UNMASKING SYNTHETIC REALITIES IN GENERATIVE AI: A COMPREHENSIVE REVIEW OF ADVERSARIALLY ROBUST DEEPFAKE DETECTION SYSTEMS

A PREPRINT

Naseem Khan
Department of Computer Science
Hamad bin Khalifa University
Qatar
nakh12498@hbku.edu.qa

Tuan Nguyen
Qatar Computing Research Institute
Hamad bin Khalifa University
Qatar
ntuan@hbku.edu.qa

Amine Bermak
Department of Computer Science
Hamad bin Khalifa University
Qatar
abermak@hbku.edu.qa

Issa M. Khalil
Qatar Computing Research Institute
Hamad bin Khalifa University
Qatar
ikhalil@hbku.edu.qa

July 30, 2025

ABSTRACT

The rapid advancement of Generative Artificial Intelligence (GAI) has fueled the proliferation of deepfakes—synthetic media encompassing both fully generated content and subtly edited authentic material—posing profound challenges to digital security, misinformation mitigation, and identity preservation. This systematic review critically evaluates state-of-the-art deepfake detection methodologies, with an emphasis on reproducible, publicly available implementations to foster transparency and scientific validation. The analysis delineates two core paradigms: (1) the detection of fully synthetic media, leveraging statistical anomalies and hierarchical feature extraction, and (2) the localization of manipulated regions within authentic content, often employing multi-modal cues such as visual artifacts and temporal inconsistencies. These approaches, spanning uni-modal and multi-modal frameworks, demonstrate notable precision and adaptability in controlled settings, effectively identifying manipulations through advanced learning techniques and cross-modal fusion.

However, a comprehensive assessment reveals a pervasive limitation: the insufficient evaluation of adversarial robustness across both paradigms. Current methods, while proficient against known generative techniques, exhibit vulnerability to adversarial perturbations—subtle, intentional alterations designed to evade detection—undermining their reliability in real-world adversarial contexts. This gap highlights a critical disconnect between methodological development and the evolving threat landscape of GAI-driven attacks. To address this, we contribute a curated GitHub repository aggregating open-source implementations of the reviewed methods, enabling researchers to replicate, extend, and stress-test these approaches. Our findings emphasize the urgent need for future work to prioritize adversarial resilience, advocating for the design of scalable, modality-agnostic architectures capable of withstanding sophisticated manipulations. This review not only synthesizes the strengths and shortcomings of contemporary deepfake detection but also charts a path toward robust, trustworthy systems amid escalating digital threats. Link to github repository: https://github.com/Magnet200/SOT_Deepfake_Detection_Mechanisms

Keywords Deepfake Detection, Generative AI, Adversarial Robustness, Multi-modal Detection, Cross-domain Generalization, Self-Supervised Learning

1 Introduction

Generative Artificial Intelligence (GAI) refers to the class of AI models designed to generate synthetic data that closely resembles real-world input. Initially developed to augment imbalanced datasets using techniques such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Autoregressive models, GAI has evolved into a powerful paradigm driving advancements across multiple domains [1, 2]. Beyond its foundational role in data augmentation, GAI has revolutionized content creation, enabling human-computer collaboration (HCC) in areas such as art, music, literature, healthcare, and scientific research [3, 4]. Its widespread adoption has significantly impacted industries by improving efficiency, reducing costs, and fostering innovation [5].

Modern GAI models generate content across diverse modalities, including text, images, audio, video, and code. Prominent applications such as ChatGPT, Bard (Gemini), Midjourney, Copilot, DALL-E, and Synthesia demonstrate its broad utility [6, 7]. While these technologies offer substantial benefits, they also introduce ethical, security, and privacy concerns. The ability of generative models to produce highly realistic yet synthetic media raises concerns about misinformation, intellectual property rights, and privacy breaches. Consequently, research efforts increasingly focus on developing ethical and regulatory frameworks to ensure responsible deployment and forensic analysis of generative systems [8, 9].

Despite their transformative potential, GAI systems pose significant risks when misused. In education, the unrestricted use of generative models undermines academic integrity and critical thinking skills [10]. In business, inadequate regulatory controls expose organizations to security vulnerabilities and ethical dilemmas [11]. In healthcare, biased training data can lead to inaccurate AI-assisted diagnostics and decision-making [12]. Moreover, the propensity of GAI models to memorize and regenerate sensitive training data raises privacy concerns, particularly in journalism and legal contexts where confidential information must be safeguarded [13]. These risks underscore the broader societal implications of GAI, including job displacement, socio-economic inequalities, and the potential weaponization of synthetic content for manipulation and disinformation campaigns [14].

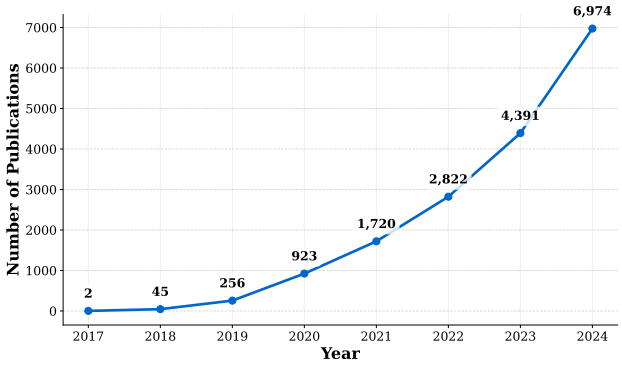

Figure 1: The graph illustrates the annual publication count in the field of DeepFakes. The data, obtained from dimensions.ai [15], highlights the development trend of Deepfakes detection from 2017 to 2024.

Among the most pressing challenges associated with GAI is the proliferation of synthetic media, commonly known as DeepFakes. DeepFakes leverage generative models to manipulate visual, auditory, and textual content, posing substantial threats to digital security, democratic stability, and public trust. As illustrated in Figure 1, research interest in DeepFake detection has surged in response to the increasing sophistication and accessibility of these technologies [15]. Malicious use cases include disinformation campaigns, identity fraud, extortion, and non-consensual explicit content generation, as exemplified in Figure 2. These developments highlight the urgent need for robust forensic techniques capable of identifying and mitigating synthetic media threats. In particular, advancing adversarially robust detection mechanisms is crucial to ensuring the integrity and trustworthiness of digital content [16].

This systematic review explores the current state of DeepFake detection methodologies, with a focus on uni-modal and multi-modal approaches, adversarial robustness, and cross-domain generalization. By critically analyzing existing frameworks, we aim to identify gaps in detection strategies and propose directions for future research toward more resilient and scalable solutions.

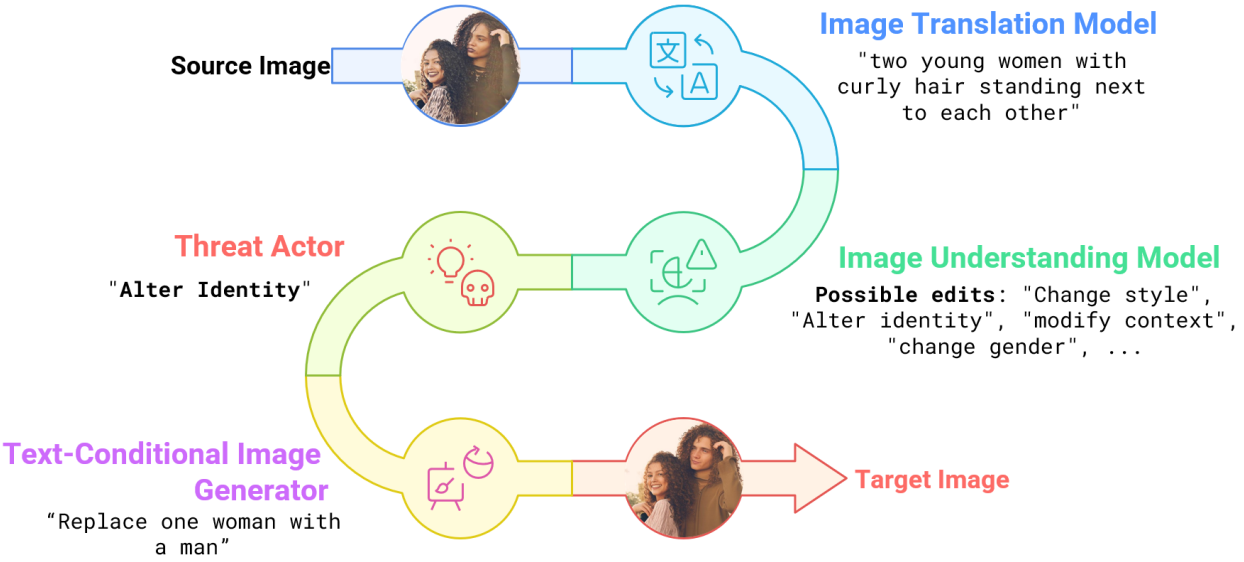

Figure 2: Illustration of a multi-stage pipeline in which a threat actor manipulates a source image using a text-conditional image generator, guided by identity-altering instructions, to produce a synthetic target image with modified personal attributes.

2 Background and Technical Foundation

The evolution of deepfake technology is intrinsically linked to advancements in generative artificial intelligence (GAI). Since its emergence in 2017, deepfake technology has rapidly progressed from rudimentary face-swapping applications to sophisticated synthetic multimedia generation systems. This advancement is primarily driven by developments in generative models, notably Generative Adversarial Networks (GANs) [17], Diffusion Models (DMs) [18], and Variational Autoencoders (VAEs) [19], which collectively underpin modern deepfake creation.

