face (I_F) using facial landmarks, applying random shape deformation and Gaussian blur to generate a soft blending mask (M), and performing alpha blending with color correction, the framework generates millions of realistic "blended" training samples alongside exact ground-truth X-ray masks.

Strengths & Limitations:

- *Strengths:* Face X-ray introduced exceptional generalization capabilities. Because it detects the universal physical discrepancy across blending boundaries (arising from hardware/software acquisition inconsistencies between two images) rather than specific generator fingerprints, it maintained high accuracy when tested on completely unseen forgery techniques where traditional classifiers experienced catastrophic performance drops.
- *Limitations:* Face X-ray assumes the explicit presence of a blending boundary. Consequently, if a generation technique evolves to synthesize an entire image cohesively from scratch (such as unconditional GANs or fully integrated diffusion models where no foreign face is spliced into a background), the blending boundary disappears, rendering Face X-ray ineffective.

3.4 Capturing Heavy Compression Resilience via Frequency Analysis (F^3 -Net)

Methodology & Framework: While spatial and boundary-based detectors perform well on high-quality media, heavy video compression (such as H.264 compression in social media pipelines) severely erodes high-frequency spatial details and blurring blending boundaries. To address this, Yuyang Qian et al. proposed **Frequency in Face Forgery Network (F^3 -Net)**, establishing that mining forgery patterns with frequency awareness is vital for handling compressed media.

F^3 -Net is built upon two complementary frequency-aware forgery clues that are fully compatible with Convolutional Neural Networks:

1. **Frequency-aware Image Decomposition (FAD):** Rather than using hand-crafted filter banks, FAD adaptively partitions the input image into frequency bands (low, middle, and high) using a set of binary base filters combined with learnable weights applied in the Discrete Cosine Transform (DCT) domain. These components are inversely transformed back to the spatial domain and fed into a CNN stream, allowing the model to capture subtle high-frequency artifacts resilient to compression.
2. **Local Frequency Statistics (LFS):** To capture statistical discrepancies between real and fake faces, LFS applies a Sliding Window DCT (SWDCT) locally and counts mean frequency responses across learnable frequency bands. These statistics form a multi-channel spatial map that mirrors natural image layouts.

To fuse these streams, the authors introduced **MixBlock**, a cross-attention module that enables bidirectional message passing and feature interaction between the FAD and LFS branches at mid and high network levels.

Strengths & Limitations:

- *Strengths:* F^3 -Net demonstrated state-of-the-art performance across all compression levels in FaceForensics++, exhibiting a massive performance lead (over 3.5% improvement in accuracy) specifically on low-quality (heavy compression) video tasks where standard RGB spatial networks completely failed. It proved that frequency-domain analysis effectively preserves forgery traces despite severe spatial degradation.
- *Limitations:* While highly robust to compression, frequency decomposition models like F^3 -Net are computationally intensive and can be sensitive to aggressive spatial downscaling or cropping that distorts localized DCT coefficient distributions.

3.5 Confronting the Next-Generation Threat: Detection of Diffusion Models

Methodology & Framework: As Generative Adversarial Networks began to be eclipsed by cutting-edge Text-to-Image Diffusion Models (e.g., Stable Diffusion, DALL-E 2, GLIDE, ADM), a new forensic crisis emerged. Riccardo Corvi et al., in their seminal work **On the Detection of Synthetic Images Generated by Diffusion Models**, investigated whether current state-of-the-art detectors—historically developed for and trained on GAN images—were suitable for identifying diffusion-generated media.

The authors conducted an extensive empirical audit by generating large test suites across diverse GANs, VQ-GANs, and Diffusion models using COCO prompts. By extracting noise residuals using a denoising filter and transforming them into the Fourier domain (following PRNU and GAN fingerprint extraction pipelines), they performed a spectral analysis across architectures. Their findings revealed a critical dichotomy: while older diffusion models (like GLIDE, Latent Diffusion, and Stable Diffusion) exhibited strong periodic spectral peaks similar to GAN fingerprints, advanced architectures like ADM and DALL-E 2 displayed extremely weak or absent spectral signatures. Consequently, classifiers trained exclusively on ProGAN or Latent Diffusion suffered severe generalization failures when tested across heterogeneous generative families, especially under social media laundering (resizing and JPEG compression).

Strengths & Limitations:

- *Strengths:* This work exposed the severe limitations of existing GAN-centric detectors in the face of the diffusion model revolution. It provided a rigorous forensic evaluation protocol (incorporating randomized cropping, scaling, and JPEG quality factors) that exposed the fragility of fixed-threshold detectors.
- *Limitations:* The study underscored that no single universal detector exists for diffusion models; models like DALL-E 2 generate images with minimal high-frequency artifacts or spectral grid patterns, rendering traditional fingerprint-based and frequency-based detection extremely challenging and calling for entirely new detection paradigms.

