Detecting Synthetic Reality: A Technical Report on AI-Generated Image Forensics

Part I: The Strategic Landscape of Synthetic Media Detection

The capacity to distinguish authentic digital media from synthetically generated content has become a paramount challenge in computer science and AI security. The rapid, democratized proliferation of powerful generative models necessitates the development of robust, scalable, and reliable detection systems. This report provides a comprehensive technical analysis of the state-of-the-art in AI-generated image detection, offering a detailed survey of foundational and advanced methodologies, practical tools and datasets, and the core challenges that define the research frontier. It is intended to serve as a foundational document for technical teams engaged in the development of synthetic media detection projects.

1.1 The Generative Revolution: From GANs to Large AI Models (LAIMs)

The field of synthetic media forensics is defined by the continuous evolution of the generative technologies it seeks to detect. Understanding this evolution is critical to developing detection strategies that remain effective over time.

The Genesis of High-Fidelity Forgery

The modern era of realistic image generation was catalyzed by the development of Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs).[1] GANs, in particular, introduced a novel training paradigm where a

generator network learns to produce realistic data to fool a *discriminator* network, which simultaneously learns to distinguish real data from the generator's creations.[2] This adversarial process proved remarkably effective at synthesizing high-quality multimedia, leading to an explosion of research and applications.[1] However, these earlier models, while powerful, often left behind subtle but systematic artifacts. These tell-tale "fingerprints," frequently stemming from the up-sampling components in their architecture, became the primary target for the first generation of detection methods.[3]

The Paradigm Shift to LAIMs

More recently, the landscape has been fundamentally reshaped by the emergence of Large AI Models (LAIMs), a category that includes Diffusion Models (DMs) and the Large Language Models (LLMs) that often guide them in text-to-image synthesis.[5] LAIMs are distinguished from their predecessors by several key characteristics: exceptionally high parameter counts (often in the billions), training on vast and diverse web-scale datasets, and a marked improvement in the quality, realism, and controllability of the generated content.[5] Diffusion models, which operate by iteratively de-noising a random signal into a coherent image, have demonstrated a superior ability to generate highly realistic and complex scenes, significantly closing the "realism gap" and making casual detection far more difficult.[5]

The advancement from GANs to LAIMs represents a significant escalation in the detection challenge. Many detection techniques and datasets that were effective against GAN-generated images are now considered outdated or insufficient for tackling the output of modern diffusion models.[8] This technological shift necessitates a corresponding evolution in detection strategy, moving away from methods that rely on the specific, known artifacts of older architectures.

Societal and Ethical Implications

The proliferation of these advanced generative tools is a dual-edged sword. While they offer immense benefits in fields like creative arts, education, and business process optimization [5], they also introduce profound societal risks. The ability to

create photorealistic images and videos on demand has been increasingly leveraged for malicious purposes, including propaganda, disinformation campaigns, and other forms of societal disruption.[1] The potential for misuse in sensitive contexts, such as generating fake evidence or influencing election cycles, has elevated the development of effective detection technologies from a purely academic pursuit to a critical component of global AI security and information integrity efforts.[5]

1.2 The Asymmetry of Perception: Human vs. Machine Detection

A key justification for the development of automated detection systems is the well-documented unreliability of human perception in distinguishing real from AI-generated imagery. While one might intuitively expect humans to possess a superior ability to spot "fakeness," large-scale empirical studies have consistently shown this is not the case.

Quantifying Human Inaccuracy

Research has demonstrated that human accuracy in this task is alarmingly low, often only slightly better than random chance. A significant study involving over 12,500 global participants who performed approximately 287,000 image evaluations found an overall success rate of just 62%.[8] Other studies have reported even lower performance, with average accuracies hovering between 49% and 61%.[11] This fallibility persists even when the synthetic images are of average quality and produced by generators that are now considered outdated.[8] The hyper-realistic quality of modern generated content is such that it can be indistinguishable to the human eye, even for experts.[7]

Content-Dependent Human Performance

Human performance is not uniform across all types of content. Studies indicate that participants are most accurate when identifying synthetic human portraits but

struggle significantly with more complex scenes like natural and urban landscapes.[8] This suggests that human detection relies heavily on semantic cues and familiarity with the subject matter. We are highly attuned to subtle errors in human faces, but our intuition is less reliable when assessing the physical plausibility of a landscape's lighting or a building's perspective. When attempting to identify fakes, humans tend to focus on specific objects or a "general impression," whereas for real images, the focus is broader.[13] This difference in strategy highlights a reliance on high-level semantics that generative models are increasingly adept at replicating.

The Imperative for Automation

The demonstrated inability of humans to reliably detect synthetic images, coupled with the escalating sophistication of generative models, creates a critical need for robust and scalable automated detection tools.[8] As malicious actors exploit these technologies to sow discord and spread misinformation, relying on human vigilance alone is an untenable strategy. Automated systems, capable of analyzing subtle, low-level statistical artifacts and physical inconsistencies that are invisible to the human eye, are essential for mitigating these risks and helping to restore trust in digital content.[5]

1.3 A Taxonomy of Forgery and Detection

To systematically address the problem of synthetic media, it is crucial to establish a clear taxonomy for both the types of forgeries being created and the methods used to detect them. This provides a structured framework for understanding the problem space and aligning project goals with appropriate technical solutions.

Categorizing Forgery Types

AI-driven image manipulation is not a monolithic category. It encompasses a spectrum of alterations, from the creation of entirely new scenes to subtle modifications of

existing content. Key forgery types include:

- **Full Image Synthesis:** Generating an entire image from scratch, typically guided by a text prompt or another image. This is the primary capability of models like Stable Diffusion and Midjourney.[14]
- **Deepfakes:** A sub-category of synthetic media that typically involves manipulating faces. This includes several distinct operations:
    - **Face Swapping:** Transferring the identity of a source person onto a target person in an image or video.[15]
    - **Face Reenactment:** Transferring the expressions and head movements of a source person to a target person.[15]
    - **Talking Face Generation:** Synthesizing a video of a person speaking from a single portrait and an audio track.[15]
    - **Facial Attribute Editing:** Modifying specific attributes of a face, such as age, hair color, or expression.[15]
- **Image Inpainting/Outpainting:** Generating or replacing only a specific portion of an image, which can be used to seamlessly add or remove objects.[16]

A Framework for Detection Methods

Recent comprehensive surveys have introduced a novel taxonomy for detection methodologies that provides a powerful lens through which to view the field.[5] This framework organizes methods along two primary axes: media modality and detection perspective.

- **Media Modality:** This axis categorizes detectors by the type of content they analyze:
    - Text
    - Images
    - Videos
    - Audio
    - Multimodal (e.g., analyzing video and audio streams concurrently)
- **Detection Perspective:** This axis categorizes detectors by their primary objective, moving beyond simple classification accuracy:
    - **Pure Detection:** This perspective includes methods whose main goal is to enhance raw detection performance, such as accuracy and inference speed.
    - **Beyond Detection:** This perspective encompasses methods that aim to add crucial attributes to the detector, recognizing that real-world deployment

requires more than just high accuracy on a test set. These attributes are:

- **Generalizability:** The ability of a detector to perform well on synthetic media generated by models it was not exposed to during training.
- **Robustness:** The resilience of a detector to common real-world image perturbations (e.g., compression, resizing, noise) and to deliberate adversarial attacks.
- **Interpretability (or Explainability):** The ability of a detector to provide a rationale for its decision, which is critical for applications in forensics, journalism, and legal contexts.