Early generative models relied on standalone architectures, such as basic encoder-decoder frameworks. However, the field has since shifted toward sophisticated conditional generation paradigms [20], enabling precise control over synthetic outputs. This shift has facilitated the integration of multiple modalities—visual, auditory, and textual—enhancing both the quality and versatility of generated media. For example, Conditional GANs [21, 22, 23] allow generators to produce tailored synthetic content based on specific input conditions, such as facial landmarks or text prompts. Similarly, Diffusion Models [24, 25, 26] employ iterative denoising processes to create high-fidelity synthetic media guided by user-defined constraints, offering a robust alternative to GAN-based approaches.

Recent innovations in multi-modal learning have further elevated synthetic media generation, particularly in text-to-image (T2I) and text-to-video (T2V) domains. Models like DALL-E [27] exemplify this capability by generating photorealistic images from textual descriptions, while T2V advancements leverage pre-trained visual representations to produce coherent video sequences without requiring paired text-video training data [28]. These developments are supported by spatial-temporal architectures that enhance motion modeling and resolution, resulting in highly realistic video content [28]. Such progress has significantly expanded the scope of deepfake applications across modalities.

In the visual domain, deepfakes benefit from conditional image synthesis and high-resolution video generation techniques. Audio deepfakes, driven by models such as WaveNet [36] and Tacotron [37], utilize generative architectures trained on speech data to produce realistic synthetic voices, often synchronized with visual elements for enhanced authenticity. Text-based synthetic content, powered by large language models like GPT [38], generates human-like narratives that can complement other modalities. The convergence of these advances has given rise to multi-modal deepfakes, where synchronized audio, visual, and textual components create highly convincing synthetic media. High-profile cases, such as political misinformation campaigns and identity spoofing [39], highlight the growing realism and accessibility of this technology.

This increased sophistication, however, presents formidable challenges for detection systems. Traditional methods, which often exploit modality-specific artifacts—such as those in GAN-generated images [40]—struggle against modern deepfakes. Diffusion-based models, for instance, produce fewer detectable traces, complicating forensic analysis [41]. Moreover, these generative frameworks can be exploited through adversarial techniques, crafting deepfakes designed to

Table 1: Comparative analysis of prior systematic reviews on deepfake detection, highlighting their modality coverage, detection paradigm and evaluation focus.

Coverage Aspects	[29]	[30]	[31, 32]	[33]	[34]	[35]	Our Study
Detection Modalities							
Image-based	✓	●	✗	✗	●	✓	✓
Video-based	✓	✓	✗	✗	●	✓	✓
Audio-based	✗	✗	✓	✗	●	✗	✓
Text-based	✗	✗	✗	✓	●	✗	✓
Multi-modal	✗	✗	✗	✗	✓	✓	✓
Detection Paradigms							
Fully Synthetic	✓	✓	✓	✓	✓	✓	✓
Edited Region Localization	●	✗	✗	✗	✗	●	✓
Evaluation Metrics							
Cross-dataset Generalization	✓	●	●	●	●	✓	✓
Natural Perturbations	●	✗	●	✗	✗	●	✓
Adversarial Robustness	✗	✗	✓	✗	●	✗	✓

✓ Comprehensive coverage, ● Partial coverage, ✗ Limited/No coverage

evade detection algorithms. Subtle perturbations, as explored in adversarial attack research [42], can deceive classifiers, exposing vulnerabilities in existing detection frameworks, especially in real-world adversarial settings.

The seamless integration of multi-modal cues and the emergence of adversarially crafted threats underscore the urgent need for advanced detection strategies. These must address not only the realism of modern deepfakes but also their intentional manipulations across modalities. The remainder of this review critically evaluates state-of-the-art uni-modal and multi-modal deepfake detection approaches, with a focus on enhancing adversarial robustness to counter these evolving synthetic challenges.

3 Systematic Review Methodology

To systematically assess advancements in deepfake detection, we conducted a structured literature review, emphasizing model robustness, transferability, and multi-modal approaches. Traditional detection methods rely on publicly available datasets, limiting their effectiveness against novel generative architectures. The rapid evolution of deepfake generation, particularly with techniques such as Low-Rank Adaptation (LoRA) [43], has increased the diversity of synthetic content, making exhaustive model-specific training impractical. This review, therefore, examines the generalization capability of detection models across datasets and their adaptability to emerging threats.

3.1 Search Strategy and Inclusion Criteria

Our review follows a systematic approach for selecting relevant literature. We queried the following databases:

- **Databases**: Dimensions.ai, Semantic Scholar, IEEE Xplore, ACM Digital Library, and arXiv.
- **Search Terms**: "Deepfake detection," "Image/Audio/Video/Text-based deepfake detection," "adversarial robustness in AI forensics," "uni-modal deepfake detection," "multi-modal deepfake detection," "cross-domain generalization in synthetic media detection."
- **Timeframe**: Studies published from January 2023 to early 2025 were prioritized to ensure coverage of the latest advancements.
- **Inclusion Criteria**: Peer-reviewed journal and conference papers focusing on image, video, audio, text-based, and multi-modal deepfake detection.
- **Exclusion Criteria**: Papers without empirical results, non-English publications, and theoretical discussions without publicly available implementation.

Each study underwent a systematic quality assessment evaluating: (1) relevance to deepfake analysis domains, (2) availability of public implementation repositories, (3) experimental validation on contemporary benchmark datasets,

and (4) methodological innovation relative to existing approaches—ensuring our review encompasses technically sound, reproducible, and state-of-the-art contributions.

3.2 Scope of Existing Systematic Reviews

The field of deepfake detection has seen extensive research, leading to multiple systematic reviews. However, existing surveys often focus on specific modalities rather than comprehensively addressing deepfake detection across different forms of synthetic media. Table 1 provides an overview of prior surveys, highlighting their scope, strengths, and limitations.

Most prior reviews target image and video-based detection, with notable contributions focusing on facial deepfake analysis and fully synthetic content identification [29, 30]. While these studies provide a strong foundation, they typically exclude audio and text-based manipulations, limiting their applicability in multi-modal contexts. Similarly, surveys on audio deepfake detection focus on speech synthesis and voice conversion techniques but do not address cross-modal threats that combine visual and auditory elements [31, 32].

Text-based deepfake detection, particularly in misinformation and fake news detection, has been explored in recent studies [33]. However, these works primarily analyze linguistic patterns and do not integrate insights from other modalities. Some broader surveys attempt to cover multiple modalities [34, 35], yet they lack a comprehensive adversarial robustness evaluation and fail to systematically assess cross-dataset generalization and real-world transferability.

While these reviews have significantly advanced the field, none provide a unified framework that integrates deepfake detection across all media formats, including image, video, audio, text, and multi-modal systems. Additionally, most prior studies focus on conventional detection approaches without addressing emerging challenges such as adversarial attacks, generative fine-tuning, and imperceptible content manipulations.

3.3 Contributions of This Review

This systematic review provides a comprehensive synthesis of deepfake detection methodologies, encompassing all primary modalities—image, video, audio, text, and multi-modal systems. It bridges critical gaps in the literature by integrating uni-modal and multi-modal approaches while offering a structured evaluation of their resilience to adversarial threats and their adaptability across diverse synthetic media contexts.

A key contribution lies in the classification of detection strategies into two distinct paradigms: (i) the identification of entirely synthetic media and (ii) the localization of manipulated regions within authentic content. The latter, an emerging and underexplored domain, demands sophisticated forensic techniques and adaptive analytical methods, which this review systematically explores.

Additionally, this study delivers an in-depth examination of cross-domain generalization, a crucial attribute for ensuring detection systems remain effective against evolving generative technologies. By assessing the transferability of detection approaches across varied synthetic media types, this review establishes a foundation for understanding their robustness in dynamic, real-world scenarios.

Adversarial robustness is a cornerstone of this analysis, with a critical evaluation of how contemporary detection systems withstand real-world perturbations and adversarial modifications. This includes a discussion of countermeasures—such as adversarial training and feature-space regularization—designed to bolster resilience against advanced manipulation techniques, without reliance on specific attack frameworks.

To uphold scientific integrity and support practical implementation, this review emphasizes studies with publicly accessible implementations, fostering reproducibility and enabling comparative assessments. A consistent evaluative framework is applied across all methodologies, scrutinizing their strengths, limitations, and applicability to operational contexts.

Finally, this review delineates key research gaps and proposes future directions, including the incorporation of explainable AI for enhanced forensic transparency, the advancement of real-time detection capabilities, and the exploration of self-supervised learning to improve generalization and robustness. By consolidating current knowledge and outlining a trajectory for future inquiry, this study serves as a vital resource for developing resilient detection systems to counter the growing sophistication of generative AI-driven synthetic media.