3.6 Synthesis and Evolutionary Narrative

The trajectory of deepfake and synthetic image detection reveals a clear paradigm shift driven by the continuous adaptation of generative algorithms:

1. **The Spatial-Overfitting Era (FaceForensics++):** Detection began as a supervised spatial classification problem. Models learned to identify pixel-level anomalies of specific generation techniques, achieving high accuracy in controlled environments but failing catastrophically on unseen manipulations.
2. **The Universal Artifact Era (Universal CNN-Artifact Detection):** Researchers discovered that common CNN upsampling mechanics leave universal structural and checkerboard traces across different GAN architectures. By incorporating aggressive data augmentations (blurring and JPEG compression), classifiers learned to generalize across multiple unseen CNN models.
3. **The Physical Boundary Era (Face X-ray):** Recognizing that generator architectures constantly change, forensic science shifted from analyzing internal generation textures to exploiting a universal physical constraint: the splicing and blending boundary. By training on pure real-image compositions, models achieved high generalization to completely unseen forgery techniques.
4. **The Frequency Resilience Era (F^3 -Net):** To counter real-world degradation such as heavy social media video compression, detection moved into the frequency domain. Utilizing adaptive DCT decomposition and local frequency statistics, models successfully exposed subtle manipulation traces hidden from RGB spatial views.
5. **The Diffusion Crisis (Detection of Diffusion Models):** The recent rise of diffusion models—which generate images via iterative noise-refinement rather than upsampling convolutions—has disrupted traditional frequency and spatial fingerprints. Advanced models like DALL-E 2 produce media devoid of traditional GAN grids, forcing the forensic community to look beyond standard deep neural network classifiers toward zero-shot, multi-modal, and non-parametric feature spaces.

4. Taxonomy of Detection Methodologies

The rapid evolution of artificial intelligence-synthesized media has catalyzed the development of an extensive array of deepfake detection methodologies. In order to counter the sophisticated forgery mechanisms employed by modern generative models, researchers have proposed diverse forensic approaches that exploit different categories of artifacts. This section presents a comprehensive taxonomy of deepfake detection techniques, systematically divided into three primary sub-domains: Spatial-Domain Detection, Frequency-Domain Detection, and Biological/Temporal Signals.

4.1 Spatial-Domain Detection

Early forensic techniques for deepfake detection predominantly operated within the spatial domain, treating the problem as a standard binary classification task. These approaches leverage pixel-level artifacts, statistical image properties, and deep convolutional neural networks (CNNs) to discriminate between genuine and forged media.

At the foundational pixel level, researchers have adapted traditional image forensics techniques to identify AI-synthesized content. For example, the extraction of Photo Response Non-Uniformity (PRNU) patterns—a noise pattern uniquely caused by the manufacturing imperfections of camera light sensors—has been heavily utilized. Because the

face-swapping process alters local PRNU patterns within the facial region, normalized cross-correlation scores between different image blocks can accurately identify tampered regions. Furthermore, pixel co-occurrence matrices, saturation cues, and differences in chrominance components have been fed into classifiers like Support Vector Machines (SVM) to detect anomalies in generated pixel distributions.

To automate feature extraction, various CNN-based classifiers have been employed. Architectures such as ResNet, XceptionNet, and EfficientNet, pre-trained on large-scale datasets like ImageNet, are routinely fine-tuned to classify images as real or fake. However, traditional deep CNNs frequently suffer from overfitting to manipulation-specific artifacts, precipitating significant performance degradation when confronting unseen generative techniques or compressed media. To counteract this, architectures focusing on mesoscopic properties have been proposed. MesoNet, comprising variants like Meso-4 and MesoInception-4, utilizes a minimal number of convolutional layers to constrain the network's focus to mesoscopic (intermediate) image properties, mitigating the effects of extreme video compression. Additionally, Capsule Networks, coupled with dynamic routing algorithms, have been introduced to explicitly model the hierarchical pose relationships between object parts, offering superior detection performance over traditional CNNs that lose spatial hierarchies due to pooling operations.

A paradigm shift within spatial-domain detection abandons the search for generative artifacts in favor of detecting the mechanical splicing traces common to most face manipulation pipelines. Face X-ray, a seminal representation, leverages the observation that almost all deepfake algorithms must blend a synthesized face into a target background, invariably leaving intrinsic image discrepancies across the blending boundary. A Face X-ray is a calculated grayscale image that reveals these blending boundaries, operating independently of the specific generative model used. Mathematically, for a given image I with a manipulation mask M , the Face X-ray B is calculated as $B_{i,j} = 4 \cdot M_{i,j} \cdot (1 - M_{i,j})$, ensuring the pixel values strictly peak at the boundary. Because Face X-ray can be trained using a vast corpus of pseudo-fakes composited entirely from pristine images, it exhibits unprecedented generalization across unseen deepfake methodologies. Expanding upon this concept, Self-Blended Images (SBIs) further enhance generalization by generating pseudo source and target images from a *single* pristine image. SBIs actively reproduce forgery artifacts such as blending boundaries, color mismatch, and frequency inconsistencies via localized transformations, forcing the detection network to learn robust, generic representations rather than overfitting to specific GAN artifacts.