This "Beyond Detection" framework is particularly valuable for project planning, as it forces a consideration of the practical requirements for a successful deployment. A model that achieves 99% accuracy on a clean, known dataset but fails catastrophically when faced with a compressed image from a new generator is of little practical use. Therefore, a successful project must define its objectives and evaluate its performance against these broader attributes.

Part II: A Compendium of Detection Methodologies

The technical approaches for detecting AI-generated images are diverse, ranging from foundational deep learning techniques that learn from raw pixels to highly specialized methods that analyze the underlying physics and geometry of a scene. A state-of-the-art detection system may benefit from integrating multiple methodologies, as the "artifact" left by a generative model is not a single, static signal but a complex, evolving set of potential clues. The evolution of these artifacts, from low-level statistical traces common in older GANs to high-level semantic inconsistencies in images from modern LAIMs, dictates a corresponding evolution in detection strategy. An effective system should not rely on a single technique but rather employ a multi-pronged defense, combining analyses of different signals to create a detector that is more robust and harder to fool.

2.1 The Foundational Approach: Deep Learning Classifiers

The dominant paradigm in synthetic image detection, as in many computer vision tasks, is the use of deep learning models as binary classifiers. These models are trained on large datasets of real and synthetic images to automatically learn the distinguishing features between the two classes.

CNN-Based Architectures

Convolutional Neural Networks (CNNs) have long been the workhorse of image forensics. Architectures such as ResNet-50, XceptionNet, and DenseNet are frequently employed as backbones for detection models.[4] These networks excel at learning hierarchical feature representations directly from pixel data. For instance, a simple classification model built on a standard ResNet-50 architecture can be trained to achieve near-perfect accuracy in separating synthetic and real images

from a specific generator it was trained on.[7] However, this high performance is often brittle; the model may overfit to the unique artifacts of that particular generator, leading to poor performance on images from unseen models. This highlights a central challenge in the field: the generalization gap.

Vision Transformers (ViT) and Large Vision-Language Models (LVLMs)

More recently, the field has begun to adopt more advanced architectures that have shown superior performance in general computer vision tasks. These include Vision Transformers (ViT) and Large Vision-Language Models (LVLMs).

- **ViT-based Models:** Architectures like the Vision Transformer are being used to gather global and spatial features from images. The UnivFD detector, for example, uses features from a pre-trained CLIP ViT encoder combined with nearest-neighbor search to detect synthetic images from various generative architectures.[18] Another example,
  UnivCLIP, also utilizes the ViT architecture for its feature extraction capabilities.[19]
- **LVLMs for Detection:** The powerful representations learned by LVLMs, such as BLIP-2 and ViTGPT2, are being harnessed for synthetic image detection in several innovative ways.[20] Some approaches fine-tune these models to act as direct artifact classifiers.[21] Others leverage the multimodal nature of these models. The

DeFake method, for instance, operates by exploiting the subtle misalignments that can occur between an image's content and its generative prompt, as measured by the difference in CLIP's image and text encoder outputs.[18] Another LVLM-based approach,
LEGION, aims not only to detect but also to ground its decision in the image and provide textual explanations for the detected artifacts.[14]

Self-Supervised and Contrastive Learning

Novel training paradigms are also being explored. Self-supervised learning methods like SimCLR and Masked Autoencoders (MAE) are being applied to this problem space. Intriguingly, research has shown that training these models on purely synthetic images can, under certain conditions, lead to representations that match or even outperform those learned from an equivalent set of real images.[22] The key appears to be careful control over the generative process, such as tuning the classifier-free guidance scale in Stable Diffusion. This suggests that synthetic data, with its potential for controlled diversity and content, could become a valuable asset for training more robust detectors, not just a target for them.

2.2 Unmasking Artifacts in the Frequency Domain

A powerful and widely studied class of detection methods operates not in the spatial (pixel) domain but in the frequency domain. The core principle is that the mathematical operations inherent in the image generation process, particularly the up-sampling layers common to GANs, VAEs, and diffusion models, introduce systematic, high-frequency artifacts that are often invisible to humans but can be readily identified with signal processing techniques.[3]

Key Techniques

Several methods are used to transform images into the frequency domain to reveal

these artifacts:

- **Discrete Cosine Transform (DCT):** The DCT is a standard technique in image processing, most famously used in JPEG compression. Research has shown that a comprehensive frequency-domain analysis using DCT reveals severe and common artifacts across a wide range of GAN architectures.[24] Based on this, simple yet effective detectors like
  FreqDetect use the DCT of an image as the direct input to a classifier, such as logistic regression, to distinguish real from fake.[18]
- **Fast Fourier Transform (FFT):** The FFT is another fundamental tool for frequency analysis. Studies using FFT have confirmed that genuine and counterfeit images exhibit distinguishable spectral differences, particularly at higher frequencies.[25] Some approaches even propose using frequency-domain upsampling techniques, like Deep Fourier Up-Sampling, to enhance the signals used for detection.[25]
- **Wavelet Transform:** For scenarios involving low-quality or heavily compressed images where high-frequency details might be lost, the Haar wavelet transform offers a robust alternative. By decomposing the image into different frequency sub-bands, it's possible to isolate the mid-to-high frequency information. This information can be captured in a "residual map" and fused with the original RGB image to create an input that is more robust for a CNN classifier, especially in challenging, low-quality conditions.[26]

Advanced Models

Building on these principles, researchers have developed specialized network architectures designed to learn directly in the frequency domain.

- **FreqNet** is a lightweight CNN that explicitly integrates frequency domain learning. It forces the detector to continuously focus on high-frequency information from both the image and the features extracted by the model itself. This approach is designed to enhance generalization by learning a more fundamental signal of forgery rather than overfitting to source-specific artifacts in the spatial domain.[27]
- **Deep Frequency Filtering (DFF)** is a novel technique that operates on the feature maps within a neural network. It performs an FFT on the latent space representations at different layers and uses a learned attention mask to enhance frequency components that are transferable across domains while suppressing

those that are not, thereby improving generalization.[28]

2.3 The Physical World as a Ground Truth: Inconsistency Analysis

A promising and rapidly developing frontier in detection involves moving beyond low-level statistical artifacts to analyze high-level semantic and physical inconsistencies. The rationale is that generative models, trained on vast datasets of 2D images, learn statistical patterns of appearance but often fail to develop a true, underlying model of 3D world physics. This failure results in the generation of images that, while visually plausible at first glance, contain subtle or overt violations of physical laws related to geometry, lighting, and shadow.[29] These high-level cues may prove more robust and generalizable than low-level artifacts, which can be easily removed by simple perturbations or eliminated in future generations of models.