3.4 Conclusion and Transition to Detection Methodologies

This review consolidates prior research on deepfake detection, emphasizing its strengths and gaps. Unlike previous surveys, it systematically integrates multi-modal detection, adversarial robustness analysis, and cross-dataset generaliza-

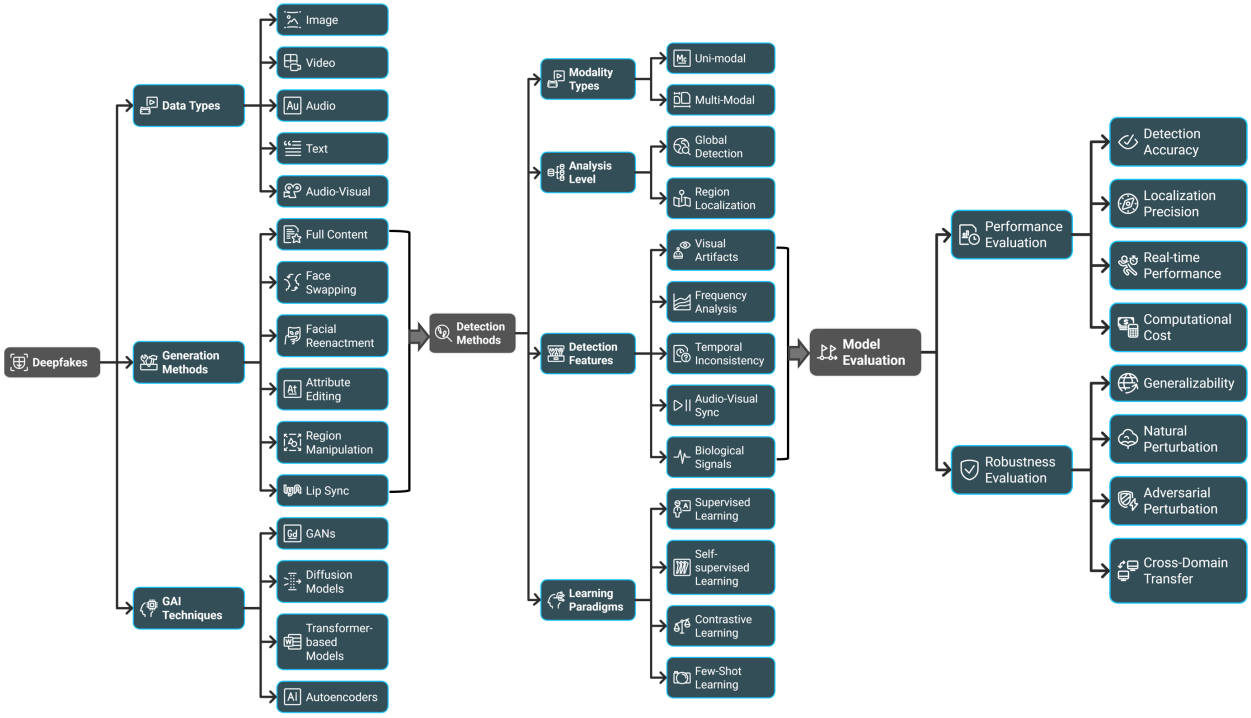

Figure 3: The broad taxonomy of Deepfake generation and detection strategies.

tion. The following section explores state-of-the-art detection methodologies, evaluating their efficacy in combating the increasing sophistication of GAI.

4 Deepfake Detection

Deepfake detection methods can be broadly categorized into *Uni-modal* and *Multi-modal* approaches, depending on whether they analyze a single data type or integrate multiple modalities for classification. Uni-modal methods focus on domain-specific artifacts within images, audio, or text, leveraging spatial inconsistencies, frequency distortions, or linguistic anomalies to detect manipulations. In contrast, multi-modal approaches enhance robustness by combining complementary information across modalities, such as synchronizing facial expressions with speech in audio-visual deepfake detection. Advanced fusion techniques, including attention mechanisms and contrastive learning, further improve cross-modal consistency analysis, enabling more reliable detection of synthetic content. The broad taxonomy of Deepfake generation and detection approaches depicted in Figure 3.

4.1 Feature Learning Strategies in Deepfake Detection

Beyond modal categorization, deepfake detection techniques can be classified based on their *feature extraction and learning mechanisms*. Traditional approaches primarily focus on spatial pattern learning, frequency spectrum analysis, and temporal consistency modeling. Spatial-based methods detect visual artifacts, inconsistencies in texture, or pixel-level anomalies introduced during synthesis [44]. Frequency-based techniques exploit traces left by generative models in the frequency domain, leveraging Fourier or wavelet transforms to capture imperceptible patterns [45]. Temporal consistency learning extends these methods by analyzing motion coherence across frames to detect artifacts in synthesized videos [46]. Textual-based methods identify linguistic anomalies or contextual inconsistencies in AI-generated text, such as fake news, using deep learning models like BERT [47].

Recent advancements have explored physiological signal analysis, such as heart rate variations (rPPG) and facial micro-expressions, to differentiate real and fake content [48]. Another emerging approach involves detecting semantic and contextual inconsistencies, focusing on unnatural facial expressions, lip-sync mismatches, or logical errors in AI-generated text [49]. To enhance robustness, adversarial perturbation analysis is employed, evaluating how detection models respond to adversarial attacks or subtle perturbations designed to bypass classifiers [42]. Furthermore, self-supervised and contrastive learning methods have gained attention for their ability to generalize across different

generative models by learning invariant feature representations [50]. Explainability-driven strategies, leveraging techniques like Grad-CAM, SHAP, TSNE, or textual description, also play a role in improving interpretability by highlighting manipulated regions in images or videos [51].

These diverse learning strategies collectively contribute to improving deepfake detection, enhancing model generalizability, and addressing evolving generative AI techniques. To systematically assess the strengths and limitations of these detection strategies, the following sections provide an in-depth review of *Uni-modal detection approaches*, evaluating their *model architectures, benchmark datasets used, performance metrics in both natural and adversarial adaptability, and their generalization ability against unseen generative models*.

4.1.1 Uni-modal Deepfake Detection

Image deepfakes: Image-based deepfake detection has been a primary focus due to the proliferation of generative models producing realistic synthetic images, ranging from fully synthetic faces to subtle manipulations. Early methods relied on *spatial pattern learning* using convolutional neural networks (CNNs) to detect pixel-level inconsistencies [52, 53]. However, their limited adaptability to diverse generative sources prompted a shift to *frequency-domain analysis*, leveraging statistical traces to enhance generalization [54, 55, 56, 57, 58, 59, 60, 61]. Recent approaches integrate *self-supervised* and *contrastive learning* to improve feature invariance and robustness [62, 63, 64, 65, 66], while lightweight architectures prioritize efficiency [67, 68]. Robustness against adversarial perturbations is addressed through watermarking and adaptive frameworks [69, 70, 71, 72, 73], with explainability-driven strategies enhancing interpretability [74, 75, 76].

Table 2: Uni-modal Deepfake Detection: Synthesis of Mechanisms and Insights

Modality	Approach	References	Strength	Challenge
Image-based Detection				
Datasets: CelebA, FFHQ, ForenSynths, LSUN, DiffusionForensics, CNNSpot, ProGAN, COCO, ImageNet, LAION, GANGen-Detection, UnivFakeDetect, MSCOCO				
Image	Frequency-Based	[77, 54, 55, 85, 86, 87, 88, 89, 84]	Uncovers subtle synthesis artifacts	Limited by evolving generative techniques
Image	Spatial-Based	[52, 53, 78, 75, 90, 79, 80, 77]	Captures pixel-level inconsistencies	Sensitive to image degradation
Image	Adaptive Learning	[58, 65, 63, 76, 82, 68, 64, 62, 81]	Adapts to diverse generative models	Requires extensive training data
Image	Robust Learning	[74, 67, 91, 83, 81]	Enhances stability under disruptions	High computational complexity
Image	Watermarking	[69, 70, 71, 72, 73]	Embeds detectable authenticity cues	Vulnerable to sophisticated attacks
Video-based Detection				
Datasets: FF++, Celeb-DF, DFDC, WildDeepfake, FakeAVCeleb, DeeperForensics, ForgeryNet, DFD, Seq-DeepFake, KODF-LS, LSR+W2L, DF40				
Video	Frame-Based	[92, 93, 94, 95, 96, 97, 98, 99]	Leverages static visual cues	Misses temporal inconsistencies
Video	Temporal-Based	[100, 101, 102, 103, 104, 105, 106, 107, 108]	Detects motion-based anomalies	Dependent on video quality
Video	Temporal+Graph	[109, 110, 111, 112]	Models relational dynamics	Complex model training
Video	Adaptive Learning	[113, 114, 115, 116, 117, 94, 118]	Handles varied forgery types	Limited by dataset diversity
Video	Robust Learning	[119, 120, 118, 121, 122]	Improves reliability in real conditions	Resource-intensive processing
Video	Hybrid/Advanced	[123, 124, 125]	Combines local and global features	Sensitive to low-quality inputs
Audio-based Detection				
Datasets: ASVspoof 2015, ASVspoof 2019, ASVspoof 2021, In-the-Wild, FakeAVCeleb, CVoiceFake, SONICS, WaveFake, EVDA, CLEAR, VSA, GRID, CD-ADD				
Audio	Artifact-Focused	[126, 127, 128, 129, 130]	Identifies synthesis imperfections	Fails with high-fidelity fakes
Audio	Temporal Modeling	[131, 132, 133]	Captures sequential patterns	Vulnerable to noise interference
Audio	Adaptive Learning	[134, 135, 136, 137, 138]	Adapts to new synthesis methods	Relies on broad data coverage
Audio	Robust Learning	[139, 140, 141, 142, 143, 144, 145]	Ensures stability across environments	Struggles with atypical audio
Text-based Detection				
Datasets: ISOT, TweepFake, OpenLLMText, PHEME, FA-KES, WebText, Ch-9, RealNews, Enhanced TweepFake, SynSciPass				
Text	Linguistic-Based	[146, 147, 148]	Exploits syntactic anomalies	Limited by style variations
Text	Transformer-Based	[47, 149, 150, 151, 152]	Leverages contextual understanding	High computational demands
Text	Hybrid/Advanced	[153]	Integrates multiple features	Challenged by evolving LLMs

Notes: FF++ = FaceForensics++, DFDC = DeepFake Detection Challenge.