4.2 Frequency-Domain Detection

Despite the success of spatial-domain methodologies, severe media compression and aggressive down-sampling on social platforms often obliterate pixel-level artifacts, rendering spatial CNNs highly vulnerable. Consequently, contemporary forensics has pivoted toward the frequency domain, where deepfake artifacts manifest as pronounced spectral anomalies that resist standard image degradation.

The fundamental efficacy of frequency-domain detection stems from the upsampling operations (e.g., transposed convolutions, nearest-neighbor interpolations) ubiquitously employed in generative networks mapping low-dimensional latent spaces to high-dimensional pixel matrices. These upsampling mechanisms violate classical sampling theorems, resulting in distinct, grid-like periodic checkerboard patterns. When transformed via the 2D Discrete Cosine Transform (DCT), natural images generally exhibit a rapid concentration of energy in low-frequency bands. In stark contrast, images synthesized by architectures like StyleGAN and ProGAN display severe, persistent artifacts and unnaturally high coefficients in high-frequency spectral regions. These spectral disparities are so distinct that simple linear ridge regression models trained exclusively on DCT coefficients can perfectly separate specific GAN-generated images from pristine data.

To integrate these potent frequency clues directly into deep learning architectures, the Frequency in Face Forgery Network (F^3 -Net) was proposed. The F^3 -Net framework identifies two complementary frequency-aware clues: Frequency-aware Decomposition (FAD) and Local Frequency Statistics (LFS). Because conventional frequency representations like DCT lack the local consistency and shift-invariance required by standard CNNs, FAD dynamically partitions the input image's frequency domain using a set of learnable frequency band filters. Specifically, FAD utilizes base filters combined with learnable filters applied via element-wise multiplication in the DCT domain, after which the Inverse Discrete Cosine Transform (IDCT) reconstructs frequency-aware spatial image components. This isolates subtle manipulation patterns hiding in high-frequency bands.

Concurrently, the LFS module extracts localized statistical discrepancies by applying a Sliding Window Discrete Cosine Transform (SWDCT). For each densely sampled local spatial patch, LFS calculates the mean frequency responses across different frequency bands and reassembles these statistics into a multi-channel spatial map. This enables the network to detect unnatural local frequency distributions while preserving spatial layouts. Finally, F^3 -Net seamlessly combines the FAD and LFS streams using a cross-attention fusion mechanism termed MixBlock, which calculates attention weights between the two feature streams to encourage rich feature interaction. This fusion of structural frequency isolation and local statistical analysis allows F^3 -Net to achieve state-of-the-art performance on highly compressed, low-quality deepfake video datasets.

4.3 Biological and Temporal Signals

While deep learning models can hallucinate photorealistic single frames, they persistently struggle to synthesize anatomically correct biological signals or maintain temporal coherence across video sequences. Therefore, exploiting biological markers and temporal dynamics has emerged as a robust countermeasure against deepfake videos.

One of the earliest physiological markers exploited for deepfake detection was eye-blinking. A healthy adult typically blinks every 2 to 10 seconds; however, early deepfake algorithms generated faces with significantly diminished blinking frequencies. This deficiency occurred because the algorithms were trained on massive repositories of internet images, which predominantly feature individuals with open eyes. To detect this anomaly, researchers utilized Long-term Recurrent Convolutional Networks (LRCN), composed of CNN feature extractors paired with Long Short-Term Memory (LSTM) cells, to learn sequence dynamics and predict the probabilities of eye open and close states. Other biological inconsistencies leveraged by detectors include unnatural head poses—derived by comparing the 3D head orientation estimated from full facial landmarks against the orientation of the spliced central face region—as well as the absence of accurate reflections in the eyes and teeth.

More advanced physiological detection models rely on extracting implicit biological signals, such as heart rate or pulse, which cannot be accurately replicated by GANs. Approaches like DeepRhythm and FakeCatcher utilize remote photoplethysmography (rPPG) to analyze the subtle, periodic color changes in human skin caused by blood flow. By extracting motion-magnified spatial-temporal representations (MMSTR) or combining spatial and temporal coherence of these biological signals, detectors can accurately authenticate real videos taken by cameras against deepfake synthesis.

Beyond biological markers, deepfake generation is typically executed on a frame-by-frame basis, inherently introducing low-level temporal artifacts and inter-frame inconsistencies. To exploit these flaws, detection pipelines combine convolutional networks with recurrent neural networks. Frame-level features extracted by architectures like DenseNet or InceptionV3 are fed into LSTM or Gated Recurrent Unit (GRU) cells to construct temporal sequence descriptors that classify the entire sequence's authenticity. Alternatively, optical flow fields have been heavily adopted to gauge temporal

disruptions. Optical flow calculates the apparent motion of pixels between consecutive frames; because deepfakes splice synthesized faces into dynamic scenes, they generate unnatural vector fields around the lips, eyes, and blending boundaries. Feeding these optical flow discrepancies into networks like ResNet50 or PWC-Net provides a robust temporal detection mechanism.