2.3.1 Illumination and Shadow Incoherence

The interaction of light and shadow in a scene is governed by strict physical principles. Generative models frequently violate these principles.

- **Physics-Based Forensics:** Traditional physics-based forensic methods have long used cues like illumination direction and color to detect image manipulations.[31] These methods model the interaction of light sources with object surfaces to find inconsistencies.
- **Detecting Flawed Shadows:** Recent work has demonstrated that state-of-the-art generative models systematically produce images with incorrect object-shadow relationships.[29] These errors can manifest as discrepancies in shadow direction, length, softness, or color that are inconsistent with the scene's inferred light sources. Specialized classifiers can be trained to detect these errors by analyzing object and shadow instance masks extracted from the image, effectively learning to spot physically implausible illumination.[30]

2.3.2 Flawed Projective and 3D Geometry

Generative models also struggle to consistently replicate the rules of 3D perspective and geometry as they are projected onto a 2D image plane.

- **Projective Geometry Errors:** These are fundamental errors in how a 3D scene is represented. Analysis shows that generated images often exhibit inconsistent vanishing points (where parallel lines fail to converge correctly), distortion of geometric figures (e.g., squares rendered as incorrect trapezoids), and scale discrepancies (where objects do not shrink appropriately with distance).[30] Classifiers can be trained to detect these anomalies by analyzing features like line segments or perspective fields derived from the image.[30]
- **3D-Aware Detection:** The development of 3D-aware generative models has, paradoxically, opened up new avenues for detection. These models attempt to learn an underlying 3D representation, which can then be checked for consistency. For deepfakes, this allows for methods that analyze the 3D geometry of a face to detect swaps that might be seamless in 2D but are inconsistent in 3D, making them robust to changes in pose.[34] Other methods extend this to analyze the physical consistency of human subjects in a scene, looking for inconsistencies in biometric signals like gaze direction, mismatches between face and body characteristics (e.g., age or gender), and incompatible appearance (e.g., resolution, color) between different faces in a multi-face scene.[36]

2.4 Proactive Defense: Watermarking and Content Authentication

The detection methods discussed thus far are primarily *passive* or *reactive*; they analyze a piece of media "as is" to determine its authenticity. A complementary and increasingly important approach is *active* or *proactive* defense, which involves embedding an invisible signal or "watermark" into content *before* it is distributed.[19] This shifts the burden of proof from demonstrating forgery to proving authenticity.

Watermarking Techniques

Digital watermarking for synthetic media detection is a burgeoning field with several distinct approaches. Watermarks can be embedded at different stages of the

generative process [19]:

- **Pre-generation:** The watermark is embedded in the initial seed or noise vector before the generation process begins. An example is the Tree-Ring method for Stable Diffusion.[19]
- **In-generation:** The watermark is embedded during the model's iterative operations, making it an intrinsic part of the final output. Stable Signature is a method that follows this approach.[19]
- **Post-generation:** The watermark is added to the image after it has been fully generated. StegaStamp and its variants use a neural network to encode and decode a robust watermark in this manner.[19]

Benchmark comparisons have shown that, in general, watermark-based detectors consistently outperform passive detectors. They tend to be more accurate and, crucially, more resilient to the types of perturbations (like compression and resizing) that often defeat passive methods.[19]

Identity Watermarking

A particularly innovative proactive technique is **identity watermarking**.[38] Instead of embedding a signal in the pixels of an image, this method embeds a watermark into the high-level

semantic identity vector of a face, which is extracted using a face recognition network. This embedded signal is designed to be robust to common, non-identity-altering image modifications (e.g., changing brightness, compression). However, if the image is used in a deepfake face swap, the original identity vector is destroyed, and the watermark is lost along with it. A verification step can then check for the presence of the watermark; its absence is a strong indicator of a deepfake manipulation.

Legal and Practical Context

Proactive authentication has significant implications beyond technical performance. In legal contexts, it can fundamentally change the evidentiary burden. Instead of forcing

a party to prove that a piece of evidence is fake, the proponent of the evidence can use the embedded authentication signal to proactively prove its integrity.[39] In enterprise settings, this is critical for preventing sophisticated fraud, such as voice phishing attacks or impersonations of executives, where authenticating the source of a communication is paramount.[40]

2.5 Beyond Static Images: Multimodal and Temporal Analysis for Video

While many detection principles apply to both images and videos, video deepfakes present unique challenges and opportunities. Videos are multimodal, containing both visual and auditory information, and they have a temporal dimension. The most effective video deepfake detectors exploit these additional dimensions.

The Challenge of Video Deepfakes

Creating a convincing video deepfake often requires manipulating both the video stream (e.g., lip-syncing) and the audio stream (e.g., voice conversion).[41] Detecting these manipulations can be especially difficult in short video clips (e.g., 200 milliseconds to 1 second), where the amount of data available for analysis is severely limited.[41]

Temporal Inconsistency

A key signal in fake videos is temporal inconsistency. Real-world physics and human biometrics are generally consistent over time, but generative models can struggle to maintain this coherence from one frame to the next. Detection methods can exploit this by using architectures that process sequences, such as combining a CNN feature extractor with a Recurrent Neural Network (RNN) like an LSTM or GRU.[42] A more targeted approach is taken by

TI2Net (Temporal Identity Inconsistency Network), which specifically captures

dissimilarities in the facial identity vectors extracted from different frames of the same video. While a real person's identity vector should remain highly consistent, a deepfake may exhibit subtle fluctuations, providing a strong detection signal.[43]

Audio-Visual Inconsistency

The most powerful video detection methods are often multimodal, analyzing the consistency *between* the audio and video streams. For example, a model can be trained to detect subtle asynchronies or mismatches between a person's lip movements and the phonemes being spoken in the audio track. The FakeAVCeleb dataset was created specifically to benchmark these types of multimodal detectors.[44] Research has shown that a multimodal approach is more effective than a monomodal one, providing more robust predictions.[45]

Significantly, a recent study demonstrated that an effective multimodal detector can be developed even when trained on separate, disjoint monomodal datasets (i.e., one dataset of video-only fakes and another of audio-only fakes).[45] This is a crucial finding, as large-scale, high-quality, and diverse multimodal deepfake datasets are still relatively scarce. This frees developers from the constraint of needing perfectly paired audio-visual fakes for training, broadening the range of data that can be used to build robust systems.

Part III: The Practitioner's Toolkit: Datasets, Benchmarks, and Tools

The theoretical understanding of detection methodologies must be grounded in the practical resources required to build, train, and validate a high-performance system. The selection of appropriate datasets, adherence to standardized evaluation protocols, and the use of reliable open-source or commercial tools are critical determinants of a project's success. The most significant ongoing engineering challenge for any detection project is not merely the initial choice of a model architecture but the development of a dynamic and evolving dataset strategy. A "train once, deploy forever" approach is fundamentally flawed in a field where new

generative models are released on a streaming basis.[16] The constant emergence of new generators guarantees that any detector trained on a static dataset will eventually become obsolete.[9] Therefore, long-term success hinges on establishing a robust pipeline for continuously monitoring the generative landscape, acquiring and curating new data, and regularly re-evaluating and re-training the detection system. This data-centric engineering effort is as crucial as the model development itself.