New advancements include *Data-Independent Operator (DIO)* [77], a training-free method using handcrafted filters for artifact extraction, and *Neighboring Pixel Relationships (NPR)* [78], which captures local pixel correlations for source-invariant detection. *Diffusion Reconstruction Error (DIRE)* [79] and its distilled variant [68] leverage reconstruction errors to detect diffusion-generated images, while *hierarchical frameworks* [80] classify images across multiple levels. *Subudhi et al.* [81] introduce a meta-learning approach for adaptability, and *Abdullah et al.* [82] propose ensemble methods with content-agnostic features to counter fine-tuned generative models. *Chen et al.* [83] augment datasets with masked diffusion models, and *Li et al.* [84] optimize spatial-frequency collaboration for IoT security. Challenges persist in generalizing to unseen models and resisting sophisticated adversarial attacks.

Video deepfakes: Video-based detection targets temporal inconsistencies in manipulated sequences. Early frame-based methods detected spatial artifacts [97, 102, 103, 93], but overlooked inter-frame dynamics. *Temporal consistency modeling* using 3D CNNs and transformers captures lip-sync mismatches and motion irregularities [104, 100, 105], while *graph-based modeling* identifies relational anomalies [109, 95, 110]. Generalization is enhanced through *self-supervised* and *contrastive learning* [113, 111, 114], and pre-trained models with adapters [119, 117, 116]. Robustness is improved via hybrid architectures [123, 122, 120], with interpretability addressed through explainable techniques [125, 121, 124].

Recent contributions include *Lin et al.* [94], ensuring fairness across demographics, and *Zhu et al.* [99, 92], decomposing frames into 3D components. *SeqFakeFormer++* [106] detects sequential manipulations, while *Song et al.* [101] use quality-centric enhancements. *Peng et al.* [118] assess perceptual fidelity, and *Ba et al.* [112] employ information bottleneck principles. *Dong et al.* [115] mitigate identity leakage, and *Nguyen et al.* [96] focus on localized artifacts. Challenges remain in handling low-quality videos and unseen forgeries.

Audio Deepfakes: Audio detection targets synthetic speech and songs, evolving from artifact-based methods [126] to robust techniques leveraging *non-verbal cues* like breathing [129, 136] and *temporal dependencies* via transformers [132, 133]. Generalization is improved through continual learning and domain generalization [127, 128, 134, 141, 137], while robustness to perturbations employs augmentation [139, 140, 144]. New methods include *HM-Conformer* [131] for hierarchical feature extraction, *TDVSA-Net* [108] for lip-based authentication, and *DeMamba* [107] for spatio-temporal analysis. *Yang et al.* [135] fuse multi-view features, and *Oiso et al.* [138] optimize prompt tuning. *SafeEar* [143] ensures privacy, while *Klein et al.* [130] trace sources. *CtrSVDD* [154] targets singing voices, and *Weizman et al.* [155] enhance ASV systems. *Xie et al.* [142] address ALM-based audio, and *Combei et al.* [145] use ensemble learning. Adversarial robustness and cross-domain performance remain underexplored.

Text Deepfakes: Text detection targets synthetic content like fake news, advancing from *linguistic feature extraction* [146, 47] to *contextual analysis* [147, 149] and *domain-specific detection* [151, 148]. Generalization employs multi-task and zero-shot learning [148, 153], with robustness enhanced by adversarially fine-tuned models [150, 153, 149]. New methods like *Mc-DNN* [146] process multi-channel text, *OBi-LSTM-CNN* [147] optimize rumor detection, and *MAGE* [152] tackle diverse LLMs. Challenges include adapting to evolving LLMs and resisting adversarial text modifications.

Conclusion: Uni-modal deepfake detection has advanced considerably across modalities. Image-based methods leverage frequency analysis and self-supervised learning, video detection emphasizes temporal and graph-based modeling, audio techniques exploit non-verbal cues and transformers, and text detection refines contextual and domain-specific approaches. Despite progress, challenges in generalization to unseen sources, robustness against adversarial attacks, and cross-modal adaptability underscore the need for continued research.

4.1.2 Multi-Modal Deepfake Detection Mechanisms

The evolution of generative AI has escalated deepfake synthesis beyond uni-modal alterations—such as facial manipulations in static images or video frames—to intricate multi-modal fabrications integrating visual, auditory, and textual elements. Uni-modal detection strategies, while proficient in isolating modality-specific anomalies (e.g., pixel-level distortions or spectral irregularities), exhibit limited efficacy against cross-modal inconsistencies inherent in advanced generative systems, such as text-driven image synthesis or audio-visual misalignment. Multi-modal detection mechanisms address this deficiency by employing integrated learning frameworks—contrastive representation, modality fusion, and inconsistency modeling—to discern interdependencies and deviations across heterogeneous data streams, demonstrating superior detection accuracy compared to uni-modal counterparts. Illustrative examples include the identification of lip-audio desynchronization [156, 157, 158] and text-visual incongruities [159, 160], which evade uni-modal scrutiny.

A systematic synthesis of recent advancements, as presented in Table 3, delineates seven paradigmatic categories of multi-modal detection, each leveraging distinct learning strategies and modality combinations. Vision-language models (VLM) harness supervised learning to align image-text representations (Img+Txt), with frameworks like, DE-FAKE [169], MM-Det [161], Prompt2Guard [162], AntifakePrompt [160], and Bi-LORA [173, 174] optimizing pre-trained

Table 3: Overview of Multi-Modal Deepfake Detection Mechanisms

Category	References	Modalities	Strengths and Challenges
VLM	[159, 161, 162, 163, 160, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174]	Image+Text	Interpretable detection, generalizable across generators; limited by prompt dependency, high-fidelity fakes
AV Sync	[156, 157, 158, 175, 176, 177, 178, 179, 180, 181, 182]	Audio+Video	Robust to natural perturbations, sync-focused; struggles with non-frontal faces, unsynchronized inputs
MV Fusion	[183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194]	Aud+Vid+Txt	Broad modality coverage, strong transferability; computationally intensive, alignment issues
SS Learning	[158, 176, 177, 187, 195]	Audio+Video	Effective with limited fake data; sync-dependent, lacks adversarial robustness
ZS Approaches	[179, 181, 196, 171, 166]	Aud+Vid+Txt	Adapts to unseen methods, minimal training; challenged by high-fidelity fakes, semantic reliance
STF Analysis	[197, 198, 199, 200, 201]	Spat+Temp+Freq	Captures subtle artifacts; high computational cost, latency concerns
Exp Detection	[163, 201, 164, 166, 167]	Img+Txt+Vid	Enhances interpretability; limited reasoning depth, dataset-specific
Datasets by Modality:			
Image+Text	D3, CDDB-Hard, FakeClass, FakeClue, FakeQA, FF++, Celeb-DF, WildDeepfake, DGM4, StyleGAN2, Latent Diffusion, COCO, Flickr, SD2, SD3, SDXL, DALL-E, ProGAN, Twitter		
Audio+Video	FF++, DFDC, FakeAVCeleb, AVLips, DeepfakeTIMIT, LRS2, LRS3, KoDF, pDFDC, FaceForensics++, AV-Deepfake1M, LAV-DF, VidTIMIT		
Aud+Vid+Txt	FakeAVCeleb, DFDC, ASVS2015, ASVS2021LA, ASVS2021DF, MUSIC-21, DF-TIMIT, FakeOrReal, InTheWild, DefakeAVMiT, RAFV		
Spat+Temp+Freq	FF++, Celeb-DF, WildDeepfake, DFD, DFDCP, DF-v1.0, CDF1, CDF2, Celeb-DFv2, DFW		

Note: Abbreviations: VLM = Vision-Language Models, AV Sync = Audio-Visual Sync, MV Fusion = Multi-View Fusion, SS Learning = Self-Supervised Learning, ZS Approaches = Zero-Shot Approaches, STF Analysis = Spatial-Temporal-Frequency Analysis, Exp Detection = Explainable Detection; Aud+Vid+Txt = Audio + Video + Text, Spat+Temp+Freq = Spatial + Temporal + Frequency. Datasets are grouped by modality combinations, reflecting common evaluation contexts.

architectures (e.g., CLIP) for cross-modal feature extraction, achieving robust generalization across diffusion-based datasets. Audio-visual synchronization (AV Sync) employs supervised temporal modeling, exemplified by AMSDF [183], LipFD [157], AVSecure [185] and AV-MAE [182], which integrate Aud+Vid signals via correlation analysis or watermarking to mitigate natural perturbations. Multi-view fusion (MV Fusion) consolidates audio, video, and text (Aud+Vid+Txt) through supervised transformer-based architectures, as in TMI-Former [184], FakeSTormer [200], and AVT²-DWF [191], enhancing cross-dataset transferability (e.g., DFDC, FakeAVCeleb) by modeling multi-dimensional feature interactions.

Complementary paradigms extend this methodological diversity. Spatial-temporal-frequency (STF) analysis, including SFDG [198], 3D ConvNet [197], and DBNet [199], utilizes supervised dynamic graph or 3D-temporal learning to capture Spat+Temp+Freq inconsistencies, excelling at subtle artifact detection. Self-supervised learning (SS Learning) frameworks, such as SpeechForensics [176], Feng et al.'s anomaly detection [177], AVFF [187] and AV-HuBERT [195], leverage Aud+Vid real-data distributions with contrastive or similarity-based objectives, offering efficacy in low-resource settings like FakeAVCeleb. Zero-shot (ZS) approaches, exemplified by CCFD [181], FACTOR [196], and Jia et al.'s LLM integration [171], exploit pre-trained models or intrinsic Aud+Vid+Txt consistency for adaptability to novel manipulations sans retraining. Explainable detection (Exp Detection) methods, such as FakeBench [163], DD-VQA [164], and HAMMER [167], employ supervised or zero-shot reasoning over Img+Txt+Vid to elucidate forgery mechanisms, validated across FF++ and Celeb-DF benchmarks (Table 3).