Finally, sophisticated multi-modal methodologies exploit visual-audio inconsistencies, leveraging the dissonance between lip movements and spoken audio tracks. Lip-syncing algorithms often fail to align specific visemes (mouth shapes) with their corresponding spoken phonemes. For instance, pronouncing words beginning with the phonemes 'M', 'B', and 'P' requires complete mouth closure, a physical constraint that deepfake synthesis frequently fails to replicate accurately. By calculating a modality dissonance score through contrastive loss networks, detectors can precisely measure the dissonance between audio and visual modalities, effectively identifying artificially dubbed or lip-synced deepfakes even when visual artifacts are negligible.

5. Benchmark Datasets and Evaluation Metrics

The accelerated evolution of deep learning-based facial manipulation algorithms has necessitated the parallel development of large-scale, high-quality benchmark datasets. For deepfake detection models to be effective in real-world scenarios, they require massive and diverse corpora of manipulated media to learn robust forensic representations. Over the past few years, the research community has iteratively constructed multiple datasets, which can be systematically categorized into three distinct generations based on their scale, visual quality, diversity of manipulation methods, and ethical data sourcing. Furthermore, the metrics utilized to evaluate detection performance have evolved to better reflect the extreme class imbalances inherent to "in-the-wild" deepfake detection.

5.1 Evolution of Deepfake Datasets

5.1.1 First-Generation Datasets: Early Forgeries and Foundational Baselines The first generation of deepfake datasets emerged as immediate responses to the initial proliferation of face-swapping software, generally containing fewer than 1,000 videos and less than one million total frames. A primary example is the UADFV dataset, which consists of only 49 real YouTube videos and 49 corresponding fake videos generated using the FakeApp mobile application. The manipulations in UADFV are highly constrained, replacing the original subjects exclusively with the face of actor Nicolas Cage. Another prominent first-generation dataset is DeepfakeTIMIT (DF-TIMIT), which comprises 620 fake videos depicting 32 subjects derived from the VidTIMIT database. DF-TIMIT videos were generated using a GAN-based face-swapping algorithm (faceswap-GAN) and were split into two subsets based on visual quality: Low Quality (LQ) synthesizing faces at 64×64 pixels, and High Quality (HQ) synthesizing faces at 128×128 pixels.

While these first-generation datasets provided foundational benchmarks for early detection models, they suffered from significant limitations. The manipulated videos exhibited severe, easily discernible visual artifacts, including low-quality synthesized faces, stark color contrasts between the fake mask and the original skin, visible splicing boundaries, and erratic temporal flickering. Furthermore, the lack of subject diversity—often restricted to a handful of identities sourced without explicit consent from YouTube—significantly constrained the generalization capabilities of models trained on them, rendering them prone to severe overfitting.

5.1.2 Second-Generation Datasets: Improved Quality and Methodological Diversity To address the low visual quality and limited scope of early datasets, second-generation datasets expanded the scale to between 1,000 and 10,000 videos and incorporated more sophisticated synthesis techniques. FaceForensics++ (FF++) is a cornerstone of this generation, comprising 1,000 real YouTube videos manipulated using four distinct state-of-the-art methods: two computer graphics-based approaches (Face2Face and FaceSwap) and two learning-based deepfake approaches

(DeepFakes and NeuralTextures). To simulate the degradation introduced by social media platforms, FF++ provides videos at three different compression levels: RAW (original quality), HQ (H.264 compression with a constant rate quantization parameter of 23), and LQ (quantization parameter of 40).

Despite the advancements in FF++, researchers identified that the visual quality of its deepfakes still lagged behind the hyper-realistic videos circulating on the internet. Consequently, the Celeb-DF dataset was introduced, featuring 590 real videos and 5,639 high-quality deepfake videos (totaling over two million frames) sourced from public YouTube interviews of 59 diverse celebrities. Celeb-DF dramatically reduced common visual artifacts through a refined synthesis algorithm. The resolution of the synthesized faces was increased to 256×256 pixels, and color mismatch was mitigated through dynamic training data augmentation and specialized color transfer algorithms. Furthermore, Celeb-DF addressed visible mask boundaries by utilizing smoothness masks based on interpolated points on the cheeks and chin, while temporal flickering was suppressed by filtering facial landmarks using a Kalman smoothing algorithm. Another dataset in this generation is the Google DeepFake Detection Dataset (Google DFD), which utilized 28 paid actors to create 3,068 deepfake videos, introducing the crucial element of participant consent. However, second-generation datasets still lacked the massive identity diversity required to achieve universal forensic generalization.