3.1 Training and Evaluation: A Guide to Benchmark Datasets

The performance and, more importantly, the generalization capability of any deep learning-based detector are fundamentally dictated by the data on which it is trained and evaluated. The historical lack of large-scale, diverse, and modern datasets has been a significant impediment to progress in the field.[46]

The Critical Role of Data

Early detectors were often trained on limited datasets containing images from only a few, now-outdated GAN architectures. While these detectors could achieve high performance on their specific training and test sets, they failed to generalize to the wider world of synthetic media. The recognition of this limitation has spurred the creation of a new generation of large, diverse, and challenging benchmark datasets.

Key Public Datasets

Selecting the appropriate dataset is a foundational strategic decision. The following table provides a comparative overview of major publicly available datasets, highlighting their primary focus, scale, and intended use cases. This allows for the strategic selection of data resources that align with specific project goals, whether that be detecting general synthetic images, video deepfakes, or assessing robustness to "in-the-wild" conditions.

Dataset Name	Primary Focus	Scale & Content	Generative Models Included	Key Features & Use Cases	Source Snippets
GenImage	General synthetic image detection, cross-generator generalization	1M+ real/fake image pairs. Broad range of ImageNet classes.	Advanced Diffusion Models (e.g., SD, Wukong, Glide) and GANs (e.g., BigGAN).	Large scale, rich content. Designed for cross-generator and degraded image classification tasks.	46
ACID	Comprehensive, modern synthetic image detection	13M samples. Fine-grained text prompts, multiple resolutions.	Over 50 different generative models, including recent ones like Stable Diffusion XL.	Very large and diverse set of modern generators. Good for testing robustness against a wide array of sources.	57
FaceForensics++	Video Deepfake Detection (Face Manipulation)	1000 original videos, manipulated with 4 methods.	Deepfakes, Face2Face, FaceSwap, NeuralTextures.	The standard benchmark for academic research on video face forgery detection.	64
DeepfakeBench Datasets	Standardized Deepfake Detection (Image & Video)	Supports 9+ datasets including FF++, Celeb-DF, etc. Pre-processed with cropped faces,	Varies by sub-dataset, covers a wide range from GANs to newer methods.	Part of a unified benchmarking platform for fair comparison. Provides pre-processed data to speed up	48

		masks.		research.	
AI-GenBench	Temporal Generalization	Dataset with images from historically-ordered generators.	Intended to be continuously updated with new models as they are released.	Unique temporal evaluation protocol to test generalization to *future*, unseen generators.	[47]
RAID	Robustness to Adversarial Attacks	72k diverse and highly transferable adversarial examples.	N/A (dataset of attacks, not generated images).	Specifically designed to provide a standardized, quick way to assess a detector's adversarial robustness.	[50]
FakeAVCeleb	Multimodal (Audio-Visual) Deepfake Detection	Videos with both real and synthetically generated/swapped audio and video.	N/A (focus is on the multimodal nature of fakes).	The primary benchmark for methods that detect fakes by analyzing audio-visual inconsistencies.	[44]

Emerging Benchmark Philosophy

The philosophy behind benchmarking is also evolving. A consensus is forming that "breadth" (having smaller samples from as many different generators and real-world scenarios as possible) is more important for evaluating modern detectors than "depth" (having millions of samples from just a few generators).[9] This shift is a direct response to the diverse and rapidly changing landscape of generative AI. Furthermore, new benchmarks like

AI-GenBench are introducing novel evaluation protocols, such as a temporal framework that trains and tests detectors on generators in their historical order of release, to more accurately measure a model's ability to generalize to future, unseen technologies.[47]

3.2 Measuring What Matters: Evaluation Metrics and Protocols

To meaningfully compare different detection methods, it is essential to use a standardized set of evaluation metrics and protocols. The lack of such standardization has historically made it difficult to assess the true state of the art, leading to unfair comparisons and potentially misleading conclusions.[48]

Standard Metrics

For the binary classification task of real vs. fake, several standard metrics are employed:

- **Accuracy:** Defined as Accuracy=(TP+TN)/(TP+TN+FP+FN), where TP, TN, FP, and FN are true positives, true negatives, false positives, and false negatives, respectively. While intuitive, accuracy can be misleading on datasets with a significant class imbalance.[49]
- **F1-Score:** The harmonic mean of precision and recall, calculated as F1=2·(Precision·Recall)/(Precision+Recall). It provides a more balanced measure of performance on imbalanced datasets.[49]
- **Area Under the ROC Curve (AUC):** The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate against the False Positive Rate at various classification thresholds. The AUC summarizes this curve into a single value, representing the model's capability to distinguish between the two classes across all possible thresholds. An AUC of 1.0 indicates a perfect classifier, while 0.5 represents a random guess.[49]
- **Equal Error Rate (EER):** This is the point on the ROC curve where the False Positive Rate is equal to the False Negative Rate. It represents the error rate at the threshold where the two types of errors are balanced.

The Need for Standardization: DeepfakeBench

To address the issue of non-standardized evaluation, the research community has developed comprehensive benchmarking platforms. The most prominent of these is **DeepfakeBench**, an open-source project designed to provide a unified and fair environment for evaluating deepfake detectors.[48] Its key features include:

- **An Integrated Framework:** It supports over 36 state-of-the-art detection methods for both images and videos.
- **A Unified Data Management System:** It ensures that all detectors are fed data from a consistent processing pipeline, eliminating a major source of variability in results. It supports over 9 major deepfake datasets.
- **Standardized Evaluation Protocols:** It implements a comprehensive suite of metrics, including frame-level and video-level AUC, accuracy, EER, and others, allowing for transparent and reproducible performance comparisons.

Using a standardized platform like DeepfakeBench is highly recommended for any serious detection project, as it ensures that performance results are comparable to the broader research field.

3.3 Implementation Pathways: Open-Source and Commercial Tools

For practitioners looking to implement a detection system, there are several pathways, ranging from building a custom model using open-source code to integrating a turnkey commercial API.

Open-Source Repositories (GitHub)

The open-source community provides a wealth of resources for building and experimenting with detection models. Key repositories include:

- **DeepfakeBench:** As mentioned, this is not just a benchmark but also a repository

containing the implementation of numerous state-of-the-art detectors.[48]

- **SSP-AI-Generated-Image-Detection:** This repository provides the code for the paper "A Single Simple Patch is All You Need for AI-generated Image Detection," offering a simple but surprisingly effective baseline model that can be trained on datasets like GenImage.[52]
- **ai-image-detector:** This tool is built around a fine-tuned Convolutional Vision Transformer (CvT-13) model trained on a massive dataset of 2.5 million images. The repository includes scripts for training, evaluation, and prediction on new images, making it a practical starting point.[53]
- **ImageAI:** This is a broader Python library for computer vision that includes pre-trained models for image classification and object detection using standard architectures like ResNet50 and InceptionV3. It also provides tools for training custom models on new datasets.[54]
- **Survey Project Pages:** The GitHub pages associated with major survey papers, such as the one from Purdue University on detecting LAIM-generated multimedia, are valuable resources that often link to the code for the various detectors they review.[5]

Commercial APIs and Services

For teams without the resources or desire to build a custom model from scratch, several commercial services offer detection capabilities via an API.