These advancements signify a paradigm shift toward holistic detection, yet critical challenges temper their potential, as cataloged in Table 3. Modality alignment remains problematic under unsynchronized conditions [190] or non-canonical perspectives [186, 201], despite robust performance against natural distortions [185, 179]. Adversarial robustness is underexplored, with notable exceptions like AVA-CL [178] and MSOC [192] employing contrastive or one-class strategies to counter occlusions. Computational overhead restricts real-time applicability in resource-intensive

frameworks [198, 186], while high-fidelity synthetics challenge ZS methods [181]. Although generalizability is evident in cross-modal regularization [175], transformer fusion [180], and knowledge distillation [193], linguistic diversity [157] and subtle manipulations [167] expose persistent gaps.

This multi-modal transition reflects the escalating sophistication of synthetic media, necessitating a strategic research agenda. Enhancing adversarial resilience through targeted training [200], optimizing computational efficiency via lightweight designs [183], and exploring nuanced forensic cues—e.g., gaze dynamics [201] or micro-expressions—represent critical imperatives. Concurrently, bolstering interpretability [164] and scalability [196] will bridge the gap between theoretical innovation and practical deployment, ensuring resilience against an evolving threat landscape. Table 3 encapsulates this state-of-the-art synthesis, delineating mechanisms, modalities, and research frontiers in multi-modal deepfake detection.

5 Deepfake Edited Regions Detection and localization

Detecting and localizing edited regions in digital content is a critical task in deepfake analysis, necessitating methodologies that precisely delineate manipulated areas across images, videos, and audio. This section systematically reviews the technical approaches developed to address this challenge, tracing the progression from traditional forensic techniques to advanced deep learning frameworks, with an emphasis on the core strategies employed for identifying tampered regions.

Initial efforts leveraged traditional forensic techniques, focusing on statistical anomalies such as noise inconsistencies and compression artifacts. Methods like those in [202, 203, 204, 205] utilized noise fingerprints and spatial rich models to detect low-level tampering traces in static images, while [206] explored counter-forensic impacts of neural compression using convolutional feature extraction. These approaches established a baseline but struggled with scalability against complex manipulations, driving the adoption of deep learning solutions.

Deep learning introduced supervised convolutional neural networks (CNNs) for localization, with hierarchical frameworks enhancing precision. [207] employed multi-branch feature extractors and localization modules for fine-grained tamper detection, while [202] integrated noise-sensitive fingerprints with RGB features to generate anomaly maps. Vision Transformers (ViTs) advanced this further: [208] combined windowed ViTs with multi-scale feature extraction, and [209] utilized sparse self-attention to target non-semantic artifacts. Lightweight strategies, such as [210], adopted state space models for efficient multi-scale analysis, improving scalability.

Noise-guided methodologies became prominent for exposing subtle edits. [211] fused denoising networks with cross-attention filters, while [98] merged ViT and CNN branches to capture local noise cues for inpainting detection. Contrastive learning approaches, like [203] and [212], applied multi-scale feature fusion and pixel-level contrast to isolate tampered regions, and [213] explored weakly-supervised localization via patch-based scoring with an Xception backbone. Specialized techniques included [214], using frequency-domain inter-intra similarity modules, and [215], employing reverse edge-attention for inpainting boundary refinement.

Proactive forensics introduced watermarking strategies. [216] and [217] embedded dual watermarks with invertible networks and adaptive transforms, while [218] mined inconsistencies through progressive feature refinement. [219] utilized a dual-stream architecture with coarse-to-fine localization, and [220] applied non-mutually exclusive contrastive learning to tackle data scarcity.

Multi-modal image-based methods integrated diverse cues for enhanced localization. [222] fused RGB and high-frequency features with object prototypes, [223] employed domain-guided visual-textual analysis, and [224] adapted CLIP with noise-assisted prompts. [225] targeted text manipulation with transformed domain features, [204] combined forensic filters via early fusion, and [205] merged noise and spatial features hierarchically. [226] introduced a multimodal VQA framework, [227] used adaptive perception with auto-annotation, and [228] orchestrated mesoscopic features with a CNN-Transformer hybrid.

Video-based localization extended these principles to temporal analysis. [229] employed spatio-temporal transformers, [230] applied co-attention fusion across frame streams, and [244] integrated temporal feature attention. [231] utilized masked autoencoders with meta-learning, [232] leveraged synthetic self-blending, and [233] adapted SAM with multiscale adapters. [234] combined ViT and timeseries transformers, and [235] fused dual-stream modalities for verification.

Multi-modal video approaches incorporated audio-visual integration. [191] utilized transformer-based dynamic fusion, [236] targeted clip-level localization, and [237] assessed anomalies via VQA. [238] combined multiscale ViTs, [239, 240] fused cross-modal interactions, and [241, 242] applied contextual attention with recurrent units. [243] embedded watermarks for dual-modal tamper detection.

Table 4: Overview of Edited Regions Detection and Localization Mechanisms

Category	References	Modalities	Strengths and Challenges
Traditional Forensic	[202, 203, 204, 205, 206]	Image	Detects statistical anomalies with noise and compression features; limited scalability to generative forgeries
Uni-modal Image DL	[207, 208, 209, 210, 211, 98, 212, 213, 214, 215, 220, 221]	Image	Enables precise localization via hierarchical, noise-guided, and contrastive techniques; faces high computational demands and poor generalization to novel manipulations
Proactive Image DL	[216, 217, 218, 219]	Image	Provides robust preemptive detection via watermarking; constrained by sensitivity to degradation and training data variability
Multi-modal Image DL	[222, 223, 224, 225, 226, 227, 228]	Img+Txt	Leverages diverse cues for enhanced localization; challenged by data quality issues and increased inference complexity
Uni-modal Video DL	[229, 230, 231, 232, 233, 234, 235]	Video	Captures temporal artifacts with spatial-temporal models; limited by short sequence processing and sparse annotations
Multi-modal Video DL	[191, 236, 237, 238, 239, 240, 241, 242, 243, 244]	Vid+Aud	Achieves strong synergy across audio-visual modalities; hindered by alignment difficulties and focus on facial regions
Audio DL	[245, 246, 247, 248]	Audio	Offers precise temporal forgery detection; restricted by limited adaptability and processing overhead
Hybrid DL	[249, 250]	Img+Vid	Integrates flexible features across static and frame analysis; confined to specific domains with reduced scalability
Datasets by Modality:			
Image	CASIA, NIST16, Columbia, Coverage, HiFi-IFDL, CelebA, FFHQ, COCO, Places365, Dolos, WildRF, CollabDif, DMID, IID-74K, DEFACTO		
Image+Text	CASIA, NIST16, Columbia, Coverage, MMTD, RTM, FF++, Multi-attack, FFA-VQA		
Video	FF++, Celeb-DF, DFDC, GRIP, VideoSham, HTVD, FaceForensics++, DFD, FMLD, DeepfakeTIMIT, SORA, Vimeo-90K, Davis		
Video+Audio	DFDC, FakeAVCeleb, LAV-DF, AV-Deepfake1M, DeepfakeTIMIT, ForgeryNet, Psynd, TVIL		
Audio	ADD2023, PartialSpoof, LAV-DF, ASVspoof, VoxPopuli, LibriSpeech, Expresso, Half-truth		

Note: Abbreviations: DL = Deep Learning, Img = Image, Vid = Video, Aud = Audio, Txt = Text. References correspond to mechanisms in Section 5. Datasets reflect evaluation contexts across modalities.

Audio-based methods addressed temporal forgery detection. [245] employed adversarial training with Wav2Vec, [246] refined proposals with coarse-to-fine learning, and [248] embedded imperceptible watermarks. [247] assessed countermeasures using multi-resolution feature extraction. Hybrid approaches included [250], localizing video frame edits with spectral analysis, [249], fusing hierarchical features for image deepfakes, and [221], decoding CLIP embeddings for tampering detection.

Across the spectrum of methodologies reviewed—from traditional forensic techniques to advanced deep learning frameworks spanning static images, videos, and audio—the proposed approaches demonstrate notable strengths in detecting and localizing edited regions within digital content. Traditional forensic methods excel at identifying statistical anomalies, while deep learning variants leverage hierarchical feature extraction, noise-guided analysis, watermarking, and multi-modal cue integration to achieve precise tamper delineation. These techniques collectively enhance the ability to pinpoint manipulations with increasing sophistication, adapting to the evolving complexity of deepfake technologies. However, a pervasive limitation emerges in their collective evaluation: an almost universal absence of rigorous testing against adversarial robustness, as summarized in Table 4. Despite their efficacy under controlled conditions, the susceptibility of these methods to adversarial perturbations—such as subtle input modifications designed

to evade detection—remains largely unaddressed, as evidenced by the consistent omission of such assessments across the referenced works. This gap is particularly critical in deepfake analysis, where adversarial attacks could exploit vulnerabilities in feature extraction or model decision boundaries, rendering localization unreliable in real-world scenarios. The absence of adversarial evaluation underscores a significant challenge to the practical deployment of these methods, highlighting the urgent need for future research to incorporate robust adversarial testing to ensure resilience against malicious countermeasures, thereby strengthening the integrity of deepfake detection and localization frameworks.