5.1.3 Third-Generation Datasets: Massive Scale, Real-World Perturbations, and Ethical Sourcing The third generation of datasets represents a monumental leap in scale, diversity, and realism, containing tens of thousands to hundreds of thousands of videos. The Deepfake Detection Challenge (DFDC) Dataset is by far the most extensive and rigorously constructed face-swapping dataset available, containing 128,154 total videos sourced from 3,426 paid actors. Overcoming the ethical shortcomings of first- and second-generation datasets, the DFDC dataset ensures that all recorded subjects provided explicit consent to have their likenesses modified. To mimic the potential harm of deepfakes targeted at average individuals rather than celebrities, the DFDC source videos were captured in highly diverse, natural settings—both indoor and outdoor—under various real-world lighting conditions and extreme head poses.

The DFDC dataset incorporates multiple generation methodologies to prevent models from overfitting to a single generator's artifacts. These methods include the Deepfake Autoencoder (DFAE), custom morphable-mask models (MM/NN face swap), Neural Talking Heads (NTH), FSGAN, and StyleGAN. The DFDC dataset also rigorously evaluates detector robustness by applying a vast array of augmentations and adversarial distractors to approximately 70% of the videos. Augmenters include Gaussian blurring, framerate alteration, varying encoding quality, and random noise. Distractors include overlying text, geometric shapes, and social media facial filters (such as dog masks and flower crowns) to simulate adversarial real-world conditions.

Another prominent third-generation dataset is ForgeryNet, which stands as the largest benchmark in terms of structural diversity. ForgeryNet encompasses 2.9 million images and 221,247 videos, utilizing 15 distinct manipulation approaches spanning entire face synthesis, identity swap, attribute manipulation, and expression swap. Furthermore, ForgeryNet introduces 36 types of mix-perturbations applied across over 5,400 subjects, solidifying the transition towards evaluating deepfake detectors under highly complex, "in-the-wild" conditions.

5.2 Evaluation Metrics for Deepfake Detection

Evaluating the efficacy of deepfake detection models necessitates rigorous quantitative metrics. Because the fundamental nature of the detection task varies—from aiding manual forensic investigation to deploying automated filters on massive social media streams—different metrics are required to provide a holistic view of a model's performance.

5.2.1 Accuracy and Equal Error Rate (EER) Accuracy, defined as the proportion of correct predictions over total predictions, is the most fundamental metric used in early forensic evaluations. However, accuracy is highly threshold-dependent and can be severely misleading in scenarios with extreme class imbalances. To mitigate threshold dependency, the Equal Error Rate (EER) is frequently reported. EER represents the specific operating point on the detection curve where the False Acceptance Rate (the proportion of fake images mistakenly classified as real) is exactly equal to the False Rejection Rate (the proportion of real images incorrectly classified as fake). A lower EER indicates a superior detection model, as it minimizes the crossover of false predictions.

5.2.2 Area Under the Receiver Operating Characteristic Curve (AUC) To provide a more comprehensive assessment across all possible classification thresholds, the Area Under the Receiver Operating Characteristic Curve (AUC) is the most widely adopted metric in deepfake benchmarking. AUC is a threshold-less, ranking-based score that plots the True Positive Rate against the False Positive Rate. This metric is highly advantageous because it is insensitive to the base rate of real and fake images within the dataset. In state-of-the-art evaluations, such as those conducted on the Celeb-DF dataset, frame-level AUC scores are preferred over video-level aggregation. Frame-level AUC avoids the inconsistencies caused by different temporal aggregation strategies (e.g., averaging scores versus taking the maximum) and obviates the necessity of calibrating classification outputs across diverse datasets.

5.2.3 Weighted Precision While AUC and EER are excellent for academic benchmarking, they often fail to capture the operational reality of automated deepfake detection at scale. In realistic social media environments, the distribution of organic traffic is overwhelmingly skewed; the ratio of deepfaked videos to unaltered, genuine videos may be less than one in a million. Under such an extreme class imbalance, a model that achieves a high overall accuracy but possesses a moderate False Positive Rate will generate an insurmountable volume of false alarms, completely drowning out the true positives and destroying the model's practical utility.

To address this critical deployment challenge, the designers of the Deepfake Detection Challenge proposed the Weighted Precision (wP) metric. Weighted Precision artificially simulates a massive class imbalance by severely penalizing false positives. Assuming the ratio of unaltered to tampered videos in organic traffic differs from the test dataset by a large factor α , Weighted Precision is mathematically defined as: $wP = \frac{TP}{TP + \alpha FP}$. Unlike standard F_β scores that assign weights to false negatives, wP heavily weights the False Positives (FP) against the True Positives (TP). By setting a high scaling factor (e.g., $\alpha = 100$), this metric strictly evaluates whether a detection model can maintain high precision without triggering alert fatigue, making it the most indicative metric for how a detector will perform against a realistic, real-world distribution of multimedia traffic.