- **LandingAI:** This company provides a platform and a suite of "Agentic Vision APIs" for various computer vision tasks, including advanced document analysis, object detection, and visual grounding. Their LandingLens platform is a low-code environment for building and deploying custom vision models.[55] While not a dedicated "fake image detector" out of the box, its powerful components could be leveraged to build such a system.
- **User-Facing Detection Tools:** A number of commercial, user-facing tools like AI Detector, Winston AI, and Detecting-AI have emerged.[56] These services often bundle image and text detection and are marketed for ease of use. However, their underlying technology is typically proprietary and non-transparent, and their performance and robustness against state-of-the-art generative models may not be rigorously benchmarked in the public domain. They represent the current state of commercially available, end-user tools but may not be suitable for high-stakes, enterprise-level applications that require deep technical validation.

Part IV: Overcoming the Hurdles: Core Challenges and Mitigation Strategies

Despite significant progress, the field of synthetic image detection is fraught with persistent challenges that can severely limit the real-world effectiveness of even the most accurate lab-tested models. A successful detection project must not only choose a powerful architecture but also develop explicit strategies to address the core issues of generalization, robustness, adversarial vulnerability, and explainability. These challenges are often in tension with one another; for instance, a model optimized for high accuracy on a known dataset may lack generalization, while a highly robust and generalizable model might operate as an unexplainable "black box." Navigating this "trilemma" is a central strategic task. A system designed for forensic evidence in court must prioritize explainability and robustness, whereas a system for high-throughput social media moderation might prioritize generalization and speed.

4.1 The Generalization Gap: Failing on Unseen Generators

The single most common and critical failure mode for synthetic image detectors is the inability to generalize. A detector can be trained to achieve over 99% accuracy on images from a specific set of generators, only to see its performance plummet when tested against images from a new, unseen generator.[7]

The Problem

This "generalization gap" arises because the detector often overfits to the unique, low-level artifacts or "fingerprints" of the generators in its training data. Instead of learning a universal, abstract concept of "fakeness," it learns to recognize the specific signature of Midjourney v5 or Stable Diffusion v1.5. When a new model like DALL-E 3 is released with a different architectural signature, the detector is easily fooled. This

makes the development of a truly future-proof detector a formidable challenge.

Mitigation Strategies

Addressing the generalization gap is a primary focus of current research. Key strategies include:

- **Training on Diverse Generators:** The most direct, if brute-force, approach is to expand the training dataset to include images from as many different generative models as possible.[7] This is the guiding principle behind the creation of massive, diverse datasets like
  ACID, which contains images from over 50 different models.[57] The goal is to expose the model to a wide variety of forgery signals, forcing it to learn more generalizable features.
- **Learning Generalizable Features:** A more sophisticated approach is to design models that inherently focus on more universal signals of forgery. This includes methods that operate in the **frequency domain**, which can capture artifacts common to entire classes of generative architectures (e.g., those using up-sampling).[27] It also includes the emerging field of
  physical inconsistency detection, which looks for violations of universal laws of physics (e.g., impossible shadows or flawed perspective) that are likely to be a common failure mode for any 2D-trained model, regardless of its specific architecture.[30]
- **Language-Guided and Contrastive Learning:** Advanced training techniques can also improve generalization. For example, the LASTED method uses language-guided contrastive learning to build a more robust feature space that is less dependent on generator-specific artifacts.[20]

4.2 The Real-World Gauntlet: Robustness to Perturbations

Beyond unseen generators, the second major hurdle is the real-world environment itself. Images shared online are rarely in their pristine, original state. They are subjected to a host of common processing operations, or "perturbations," that can severely degrade detector performance.

The Problem

The subtle, often high-frequency artifacts that many detectors rely on can be easily masked or destroyed by operations like JPEG compression, image resizing, the addition of Gaussian noise, or blurring.[7] An image uploaded to a social media platform, for example, is almost always re-compressed and resized, which can significantly limit the effectiveness of forensic analysis.[17] A detector that performs perfectly on clean data but fails on a compressed image is not viable for most practical applications.

Mitigation Strategies

Building robustness to these perturbations is a critical engineering task.

- **Data Augmentation:** The most effective and widely used strategy is to proactively train the model to be resilient to these changes. This is achieved through data augmentation, where the training images (both real and fake) are randomly subjected to the very perturbations the model is expected to encounter in the wild, such as various levels of JPEG compression, resizing, and noise addition.[7] This forces the model to learn features that survive these transformations.
- **Robust Feature Engineering:** Another strategy is to build the detector around features that are inherently more robust to degradation. Proactive **watermarking** methods are often explicitly designed to be resistant to common perturbations like compression and cropping.[19] Similarly, some high-level **physical inconsistency** cues (e.g., a shadow pointing in the wrong direction) are semantic in nature and less likely to be affected by low-level signal processing than, for example, a delicate frequency artifact.
- **Understanding Attack Vectors:** Research into attacks like SR-attack, which uses super-resolution to deliberately obscure forgery artifacts, highlights the need for detectors that are robust not just to incidental but also to intentional signal degradation.[58]

4.3 The Adversarial Arms Race: Vulnerability to Attacks

Like virtually all deep learning systems, synthetic image detectors are vulnerable to adversarial attacks. These are carefully crafted, often imperceptible modifications to an input image designed specifically to cause the model to make a mistake.

The Problem

The existence of adversarial attacks means that a motivated actor with knowledge of the detection system can often bypass it. This creates an "arms race" where defenders must constantly anticipate and patch vulnerabilities exploited by attackers. These attacks can be broadly categorized [59]:

- **Evasion Attacks:** The most common type in this context, where an attacker makes small changes to a synthetic image to make it be classified as "real" by the detector.[60]
- **Poisoning Attacks:** A more insidious attack where the adversary corrupts the detector's training data, for example by injecting subtly manipulated images with incorrect labels, thereby compromising the model's integrity from its inception.[59]
- **White-box vs. Black-box:** Attacks are classified by the attacker's level of knowledge. In a **white-box** attack, the adversary has full access to the model's architecture and parameters. In a **black-box** attack, the adversary can only query the model and observe its outputs.[50]

Mitigation Strategies

Defending against adversarial attacks is an active and challenging area of research.