6 Open Challenges and Future Directions

Despite notable advancements in deepfake detection, the field continues to grapple with several unresolved challenges that impede the creation of robust and reliable systems, particularly as generative technologies and adversarial threats evolve rapidly. These challenges encompass generalizability to emerging generative models, robustness against adversarial and natural perturbations, computational efficiency for real-time deployment, effective multi-modal integration, precise localization of subtle manipulations, and the availability of diverse datasets. Among these, this review identifies the evaluation of adversarial robustness as a critical yet underexplored gap within the existing literature. This section first outlines these broad challenges, then delves into the specific issue of adversarial robustness—examining key attack strategies, their targeted modalities, and the detection models they undermine—and concludes by proposing future directions to address these limitations, emphasizing adaptive frameworks and cross-modal knowledge sharing.

Broad Challenges in Deepfake Detection

The relentless advancement of generative models, such as diffusion-based architectures and transformer-based language models, poses significant hurdles for detection systems, which often struggle to generalize beyond the specific artifact patterns they were trained to recognize. This challenge is compounded by the need to maintain performance under natural perturbations—such as compression, environmental noise, lighting variations, or linguistic noise—that can obscure manipulation cues across image, video, audio, and text modalities. Computational efficiency remains a critical constraint, particularly for real-time applications where resource limitations demand lightweight yet effective models. The integration of multi-modal data—spanning image, video, audio, and text—introduces additional complexity, as aligning features across disparate domains requires overcoming modality-specific noise and inconsistencies. Precise localization of subtle manipulations, such as minor facial distortions, imperceptible audio splicing, or contextually inconsistent text, is hindered by insufficient labeled data, while the scarcity of diverse datasets capturing a wide range of manipulation techniques and real-world conditions restricts model training and evaluation.

These challenges vary across media types. For image-based deepfakes, detection systems must address high-resolution details and compression artifacts. Video-based systems face temporal inconsistencies and motion-based anomalies. Audio deepfakes present unique difficulties, such as identifying unnatural speech patterns or splicing in noisy environments, while text-based deepfakes require detecting subtle linguistic manipulations, like contextually inappropriate phrasing or adversarial perturbations. Recognizing these modality-specific issues is essential for developing comprehensive detection frameworks capable of addressing the multifaceted nature of synthetic media.

Adversarial Robustness: A Critical Yet Missing Evaluation

Adversarial robustness is a cornerstone of reliable deepfake detection, yet many proposed models lack thorough assessments against adversarial perturbations, leaving them vulnerable to an increasingly sophisticated array of threats. This deficiency exposes systems to attacks that exploit weaknesses across modalities—image, video, audio, and text—compromising a diverse spectrum of detection frameworks, from convolutional neural networks (CNNs) and frequency-domain analyzers to watermark-based and transformer-based systems. The following adversarial attack strategies, drawn from recent literature, highlight these vulnerabilities:

- **Image and Video Attacks**:
  - *Pixel-Space Attacks*: Subtle alterations, such as blur, noise, or exposure adjustments [251, 252], and natural shadow overlays [253], evade spatial detectors (e.g., ResNet50, EfficientNet-b4) by aligning statistical distributions or masking differences [254].
  - *Black-Box Perturbations*: Attacks targeting salient facial regions using Natural Evolutionary Strategies (NES) [74, 255], super-resolution techniques [256], or diffusion-based purification [257] challenge models like MesoNet, XceptionNet, and Swin-Small, achieving imperceptible yet potent degradation.

  - *Frequency-Domain Attacks*: Spectral manipulations [82], 2D convolutional filters [258], and frequency-based Bayesian perturbations [259] degrade detectors like DCT and FrequencyForensics, exploiting spectral inconsistencies with high transferability.
  - *Latent-Space Attacks*: Perturbations in generative model representations, such as StyleGAN2's latent space [260], customized Stable Diffusion outputs [82], or diffusion-based latent optimization [261, 262, 263, 264], bypass DNN-based and commercial detectors (e.g., Baidu, Tencent) with remarkable success rates.
  - *Backdoor Attacks*: Poisoned training data with embedded triggers [265, 266] or diffusion process manipulations [265] compromise models like WideResNet and DeiT-S, activating malicious behavior during inference.
  - *Watermarking Attacks*: Diffusion purification [257], universal spectral-domain attacks [267], and adversarial watermark fine-tuning [268] undermine watermark-based detectors (e.g., StegaStamp, RivaGAN), significantly reducing detection efficacy.
- **Audio Attacks**: Adversarial noise injection or splicing [266], white-box attacks like FGSM and PGD [269], and GAN-based transferable attacks [270] exploit temporal inconsistencies, targeting neural network detectors (e.g., LCNN, RawNet3) and end-to-end models (e.g., RawNet2, Res-TSSDNet), with detection accuracies dropping sharply (e.g., 98% to 26%).
- **Text Attacks**: Adversarial perturbations, such as synonym substitution or grammatical alterations [266], and low-cost attacks like decoding strategy shifts and DFTFooler [153], mislead text-based detectors (e.g., GROVER, BERT-Defense) by preserving semantic meaning while altering syntactic or statistical features, achieving high evasion rates (up to 91.3%).

These strategies collectively underscore systemic frailties across detection paradigms, emphasizing the urgent need for comprehensive robustness evaluations that encompass the full spectrum of modalities and adversarial threats identified in current research.

Future Directions: Adaptive Frameworks and Cross-Modal Knowledge Sharing

To address these challenges, particularly the critical gap in adversarial robustness, future research must prioritize the development of adaptive detection frameworks that leverage cross-modal knowledge sharing to enhance resilience across image, video, audio, and text modalities. Drawing on recent advancements, such frameworks could enable detectors to dynamically adjust their focus based on input characteristics and task demands, improving their ability to identify subtle manipulations and withstand diverse adversarial perturbations.

A key strategy involves designing modality-agnostic architectures that integrate and transfer knowledge across media types. For instance, image-based detectors could incorporate audio cues to detect lip-sync inconsistencies, while text-based systems might verify narrative consistency with visual elements, mitigating modality-specific attacks by diversifying feature representations [271, 272, 273]. Expert-driven architectures, with specialized modules targeting distinct manipulation types—such as facial distortions in video, voice cloning in audio, or syntactic anomalies in text—could be dynamically weighted based on input reliability, enhancing sensitivity to cross-modal anomalies [274].

To counter the rapid evolution of deepfake technologies, parameter-efficient adaptation techniques, such as lightweight projectors or adapters, could facilitate rapid recalibration to emerging manipulation methods without extensive retraining. Unsupervised or semi-supervised learning approaches could reduce reliance on scarce labeled datasets, leveraging abundant unlabeled data to improve robustness. Explainable AI techniques could enhance localization capabilities, providing interpretable insights into detected manipulations and informing targeted countermeasures [275].

Standardized benchmarks encompassing a broad spectrum of adversarial scenarios—spanning pixel-space, frequency-domain, latent-space, backdoor, and watermarking attacks across all modalities—are essential for consistent and rigorous robustness assessments. Datasets enriched with adversarial perturbations, reflecting real-world conditions, will further support model generalization. Lightweight models balancing computational efficiency with resilience will ensure practical deployment in resource-constrained environments.

Finally, the societal implications of deepfake detection failures—such as misinformation, fraud, or privacy breaches—underscore the need for robust systems. Future efforts should integrate ethical considerations, developing transparent and accountable frameworks to safeguard public trust and security.

In summary, while deepfake detection faces a multifaceted array of challenges, the insufficient evaluation of adversarial robustness remains a pivotal shortfall. By adopting adaptive frameworks with cross-modal knowledge sharing, embedding adversarial training, and establishing standardized benchmarks, the field can advance toward reliable, scalable detection systems capable of countering the dynamic and escalating threats posed by synthetic media.

7 Conclusion

The escalating sophistication of Generative Artificial Intelligence has amplified the deepfake threat across image, video, audio, and text modalities, challenging the integrity of digital systems and societal trust. This systematic review has elucidated the strengths and limitations of contemporary deepfake detection methodologies, spanning uni-modal and multi-modal frameworks adept at identifying fully synthetic media and localizing subtle manipulations within authentic content. While these approaches demonstrate commendable precision in controlled settings, their vulnerability to adversarial perturbations and limited generalizability to emerging generative techniques underscore a critical gap in real-world applicability.

Our analysis highlights the urgent need for robust, adaptable detection systems capable of withstanding the evolving landscape of synthetic media threats. By prioritizing reproducibility through a curated repository of open-source implementations, this study fosters transparency and enables rigorous validation of current methods. The integration of multi-modal cues emerges as a promising avenue, yet the pervasive shortfall in adversarial robustness demands innovative solutions beyond traditional paradigms. Future advancements must focus on scalable, modality-agnostic architectures that enhance resilience, alongside standardized evaluation protocols to ensure consistent performance across diverse scenarios.

In synthesizing these insights, this review not only consolidates the current state of deepfake detection but also delineates a strategic path forward. The development of trustworthy systems—capable of mitigating misinformation, safeguarding privacy, and maintaining digital security—hinges on addressing these identified challenges with a concerted emphasis on robustness and adaptability. This study thus provides a foundational framework for advancing next-generation detection capabilities, poised to meet the complexities of an increasingly synthetic digital era.

References

[1] T. Sakirin and S. Kusuma, "A survey of generative artificial intelligence techniques," *Babylonian Journal of Artificial Intelligence*, vol. 2023, pp. 10–14, 2023.

[2] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt," *arXiv preprint arXiv:2303.04226*, 2023.