6. Current Challenges, Generalization, and Evasion Techniques

6.1 The Adversarial Battleground The rapid and continuous evolution of artificial intelligence-synthesized media has established a highly adversarial domain, widely recognized as a "battleground" between deepfake generation methods and deepfake detection methods. In this competitive ecosystem, the "offenders" or "adversaries" represent the generation algorithms attempting to create frictionless, hyper-realistic forgeries, while the "defenders" are the detection models striving to expose them. The development of defenders is continuously accelerated by the aggressive push from adversaries, and vice versa. When a novel generation method is developed, its creators naturally aim to penetrate and bypass the latest deepfake detectors; concurrently, the resulting generator will inevitably be analyzed and blocked by newer forensic detectors down the line. This unceasing, tightly-knit interactivity forces both parties to outcompete the other side, resulting in incremental but continuous scientific progress. However, the intrinsic nature of this battleground dictates that detection methods are inherently reactive and usually lagging behind generation methods. The current state-

of-the-art deepfake detectors remain somewhat myopic; they confidently tackle deepfakes generated by existing methods but are inevitably outcompeted by forgeries generated through future, unseen generative techniques.

6.2 The Generalization Crisis: Failing on Unseen Data One of the most critical challenges plaguing modern forensic architectures is the generalization crisis. Most existing deepfake detection systems model the problem as a supervised binary classification task, wherein a deep convolutional backbone extracts global features from a suspect image to feed into a real/fake classifier. When evaluated under controlled scenarios—meaning the detector is tested on the exact same manipulation methodologies it was trained on (e.g., DeepFakes, Face2Face, or FaceSwap)—these models perform exceptionally well, routinely achieving detection accuracies of around 98%.

However, because the differences between real and photorealistic fake images are profoundly subtle and localized, this vanilla binary classification approach is highly suboptimal. Deep neural network-based detectors tend to suffer from severe overfitting, fixating on manipulation-specific artifacts present only in their training data. Consequently, their effectiveness is strictly limited to the specific generative techniques they were optimized against. When these supervised binary classifiers are deployed against deepfakes synthesized by unknown or unseen manipulation methods, they experience a catastrophic performance drop, often reverting to near-chance accuracy. For instance, classifiers heavily optimized to detect the spatial artifacts of classical autoencoder-based deepfakes fail to identify the high-level semantic anomalies introduced by entirely different manipulation paradigms.

To combat this generalization failure, recent methodologies have attempted to shift the learning objective away from generator-specific artifacts. Approaches such as "Face X-ray" focus universally on the intrinsic image discrepancies that occur across the blending boundaries when any altered face is spliced back into a background image. Similarly, frameworks utilizing Self-Blended Images (SBIs) generate pseudo source and target images from single pristine images to reproduce common forgery artifacts like blending boundaries and statistical color inconsistencies, actively encouraging the classifier to learn generic, robust representations rather than overfitting to specific manipulation traces.

6.3 Vulnerability to Image Degradation and Post-Processing Compounding the generalization crisis is the extreme vulnerability of forensic detectors to common image degradation. In realistic, real-world deployment scenarios—such as videos uploaded to and distributed across social media networks—multimedia files undergo severe, automatic post-processing. These operations include aggressive algorithmic resizing, heavy H.264 or JPEG compression, varying bit-rates, and blurring.

These standard operations act as inadvertent laundering mechanisms that obliterate the high-frequency spatial and spectral artifacts upon which deepfake detectors heavily rely. Extensive evaluations conducted on large-scale datasets, such as FaceForensics++, demonstrate an undeniable correlation between video quality degradation and the rapid decline of detection capabilities. While automated deep neural networks achieve near-perfect classification accuracy on uncompressed, RAW input data, their performance drops precipitously when evaluating low-quality (LQ) compressed video footage. This degradation is even more pronounced for forensics methods relying on hand-crafted steganalysis features and shallow CNN architectures, which struggle immensely to cope with compression, sometimes falling below the accuracy threshold of random human observers.

To mitigate the catastrophic impact of image degradation, researchers have adopted two primary strategies. First, injecting aggressive data augmentation during the training phase—such as simulating heavy JPEG compression and random Gaussian blurring—forces the neural network to learn persistent features that survive severe degradation, simultaneously improving the model's cross-generator robustness. Second, moving beyond the spatial domain,

frameworks like F^3 -Net exploit the frequency domain, arguing that subtle manipulation artifacts and compression errors are better described through localized frequency statistics. Despite these interventions, the battle against aggressive, in-the-wild media degradation remains one of the most formidable hurdles for scalable deepfake detection.

6.4 Evasion Techniques: Adversarial Attacks and Shallow Reconstruction As deepfake detectors grow more sophisticated, malicious actors actively engineer dedicated evasion techniques to bypass forensic scrutiny. By mapping a real or fake face to a new representation, attackers attempt to hide the distinct fake traces from state-of-the-art detectors. Evasion methodologies broadly fall into three categories: adversarial perturbations, natural degradation spoofing, and frequency artifact removal.