- **Adversarial Training:** The primary defense is adversarial training, which involves generating adversarial examples and explicitly including them in the training set. This forces the model to learn to be invariant to these types of malicious perturbations.[59]
- **Robust Architectures:** Some research focuses on designing model architectures that are inherently more robust. For instance, it has been shown that adversarially

robust classifiers can serve as powerful and less brittle primitives for various image manipulation and synthesis tasks, suggesting that robustness is a fundamental desirable property.[61]

- **Standardized Robustness Evaluation:** Generating adversarial examples can be computationally expensive, which has hindered the routine evaluation of adversarial robustness. To address this, projects like **RAID** (Robust evaluation of AI-generated image Detectors) provide a large-scale, pre-computed dataset of diverse and transferable adversarial examples. This allows developers to quickly and efficiently assess a detector's vulnerability without having to mount a full-scale attack themselves.[50]

4.4 The "Black Box" Problem: The Imperative for Explainable AI (XAI)

The final core challenge is one of trust and transparency. Most high-performance deep learning detectors operate as "black boxes"—they can output a highly accurate prediction, but they cannot provide a human-understandable reason for that decision.[62]

The Problem

This lack of explainability is a major barrier to the adoption of detection tools in high-stakes domains. In a court of law, for example, an expert witness cannot simply state that "the AI said this photo is fake." They must be able to explain the basis for that conclusion.[39] Similarly, a journalist debunking a piece of misinformation needs to be able to articulate

why an image is a forgery. Without this transparency, the detector's output lacks the credibility required for real-world impact.[12] This need is a central component of the "beyond detection" paradigm.[5]

Interpretability vs. Explainability

It is crucial to distinguish between two related concepts [63]:

- **Interpretability:** Refers to models that are inherently transparent and whose decision-making process is understandable by humans. A simple decision tree is an example of an interpretable model. These are often called "glass box" models.
- **Explainability (XAI):** Refers to the application of post-hoc techniques to provide an explanation for a black-box model's decision. These techniques do not reveal the model's internal workings but attempt to approximate its reasoning. A key risk is that these explanations are not always faithful to the model's actual calculations and can sometimes be misleading.[63]

For forensic applications where evidentiary standards are high, true interpretability is the gold standard.

XAI Techniques

Several techniques are used to add a layer of explanation to detection models:

- **Feature Attribution / Saliency Maps:** Methods like Grad-CAM can produce a "heatmap" that highlights the pixels or regions of an image that were most influential in the model's decision. However, these maps can be noisy and difficult for a human to interpret into a coherent reason.[64]
- **LIME and SHAP:** These are popular post-hoc explanation techniques. LIME (Local Interpretable Model-agnostic Explanations) explains an individual prediction by fitting a simpler, interpretable model in its local vicinity. SHAP (SHapley Additive exPlanations) uses a game-theoretic approach to assign an importance value to each feature for a given prediction.[62]
- **Human-Centric Explanations:** An emerging area of research aims to go beyond heatmaps and feature scores to generate natural language explanations. One approach frames detection as a **Visual Question Answering (VQA)** task. Instead of just asking "Is this image fake?", the system can be asked, "Does the person's nose look fake?", and respond with a textual explanation like, "The person has overlapped eyebrows, and there is a boundary on the person's forehead".[64] This provides a much more meaningful and actionable form of explanation.

Part V: Strategic Synthesis and Future Trajectories

The preceding analysis provides a comprehensive map of the AI-generated image detection landscape. This final section synthesizes these findings into actionable architectural recommendations for a detection project and offers an outlook on the future of this dynamic and critical field.

5.1 Recommended Architectural Blueprints for Your Project

The optimal architecture for a detection system depends heavily on its intended application. There is no single "best" model, but rather a set of strategic choices and trade-offs. Two distinct blueprints are proposed to address common use cases.

Blueprint A: The General-Purpose, Robust Detector

This blueprint is designed for broad applications like large-scale social media content moderation or general-purpose forensic scanning, where high generalization across diverse generators and robustness to real-world perturbations are the primary goals.

- **Concept:** An ensemble model that fuses signals from multiple, diverse detection methodologies to create a system that is harder to fool than any single component.
- **Architecture:** A multi-branch network architecture is recommended.
  - **Backbone:** A powerful and robust modern architecture, such as a Vision Transformer (ViT) or a recent ConvNeXt variant, pre-trained on a large-scale image dataset.
  - **Branch 1 (Frequency Analysis):** A lightweight module dedicated to frequency-domain analysis. This could be inspired by FreqNet [27], taking the Discrete Cosine Transform (DCT) of the input image and feeding it into a small, dedicated CNN. This branch is tasked with finding low-level, cross-generator artifacts.
  - **Branch 2 (Geometric Analysis):** A module trained specifically to detect high-level physical inconsistencies. This could be a classifier that analyzes

derived geometric features like perspective fields or object-shadow relationships, as demonstrated in recent research.[30] This branch targets semantic-level forgeries.
    - **Fusion:** The outputs from the main backbone and the specialized branches should be combined using a late fusion mechanism (e.g., concatenating feature vectors before a final classification layer or averaging prediction scores).
- **Training Strategy:** This model should be trained on the largest and most diverse dataset available, such as GenImage or ACID.[46] The training regimen must include aggressive data augmentation with real-world perturbations, including various levels of JPEG compression, resizing, noise, and blur, to ensure robustness.[7]

Blueprint B: The High-Assurance, Explainable Forensic Tool

This blueprint is designed for high-stakes applications, such as providing evidence in legal proceedings or supporting investigative journalism, where the explainability and trustworthiness of a result are paramount.

- **Concept:** A system that prioritizes interpretability over a single probabilistic score. The output is not just a "real" or "fake" label but a dashboard of evidence for a human analyst.
- **Architecture:** A pipeline of independent, interpretable "expert" models rather than a single end-to-end black box.
    - **Step 1 (Multi-Signal Analysis):** The input image is passed through a suite of independent detectors, each designed to check for a specific, understandable type of flaw. This suite could include:
        - A DCT-based detector for known frequency artifacts.
        - A shadow consistency checker that flags physically implausible illumination.
        - A projective geometry analyzer that checks for inconsistent vanishing points.
        - For faces, a gaze direction monitor or a 3D shape consistency model.
    - **Step 2 (Explainable Reporting):** If any of the expert models raise a flag, the system uses an **XAI-VQA (Visual Question Answering)** model to generate a human-readable report.[64] Instead of a heatmap, the output would be a series of question-answer pairs that a human analyst can use to build a case. For example:

        - *Query:* "Is the shadow cast by the lamppost consistent with the sun's position?"
        - *Answer:* "No, the shadow is inconsistent. Based on other shadows in the scene, the primary light source is high and to the right, but the lamppost's shadow is cast to the right."
- **Training Strategy:** Each expert model in the pipeline is trained on a highly specific task. The goal is not to optimize a single, global accuracy metric but to build a collection of reliable, interpretable tools that empower a human expert to make a final, justifiable determination.

5.2 Future Research Directions and Emerging Threats

The field of synthetic media detection is in a constant state of flux, locked in an "arms race" with generative technologies. Several key trends will define its future trajectory.