[3] S. C. Tan, W. Chen, and B. L. Chua, "Leveraging generative artificial intelligence based on large language models for collaborative learning," *Learning: Research and Practice*, vol. 9, no. 2, pp. 125–134, 2023.

[4] J. Liu, S. Li, and Q. Dong, "Collaboration with generative artificial intelligence: An exploratory study based on learning analytics," *Journal of Educational Computing Research*, p. 07356331241242441, 2024.

[5] S. Ali, P. Ravi, R. Williams, D. DiPaola, and C. Breazeal, "Constructing dreams using generative ai," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 21, 2024, pp. 23 268–23 275.

[6] P. Gupta, B. Ding, C. Guan, and D. Ding, "Generative ai: A systematic review using topic modelling techniques," *Data and Information Management*, p. 100066, 2024.

[7] J. Bu, R.-L. Jiang, and B. Zheng, "Research on deepfake technology and its application," *Proceedings of the 2023 4th International Conference on Computing, Networks and Internet of Things*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260204654

[8] M. T. Baldassarre, D. Caivano, B. Fernandez Nieto, D. Gigante, and A. Ragone, "The social impact of generative ai: An analysis on chatgpt," in *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, 2023, pp. 363–373.

[9] T. Dorigo, G. D. Brown, C. Casonato, A. Cerdà, J. Ciarrochi, M. Da Lio, N. D'souza, N. R. Gauger, S. C. Hayes, S. G. Hofmann *et al.*, "Artificial intelligence in science and society: the vision of usern," *IEEE Access*, 2025.

[10] E. Simonsson, "Generative ai effects on school systems : An overview of generative ai with focus on chatgpt, what it is, what it isn't and how it works." p. 83, 2023.

[11] D. Humphreys, A. Koay, D. Desmond, and E. Mealy, "Ai hype as a cyber security risk: the moral responsibility of implementing generative ai in business," *AI and Ethics*, pp. 1–14, 2024.

[12] Y. Chen and P. Esmaeilzadeh, "Generative ai in medical practice: In-depth exploration of privacy and security challenges," *Journal of Medical Internet Research*, vol. 26, p. e53008, 2024.

[13] S. Nishal and N. Diakopoulos, "Envisioning the applications and implications of generative ai for news media," *arXiv preprint arXiv:2402.18835*, 2024.

[14] D. M. Popa, "Critical exploratory investigation of ai consumption, ai perception and ai literacy requirements," *AI perception and AI literacy requirements (November 01, 2024)*, 2024.

[15] Dimensions AI, "Dimensions AI - Research insights at your fingertips," n.d., accessed: [date]. [Online]. Available: https://www.dimensions.ai/

[16] C. Whyte, "Beyond "bigger, faster, better"," *The Cyber Defense Review*, vol. 8, no. 3, pp. 135–150, 2023.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[18] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[19] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, "Variational autoencoder for deep learning of images, labels and captions," *Advances in neural information processing systems*, vol. 29, 2016.

[20] Y. Chen, N. A. H. Haldar, N. Akhtar, and A. Mian, "Text-image guided diffusion model for generating deepfake celebrity interactions," *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 348–355, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:262825503

[21] M. Mirza, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[22] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," *IEEE transactions on circuits and systems for video technology*, vol. 30, no. 11, pp. 3943–3956, 2019.

[23] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International conference on machine learning*. PMLR, 2016, pp. 1060–1069.

[24] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[25] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[26] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.

[27] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," *arXiv preprint arXiv:2102.12092*, 2021.

[28] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet *et al.*, "Video diffusion models," *arXiv preprint arXiv:2204.03458*, 2022.

[29] B. M. Le, J. Kim, S. Tariq, K. Moore, A. Abuadbba, and S. S. Woo, "Sok: Facial deepfake detectors," *arXiv preprint arXiv:2401.04364*, 2024.

[30] A. Kaur, A. N. Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, "Deepfake video detection: challenges and opportunities," *Artif. Intell. Rev.*, vol. 57, p. 159, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270127933

[31] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, "Audio deepfake detection: A survey," *ArXiv*, vol. abs/2308.14970, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261276979

[32] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "A survey on speech deepfake detection," *ACM Computing Surveys*, 2025.

[33] M. Q. Alnabhan and P. Branco, "Fake news detection using deep learning: A systematic literature review," *IEEE Access*, 2024.

[34] A. Kumar, D. Singh, R. Jain, D. K. Jain, C. Gan, and X. Zhao, "Advances in deepfake detection algorithms: Exploring fusion techniques in single and multi-modal approach," *Information Fusion*, p. 102993, 2025.

[35] P. Liu, Q. Tao, and J. T. Zhou, "Evolving from single-modal to multi-modal facial deepfake detection: A survey," *ArXiv*, vol. abs/2406.06965, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270379957

[36] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[37] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, Q. V. Le *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Interspeech 2017*, 2017, pp. 4006–4010.

[38] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[39] M. Westerlund, "The emergence of deepfake technology: A review," *Technology Innovation Management Review*, vol. 9, no. 11, pp. 39–52, 2019.

[40] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-generated images are surprisingly easy to spot... for now," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8695–8704.

[41] F. Wang, W. Liu, D. Chen, J. Wu, Z. Ye, S. Shen *et al.*, "Diffusion models for image synthesis: A comprehensive survey," *arXiv preprint arXiv:2302.06825*, 2023.

[42] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, and J. McAuley, "Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3348–3357.

[43] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[44] 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.

[45] 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*. PMLR, 2020, pp. 3247–3258.

[46] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces (GUI)*, vol. 3, no. 1, pp. 80–87, 2019.

[47] A. T. Y. Chong, H. N. Chua, M. B. Jasser, and R. T. K. Wong, "Bot or human? detection of deepfake text with semantic, emoji, sentiment and linguistic features," *2023 IEEE 13th International Conference on System Engineering and Technology (ICSET)*, pp. 205–210, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264808782

[48] U. A. Ciftci, I. Demir, and L. Yin, "Fakecatcher: Detection of synthetic portrait videos using biological signals," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[49] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, "Detecting deep-fake videos from phoneme-viseme mismatches," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 660–661.

[50] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6458–6467.

[51] B. Malolan, A. Parekh, and F. Kazi, "Explainable deep-fake detection using visual interpretability methods," in *2020 3rd International conference on Information and Computer Technologies (ICICT)*. IEEE, 2020, pp. 289–293.

[52] R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang, and Y. Liu, "Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces," in *International Joint Conference on Artificial Intelligence*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:212976079

[53] C. Tan, Y. Zhao, S. Wei, G. Gu, and Y. Wei, "Learning on gradients: Generalized artifacts representation for gan-generated images detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 105–12 114.

[54] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Frequency-aware deepfake detection: Improving generalizability through frequency space learning," *arXiv preprint arXiv:2403.07240*, 2024.

[55] Y. Jeong, D. Kim, Y. Ro, and J. Choi, "Frepgan: Robust deepfake detection using frequency-level perturbations," in *AAAI Conference on Artificial Intelligence*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:246634415

[56] C. T. Doloriel and N.-M. Cheung, "Frequency masking for universal deepfake detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 13 466–13 470.

[57] O. Pontorno, L. Guarnera, and S. Battiato, "On the exploitation of dct-traces in the generative-ai domain," in *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2024, pp. 3806–3812.

[58] C. Tian, Z. Luo, G. Shi, and S. Li, "Frequency-aware attentional feature fusion for deepfake detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[59] B. Liu, B. Liu, M. Ding, and T. Zhu, "Detection of diffusion model-generated faces by assessing smoothness and noise tolerance," in *2024 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 2024, pp. 1–6.

[60] Y. Lu and T. Ebrahimi, "Towards the detection of ai-synthesized human face images," in *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2024, pp. 3778–3784.

[61] D. A. Coccomini, R. Caldelli, C. Gennaro, G. Fiameni, G. Amato, and F. Falchi, "Deepfake detection without deepfakes: Generalization via synthetic frequency patterns injection," *arXiv preprint arXiv:2403.13479*, 2024.

[62] L. B. Baru, R. Boddeda, S. A. Patel, and S. M. Gajapaka, "Wavelet-driven generalizable framework for deepfake face forgery detection," in *Proceedings of the winter conference on applications of computer vision*, 2025, pp. 1661–1669.

[63] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 480–24 489.

[64] S. Tang, P. He, H. Li, W. Wang, X. Jiang, and Y. Zhao, "Towards extensible detection of ai-generated images via content-agnostic adapter-based category-aware incremental learning," *IEEE Transactions on Information Forensics and Security*, 2025.

[65] H. Liu, Z. Tan, C. Tan, Y. Wei, J. Wang, and Y. Zhao, "Forgery-aware adaptive transformer for generalizable synthetic image detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 770–10 780.

[66] Q. Xu, D. Chen, J. Chen, S. Lyu, and C. Wang, "Recent advances on generalizable diffusion-generated image detection," *arXiv preprint arXiv:2502.19716*, 2025.

[67] R. Lanzino, F. Fontana, A. Diko, M. R. Marini, and L. Cinque, "Faster than lies: Real-time deepfake detection using binary neural networks," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3771–3780, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270357961

[68] Y. Lim, C. Lee, A. Kim, and O. Etzioni, "Distildire: A small, fast, cheap and lightweight diffusion synthesized deepfake detection," *ArXiv*, vol. abs/2406.00856, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270210472

[69] X. Wu, X. Liao, and B. Ou, "Sepmark: Deep separable watermarking for unified source tracing and deepfake detection," *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258588430

[70] L.-Y. Hsu, "Ai-assisted deepfake detection using adaptive blind image watermarking," *J. Vis. Commun. Image Represent.*, vol. 100, p. 104094, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267907143

[71] N. Lukas and F. Kerschbaum, "{PTW}: Pivotal tuning watermarking for {Pre-Trained} image generators," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 2241–2258.