The first type relies on adversarial machine learning attacks, wherein attackers inject mathematically optimized, visually imperceptible adversarial perturbations into the fake media. These perturbations maliciously manipulate the loss gradients of the detector, causing even the most robust, state-of-the-art deepfake classifiers to fail catastrophically under both white-box and black-box attack scenarios.

Because standard additive adversarial noise can sometimes be detected by secondary defense mechanisms, attackers have pioneered a second category: natural degradation-based adversarial attacks. These attacks leverage operations such as motion blur, light variations, and weather elements to generate realistic-blurred images that successfully mislead deep neural networks while appearing as organic environmental artifacts to human observers.

The third category focuses on neutralizing the specific forensic clues used by detectors, particularly in the frequency domain. Knowing that CNN-based upsampling leaves severe, distinct spectral peaks, attackers regard evasion as a general image generation process, employing advanced image filtering or dictionary-based shallow reconstruction to erase these frequency anomalies. Through sparse coding and shallow reconstruction, attackers can effectively smooth out the spectral fingerprints of a deepfake without compromising the visual photorealism of the spatial image, effectively blinding frequency-based detectors. Furthermore, evasion frameworks can explicitly incorporate spectral regularization objective terms into the generative adversarial network's loss function during the training phase, forcing the generator to perfectly reproduce the spectral distributions of natural training data and repairing the frequency flaws before the deepfake is even synthesized.

6.5 The Paradigm Shift: The Struggle Against Diffusion Models Perhaps the most pressing modern challenge in media forensics is the abrupt paradigm shift in generative architectures, moving away from Generative Adversarial Networks (GANs) toward Text-to-Image Diffusion Models. Over the past decade, the rapid progress in creating synthetic media was largely driven by powerful GAN architectures. Consequently, the vast majority of existing forensic algorithms, feature extractors, and benchmark datasets have been heavily optimized to detect GAN-specific artifacts.

GAN architectures generate imagery by mapping a low-dimensional latent space to a higher-dimensional pixel space. This generation relies on successive upsampling operations—such as transposed convolutions or bilinear upsampling—which are mathematically known to cause structural artifacts. These operations fail to reproduce natural spectral distributions, leaving behind severe, periodic, grid-like artifacts that are easily identifiable as bright peaks in the Discrete Cosine Transform (DCT) frequency space. Because these flaws are systemic across GAN architectures (e.g., StyleGAN, ProGAN, BigGAN), detectors trained on one GAN can often generalize to another by simply searching for these shared upsampling artifacts.

However, the recent advent of Diffusion Models (DMs)—such as DALL-E 2, Stable Diffusion, and GLIDE—has fundamentally disrupted this forensic reliance. Diffusion models do not utilize the same architectural mechanisms; they synthesize images through a gradual denoising process rather than single-pass latent upsampling. Consequently, the severe, grid-like upsampling artifacts and spectral peaks that characterize GANs are largely absent in media generated by diffusion models.

When state-of-the-art detectors developed for GAN-generated images are applied to these new synthetic images, their performance drops severely. Empirical analyses reveal that detectors trained exclusively on GAN images (such as ProGAN) work poorly when attempting to detect images generated by DALL-E 2 or the Ablated Diffusion Model (ADM). While diffusion models do leave their own, unique forensic traces in the frequency domain, these artifacts are much weaker and exhibit vastly different spectral patterns compared to GANs. Thus, achieving cross-model generalization remains the ultimate hurdle. The inability of current GAN-trained forensic models to reliably classify images from diffusion models exposes a profound vulnerability in the current forensic ecosystem, confirming that the continuous evolution of generative models requires an equally continuous, radical reinvention of deepfake detection strategies.

7. Future Horizons

As generative models continuously evolve, the forensic community must proactively anticipate future threats and pioneer next-generation detection paradigms. The current "battleground" dynamic reveals that deepfake detectors are inherently reactive, often struggling to maintain robustness against unseen generative architectures and natural image degradations. To maintain the integrity of digital media, future research must pivot toward holistic, adaptive, and predictive forensic frameworks.

7.1 Multi-Modal Detection Strategies The vast majority of current forensic algorithms operate exclusively in the visual domain, scrutinizing spatial anomalies or frequency artifacts within isolated frames. However, relying solely on visual cues is increasingly insufficient as image synthesis approaches visual perfection. Future horizons dictate a paradigm shift toward multi-modal detection, which incorporates supplementary media modalities such as voice and sound. By evaluating the interactions between these modalities—such as analyzing the synchronization of acoustic cues with visual mouth movements (lip-sync)—detectors can expose sophisticated deepfakes. Detecting these multi-modal fake traces, including standalone visual cues, standalone acoustic cues, and their complex interactions, will significantly boost overall detection performance by adhering to Liebig's barrel theory, where the weakest synthesized modality reveals the forgery.