- **The Unending Arms Race:** The dynamic between generation and detection will continue to escalate. As detection methods improve at identifying specific artifacts, generative models will be fine-tuned or adversarially trained to eliminate those very signals, making them harder to detect.[65] This necessitates a move towards detecting more fundamental and harder-to-fake inconsistencies.
- **The Primacy of Multimodal Forensics:** The future of detection for dynamic media is unequivocally multimodal. The most significant breakthroughs will come from methods that can fuse signals from multiple modalities—analyzing text, image, video, and audio streams in concert to find inconsistencies that are invisible when looking at any single modality in isolation.[5]
- **A Push for Proactive Defense and Standardization:** As the risks of synthetic media become more tangible, there will be a growing demand from industry, government, and the public for proactive solutions. Expect a greater emphasis on technologies like digital watermarking and content provenance standards that aim to authenticate content at its source rather than just detecting fakes after the fact.[6]
- **Market Growth and Commercialization:** The market for AI-powered image recognition and detection is projected to grow substantially, with some estimates projecting a market size of over $61 billion by 2025.[66] This economic driver will fuel further research and the development of more sophisticated commercial detection tools.
- **The Necessity of Continuous Benchmarking:** The rapid evolution of generative

models means that any static benchmark dataset will quickly become obsolete. The long-term health and progress of the research community will depend on the maintenance and continuous expansion of dynamic benchmarking platforms like AI-GenBench and DeepfakeBench, which are designed to incorporate new generators and detection methods as they emerge.[47]

Works cited

1. [2502.15176] Methods and Trends in Detecting Generated Images: A Comprehensive Review - arXiv, accessed on July 30, 2025, https://arxiv.org/abs/2502.15176
2. Survey of Fake Image Synthesis and its Detection - IRO Journals, accessed on July 30, 2025, https://irojournals.com/iroiip/article/view/4/4/6
3. (PDF) Detecting AI Generated Images Through Texture and Frequency Analysis of Patches, accessed on July 30, 2025, https://www.researchgate.net/publication/387792397_Detecting_AI_Generated_Images_through_Texture_and_Frequency_Analysis_of_Patches
4. AI vs. AI: Can AI Detect AI-Generated Images? - MDPI, accessed on July 30, 2025, https://www.mdpi.com/2313-433X/9/10/199
5. Detecting Multimedia Generated by Large AI Models: A Survey - arXiv, accessed on July 30, 2025, https://arxiv.org/html/2402.00045v1
6. [2402.00045] Detecting Multimedia Generated by Large AI Models: A Survey - arXiv, accessed on July 30, 2025, https://arxiv.org/abs/2402.00045
7. Synthetic Image Detection - ČVUT DSpace, accessed on July 30, 2025, https://dspace.cvut.cz/bitstream/handle/10467/113335/F3-DP-2024-Petrzelkova-Nela-_DP__Synthetic_Image_Detection.pdf
8. How good are humans at detecting AI-generated images? Learnings from an experiment - arXiv, accessed on July 30, 2025, https://arxiv.org/html/2507.18640v1
9. New Microsoft benchmark for evaluating deepfake detection prioritizes breadth, accessed on July 30, 2025, https://www.biometricupdate.com/202507/new-microsoft-benchmark-for-evaluating-deepfake-detection-prioritizes-breadth
10. [2507.18640] How good are humans at detecting AI-generated images? Learnings from an experiment - arXiv, accessed on July 30, 2025, https://www.arxiv.org/abs/2507.18640
11. Human vs. AI: A Novel Benchmark and a Comparative Study on the Detection of Generated Images and the Impact of Prompts - arXiv, accessed on July 30, 2025, https://arxiv.org/html/2412.09715v1
12. Deepfake Detection: A Comprehensive Survey from the Reliability Perspective - arXiv, accessed on July 30, 2025, https://arxiv.org/html/2211.10881v3
13. Human vs. AI: A Novel Benchmark and a ... - ACL Anthology, accessed on July 30, 2025, https://aclanthology.org/2025.genaidetect-1.2.pdf
14. Daily Papers - Hugging Face, accessed on July 30, 2025, https://huggingface.co/papers?q=synthetic%20image%20detection

15. [2403.17881] Deepfake Generation and Detection: A Benchmark and Survey - arXiv, accessed on July 30, 2025, https://arxiv.org/abs/2403.17881
16. [2310.15150] Online Detection of AI-Generated Images - arXiv, accessed on July 30, 2025, https://arxiv.org/abs/2310.15150
17. AI-Generated-Image Detection Using Deep Learning Techniques - ResearchGate, accessed on July 30, 2025, https://www.researchgate.net/publication/390721885_AI-Generated-Image_Detection_Using_Deep_Learning_Techniques
18. Evolution of Detection Performance throughout the Online Lifespan of Synthetic Images, accessed on July 30, 2025, https://arxiv.org/html/2408.11541v1
19. [Literature Review] AI-generated Image Detection: Passive or Watermark? - Moonlight, accessed on July 30, 2025, https://www.themoonlight.io/en/review/ai-generated-image-detection-passive-or-watermark
20. Synthetic Image Detection | Papers With Code, accessed on July 30, 2025, https://paperswithcode.com/task/synthetic-image-detection
21. SynArtifact: Classifying and Alleviating Artifacts in Synthetic Images via Vision-Language Model - arXiv, accessed on July 30, 2025, https://arxiv.org/html/2402.18068v2
22. NeurIPS Poster StableRep: Synthetic Images from Text-to-Image Models Make Strong Visual Representation Learners, accessed on July 30, 2025, https://neurips.cc/virtual/2023/poster/69986
23. StableRep: Synthetic Images from Text-to-Image Models Make Strong Visual Representation Learners, accessed on July 30, 2025, https://proceedings.neurips.cc/paper_files/paper/2023/file/971f1e59cd956cc094da4e2f78c6ea7c-Paper-Conference.pdf
24. Leveraging Frequency Analysis for Deep Fake Image ... - arXiv, accessed on July 30, 2025, http://arxiv.org/pdf/2003.08685
25. Enhanced Deepfake Detection Using Frequency Domain Upsampling - SciTePress, accessed on July 30, 2025, https://www.scitepress.org/Papers/2024/124737/124737.pdf
26. Frequency Domain Filtered Residual Network for Deepfake Detection, accessed on July 30, 2025, https://www.mdpi.com/2227-7390/11/4/816
27. Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Domain Learning, accessed on July 30, 2025, https://ojs.aaai.org/index.php/AAAI/article/view/28310/28609
28. Deep Frequency Filtering for Domain Generalization - CVF Open Access, accessed on July 30, 2025, https://openaccess.thecvf.com/content/CVPR2023/papers/Lin_Deep_Frequency_Filtering_for_Domain_Generalization_CVPR_2023_paper.pdf
29. CVPR Poster Shadows Don't Lie and Lines Can't Bend! Generative Models don't know Projective Geometry...for now, accessed on July 30, 2025, https://cvpr.thecvf.com/virtual/2024/poster/31180
30. Shadows Don't Lie and Lines Can't Bend ... - CVF Open Access, accessed on July 30, 2025,

https://openaccess.thecvf.com/content/CVPR2024/papers/Sarkar_Shadows_Dont_Lie_and_Lines_Cant_Bend_Generative_Models_dont_CVPR_2024_paper.pdf
31. Illumination Analysis in Physics-based Image Forensics: A Joint Discussion of Illumination Direction and Color - FAU, accessed on July 30, 2025, https://faui1-files.informatik.uni-erlangen.de/public/publications/mmsec/2017-Riess-IAP.pdf
32. Identifying Image Composites Through Shadow Matte Consistency | Request PDF, accessed on July 30, 2025, https://www.researchgate.net/publication/220177263_Identifying_Image_Composites_Through_Shadow_Matte_Consistency
33. Making Images Real Again: A Comprehensive Survey on Deep Image Composition - arXiv, accessed on July 30, 2025, https://arxiv.org/html/2106.14490v7
34. CVPR Poster 3D-Aware Face Swapping, accessed on July 30, 2025, https://cvpr.thecvf.com/virtual/2023/poster/22645
35. 3D-Aware Face Swapping - CVF Open Access - The Computer ..., accessed on July 30, 2025, https://openaccess.thecvf.com/content/CVPR2023/papers/Li_3D-Aware_Face_Swapping_CVPR_2023_paper.pdf
36. Seeing Through Deepfakes: A Human-Inspired Framework for Multi-Face Detection - arXiv, accessed on July 30, 2025, https://arxiv.org/html/2507.14807v1
37. GazeForensics: DeepFake Detection via Gaze-guided Spatial Inconsistency Learning, accessed on July 30, 2025, https://arxiv.org/html/2311.07075
38. Proactive Deepfake Defence via Identity ... - CVF Open Access, accessed on July 30, 2025, https://openaccess.thecvf.com/content/WACV2023/papers/Zhao_Proactive_Deepfake_Defence_via_Identity_Watermarking_WACV_2023_paper.pdf
39. Deepfakes in the Courtroom: Problems and Solutions - Illinois State Bar Association, accessed on July 30, 2025, https://www.isba.org/sections/ai/newsletter/2025/03/deepfakesinthecourtroomproblemsandsolutions
40. Proactively Protect Your Business with Deepfake Audits - Pindrop Security, accessed on July 30, 2025, https://www.pindrop.com/article/protect-your-business-with-deepfake-audits/
41. Multimodal Deepfake Detection for Short Videos - SciTePress, accessed on July 30, 2025, https://www.scitepress.org/Papers/2024/125573/125573.pdf
42. A Review of Deep Learning-based Approaches for Deepfake Content Detection - arXiv, accessed on July 30, 2025, https://arxiv.org/html/2202.06095v3
43. TI2Net: Temporal Identity Inconsistency Network for Deepfake Detection - CVF Open Access, accessed on July 30, 2025, https://openaccess.thecvf.com/content/WACV2023/papers/Liu_TI2Net_Temporal_Identity_Inconsistency_Network_for_Deepfake_Detection_WACV_2023_paper.pdf
44. Multimodal Forgery Detection - Papers With Code, accessed on July 30, 2025, https://paperswithcode.com/task/multimodal-forgery-detection
45. A Robust Approach to Multimodal Deepfake Detection - PMC, accessed on July

30, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10299653/
46. GenImage: A Million-Scale Benchmark for Detecting AI-Generated ..., accessed on July 30, 2025, https://proceedings.neurips.cc/paper_files/paper/2023/hash/f4d4a021f9051a6c18183b059117e8b5-Abstract-Datasets_and_Benchmarks.html
47. AI-GenBench: A New Ongoing Benchmark for AI-Generated Image Detection - arXiv, accessed on July 30, 2025, https://arxiv.org/html/2504.20865v1
48. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection - NIPS, accessed on July 30, 2025, https://proceedings.neurips.cc/paper_files/paper/2023/file/0e735e4b4f07de483cbe250130992726-Paper-Datasets_and_Benchmarks.pdf
49. Deepfake: definitions, performance metrics and standards, datasets, and a meta-review, accessed on July 30, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11408348/
50. RAID: A Dataset for Testing the Adversarial Robustness of AI-Generated Image Detectors, accessed on July 30, 2025, https://arxiv.org/html/2506.03988v1
51. SCLBD/DeepfakeBench: A comprehensive benchmark of ... - GitHub, accessed on July 30, 2025, https://github.com/SCLBD/DeepfakeBench
52. bcmi/SSP-AI-Generated-Image-Detection - GitHub, accessed on July 30, 2025, https://github.com/bcmi/SSP-AI-Generated-Image-Detection
53. guyfloki/ai-image-detector: A tool to distinguish between AI ... - GitHub, accessed on July 30, 2025, https://github.com/guyfloki/ai-image-detector
54. OlafenwaMoses/ImageAI: A python library built to empower developers to build applications and systems with self-contained Computer Vision capabilities - GitHub, accessed on July 30, 2025, https://github.com/OlafenwaMoses/ImageAI
55. LandingAI: Computer Vision Platform - Visual AI Software, accessed on July 30, 2025, https://landing.ai/
56. I Tested 30+ AI Detectors. These 9 are Best to Identify Generated Text. - Medium, accessed on July 30, 2025, https://medium.com/freelancers-hub/best-ai-detectors-2025-35a58eac86c5
57. ACID: A Comprehensive Dataset for AI-Created Image Detection - OpenReview, accessed on July 30, 2025, https://openreview.net/forum?id=1P6AqR6xkF
58. Robustness and Generalization of Synthetic Images Detectors - CEUR-WS.org, accessed on July 30, 2025, https://ceur-ws.org/Vol-3762/503.pdf
59. 6 Key Adversarial Attacks and Their Consequences - Mindgard, accessed on July 30, 2025, https://mindgard.ai/blog/ai-under-attack-six-key-adversarial-attacks-and-their-consequences
60. Adversarial AI: Understanding and Mitigating the Threat - Sysdig, accessed on July 30, 2025, https://sysdig.com/learn-cloud-native/adversarial-ai-understanding-and-mitigating-the-threat/
61. Image Synthesis with a Single (Robust) Classifier, accessed on July 30, 2025, http://papers.neurips.cc/paper/8409-image-synthesis-with-a-single-robust-classifier.pdf

62. Developing an Explainable AI System for Digital Forensics ..., accessed on July 30, 2025, https://www.forensicscijournal.com/articles/jfsr-aid1089.php
63. Interpretable algorithmic forensics - PNAS, accessed on July 30, 2025, https://www.pnas.org/doi/10.1073/pnas.2301842120
64. Common Sense Reasoning for Deepfake Detection - European Computer Vision Association, accessed on July 30, 2025, https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/12295.pdf
65. Learning Self-Consistency for Deepfake Detection - CVF Open Access, accessed on July 30, 2025, https://openaccess.thecvf.com/content/ICCV2021/papers/Zhao_Learning_Self-Consistency_for_Deepfake_Detection_ICCV_2021_paper.pdf
66. How Image Recognition Powers Modern AI Applications in 2025 - Vertu, accessed on July 30, 2025, https://vertu.com/ai-tools/how-image-recognition-powers-ai-applications-2025/