7.2 Self-Supervised Learning and Vision Transformers Currently, supervised binary classification heavily relies on massive, manually curated datasets of real and fake media. This dependency inherently limits a model's ability to generalize to unknown attacks, as the classifier becomes excessively tethered to the artifacts of the specific generators present in its training distribution. To overcome this, self-supervised learning has emerged as a major research hot topic. Free from the constraints of explicit labeling, self-supervised learning enables continuous, life-long learning on an infinite and smoothly changing data stream. Furthermore, the integration of Vision Transformers (ViTs) presents a promising frontier. Transformers, which have already disrupted natural language processing, model complex, long-range dependencies across an image far more effectively than standard Convolutional Neural Networks (CNNs). Although a dedicated ViT for deepfake detection is still in its infancy, the unprecedentedly fast-growing pace and increasing ubiquity of ViTs in computer vision will eventually transition the forensic landscape into a predominantly ViT-based one.

7.3 Approaching the 'DeepFake Singularity' As the arms race between adversaries and defenders accelerates, we must confront the ultimate existential challenge of multimedia forensics: the "DeepFake Singularity". Humanity may eventually reach a stage where DeepFakes become so genuinely flawless that they are beyond both human and machine capability to distinguish from real media. If generative models can perfectly replicate anatomical physics, temporal coherence, and natural spectral distributions, reactive post-generation detection will be rendered obsolete. Preparing for this singularity requires the adoption of proactive, structural countermeasures. Future defense strategies must embed security directly into the media lifecycle, utilizing technologies such as digital watermarking, distributed ledger technologies (blockchain) for immutable digital provenance, and responsible disclosure mandates enforced via GAN fingerprinting.

8. Conclusion

The democratization of deep generative algorithms has fundamentally disrupted digital trust, making the creation of hyper-realistic, AI-synthesized media frictionless and accessible to the masses. This comprehensive survey has chronicled the evolutionary trajectory of deepfakes, transitioning from rudimentary autoencoder-based face swaps to highly sophisticated architectures powered by Generative Adversarial Networks (GANs) and, most recently, Text-to-Image Diffusion Models. As generation techniques have grown exponentially in their ability to hallucinate photorealistic pixels, the necessity for robust, automated forensic countermeasures has never been more urgent.

We have detailed the extensive taxonomy of detection methodologies developed to combat this threat. Early detectors relied on spatial-domain CNNs to identify pixel-level artifacts, achieving near-perfect accuracy on controlled, in-distribution datasets like FaceForensics++. However, recognizing the catastrophic generalization failure of these vanilla classifiers, researchers innovated deeply. Modern detection paradigms now target intrinsic manipulation markers—such as blending boundaries via Face X-ray and Self-Blended Images (SBIs)—or exploit severe, periodic upsampling artifacts hiding in the Fourier spectrum utilizing frameworks like F^3 -Net. Additionally, leveraging biological signals, such as rPPG and temporal audio-visual dissonance, has provided resilient defenses against sequence-based deepfakes.

Despite these advancements, the adversarial battleground remains fraught with immense challenges. Forensic models continue to struggle against aggressive real-world image degradations (such as H.264 compression and resizing), adversarial evasion techniques designed to erase spectral fingerprints, and the monumental paradigm shift toward Diffusion Models, which bypass the traditional upsampling artifacts that GAN-trained detectors rely upon. To navigate these challenges, the continuous curation of massive, ethically sourced, third-generation benchmark datasets like the DFDC and GenImage is imperative to evaluate detectors under rigorous, "in-the-wild" conditions. Ultimately, winning the war against visual disinformation will require continuous algorithmic innovation, multi-modal analysis, and proactive provenance frameworks to ensure the long-term integrity of digital media.

References

- [1] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 1831–1839.
- [2] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu, "Countering malicious deepfakes: Survey, battleground, and horizon," *arXiv preprint arXiv:2103.00218*, 2022.
- [3] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," *arXiv preprint arXiv:2211.00680*, 2022.

- [4] M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, Z. Tu, H. Hu, J. Hu, and Y. Wang, "GenImage: A million-scale benchmark for detecting AI-generated image," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [5] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 24480–24489.
- [6] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18720–18729.
- [7] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2185–2194.
- [8] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.
- [9] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8695–8704.
- [10] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 86–103.
- [11] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *International Conference on Machine Learning (ICML)*, 2020, pp. 3247–3258.
- [12] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5001–5010.
- [13] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge dataset," *arXiv preprint arXiv:2006.07397*, 2020.
- [14] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3207–3216.
- [15] T. T. Nguyen *et al.*, "Deep learning for deepfakes creation and detection: A survey," *arXiv preprint arXiv:1909.11573*, 2022.
- [16] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1–11.
- [17] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2307–2311.
- [18] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1–7.
- [19] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 2672–2680.
- [20] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. International Conference on Learning Representations (ICLR)*, 2014.
- [21] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.
- [22] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.
- [23] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. International Conference on Learning Representations (ICLR)*, 2018.

- [24] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.
- [25] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8789–8797.
- [26] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial attribute editing by only changing what you want," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5464–5478, 2019.
- [27] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1831–1839.
- [28] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8261–8265.
- [29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [30] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114.