[72] T. Wang, M. Huang, H. Cheng, X. Zhang, and Z. Shen, "Lampmark: Proactive deepfake detection via training-free landmark perceptual watermarks," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 10 515–10 524.

[73] T. Qiao, B. Zhao, R. Shi, M. Han, M. Hassaballah, F. Retraint, and X. Luo, "Scalable universal adversarial watermark defending against facial forgery," *IEEE Transactions on Information Forensics and Security*, 2024.

[74] B. Gowrisankar and V. L. Thing, "An adversarial attack approach for explainable ai evaluation on deepfake detection models," *Computers & Security*, vol. 139, p. 103684, 2024.

[75] O. Pontorno, L. Guarnera, and S. Battiato, "Deepfeaturex net: Deep features extractors based network for discriminating synthetic from real images," *ArXiv*, vol. abs/2404.15697, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269330225

[76] A. Aghasanli, D. Kangin, and P. P. Angelov, "Interpretable-through-prototypes deepfake detection for diffusion models," *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 467–474, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263771394

[77] C. Tan, P. Liu, R. Tao, H. Liu, Y. Zhao, B. Wu, and Y. Wei, "Data-independent operator: A training-free artifact representation extractor for generalizable deepfake detection," *ArXiv*, vol. abs/2403.06803, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268363480

[78] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 28 130–28 139.

[79] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "Dire for diffusion-generated image detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 445–22 455.

[80] L. Guarnera, O. Giudice, and S. Battiato, "Mastering deepfake detection: A cutting-edge approach to distinguish gan and diffusion-model images," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268302141

[81] B. N. Subudhi *et al.*, "Adaptive meta-learning for robust deepfake detection: A multi-agent framework to data drift and model generalization," *arXiv preprint arXiv:2411.08148*, 2024.

[82] S. M. Abdullah, A. Cheruvu, S. Kanchi, T. Chung, P. Gao, M. Jadliwala, and B. Viswanath, "An analysis of recent advances in deepfake image detection in an evolving threat landscape," *arXiv preprint arXiv:2404.16212*, 2024.

[83] T. Chen, S. Yang, S. Hu, Z. Fang, Y. Fu, X. Wu, and X. Wang, "Masked conditional diffusion model for enhancing deepfake detection," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–7.

[84] J. Li, W. Jiang, L. Shen, and Y. Ren, "Optimized frequency collaborative strategy drives ai image detection," *IEEE Internet of Things Journal*, 2025.

[85] C. T. C. Doloriel and N.-M. Cheung, "Frequency masking for universal deepfake detection," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13 466–13 470, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:266977102

[86] O. Pontorno, L. Guarnera, and S. Battiato, "On the exploitation of dct-traces in the generative-ai domain," *ArXiv*, vol. abs/2402.02209, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267411952

[87] Y. Lu and T. Ebrahimi, "Towards the detection of ai-synthesized human face images," *ArXiv*, vol. abs/2402.08750, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267657939

[88] D. A. Coccomini, R. Caldelli, C. Gennaro, G. Fiameni, G. Amato, and F. Falchi, "Deepfake detection without deepfakes: Generalization via synthetic frequency patterns injection," *ArXiv*, vol. abs/2403.13479, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268537294

[89] B. Liu, B. Liu, M. Ding, and T. Zhu, "Detection of diffusion model-generated faces by assessing smoothness and noise tolerance," *2024 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–6, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271574103

[90] C. Zheng, C. Lin, Z. Zhao, H. Wang, X. Guo, S. Liu, and C. Shen, "Breaking semantic artifacts for generalized ai-generated image detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 59 570–59 596, 2025.

[91] R. R. Dhanakshirur *et al.*, "Herd mentality in augmentation–not a good idea! a robust multi-stage approach towards deepfake detection," *arXiv preprint arXiv:2410.05466*, 2024.

[92] X. Zhu, H. Fei, B. Zhang, T. Zhang, X. Zhang, S. Z. Li, and Z. Lei, "Face forgery detection by 3d decomposition and composition search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8342–8357, 2023.

[93] K. Xu, X. Hu, X. Zhou, X. Xu, L. Qi, and C. Chen, "Rlgc: Reconstruction learning fusing gradient and content features for efficient deepfake detection," *IEEE Transactions on Consumer Electronics*, 2024.

[94] L. Lin, X. He, Y. Ju, X. Wang, F. Ding, and S. Hu, "Preserving fairness generalization in deepfake detection," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16 815–16 825, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268031845

[95] H. She, Y. Hu, B. Liu, J. Li, and C.-T. Li, "Using graph neural networks to improve generalization capability of the models for deepfake detection," *IEEE Transactions on Information Forensics and Security*, 2024.

[96] D. Nguyen, N. Mejri, I. P. Singh, P. Kuleshova, M. Astrid, A. Kacem, E. Ghorbel, and D. Aouada, "Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17 395–17 405, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270045455

[97] R. Shao, T. Wu, and Z. Liu, "Detecting and recovering sequential deepfake manipulation," in *European Conference on Computer Vision*. Springer, 2022, pp. 712–728.

[98] L. Zhang, X. Zhu, D. He, X. Liao, and B. Sun, "Samif: Adapting segment anything model for image inpainting forensics," in *Proceedings of the Asian Conference on Computer Vision*, 2024, pp. 3605–3621.

[99] X. Zhu, H. Wang, H. Fei, Z. Lei, and S. Z. Li, "Face forgery detection by 3d decomposition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2929–2939.

[100] R. Zhang, P. He, H. Li, S. Wang, and Y. Cao, "Temporal diversified self-contrastive learning for generalized face forgery detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[101] W. Song, Z. Yan, Y. Lin, T. Yao, C. Chen, S. Chen, Y. Zhao, S. Ding, and B. Li, "A quality-centric framework for generic deepfake detection," *arXiv preprint arXiv:2411.05335*, 2024.

[102] A. A. Hasanaath, H. Luqman, R. Katib, and S. Anwar, "Fsbi: Deepfake detection with frequency enhanced self-blended images," *Image and Vision Computing*, p. 105418, 2025.

[103] M. D. Bah and M. Dahmane, "Enhanced deepfake detection using frequency domain upsampling," in *VISIGRAPP : VISAPP*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268236242

[104] S. K. Datta, S. Jia, and S. Lyu, "Exposing lip-syncing deepfakes from mouth inconsistencies," *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267035087

[105] Y. Xu, J. Liang, L. Sheng, and X.-Y. Zhang, "Learning spatiotemporal inconsistency via thumbnail layout for face deepfake detection," *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5663–5680, 2024.

[106] R. Shao, T. Wu, and Z. Liu, "Robust sequential deepfake detection," *International Journal of Computer Vision*, pp. 1–18, 2025.

[107] H. Chen, Y. Hong, Z. Huang, Z. Xu, Z. Gu, Y. Li, J. Lan, H. Zhu, J. Zhang, W. Wang *et al.*, "Demamba: Ai-generated video detection on million-scale genvideo benchmark," *arXiv preprint arXiv:2405.19707*, 2024.

[108] Y. He, L. Yang, S. Wang, and A. W.-C. Liew, "Lip feature disentanglement for visual speaker authentication in natural scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[109] Z. Yang, J. Liang, Y. Xu, X.-Y. Zhang, and R. He, "Masked relation learning for deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1696–1708, 2023.

[110] H. Xie, H. He, B. Fu, and V. Sanchez, "Grdt: Towards robust deepfake detection using geometric representation distribution and texture," in *Proceedings of the Winter Conference on Applications of Computer Vision*, 2025, pp. 734–744.

[111] Z. Sun, S. Chen, T. Yao, B. Yin, R. Yi, S. Ding, and L. Ma, "Contrastive pseudo learning for open-world deepfake attribution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 882–20 892.

[112] Z. Ba, Q. Liu, Z. Liu, S. Wu, F. Lin, L. Lu, and K. Ren, "Exposing the deception: Uncovering more forgery clues for deepfake detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 719–728.

[113] Z. Cai, S. Ghosh, K. Stefanov, A. Dhall, J. Cai, H. Rezatofighi, R. Haffari, and M. Hayat, "Marlin: Masked autoencoder for facial video representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1493–1504.

[114] N. Larue, N.-S. Vu, V. Struc, P. Peer, and V. Christophides, "Seeable: Soft discrepancies and bounded contrastive learning for exposing deepfakes," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 21 011–21 021.

[115] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, and Z. Ge, "Implicit identity leakage: The stumbling block to improving deepfake detection generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3994–4004.

[116] C. Kong, A. Luo, P. Bao, H. Li, R. Wan, Z. Zheng, A. Rocha, and A. C. Kot, "Open-set deepfake detection: A parameter-efficient adaptation method with forgery style mixture," *arXiv preprint arXiv:2408.12791*, 2024.

[117] C. Tan, R. Tao, H. Liu, G. Gu, B. Wu, Y. Zhao, and Y. Wei, "C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection," *ArXiv*, vol. abs/2408.09647, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271903069

[118] C. Peng, H. Guo, D. Liu, N. Wang, R. Hu, and X. Gao, "Deepfidelity: Perceptual forgery fidelity assessment for deepfake detection," *arXiv preprint arXiv:2312.04961*, 2023.

[119] R. Shao, T. Wu, L. Nie, and Z. Liu, "Deepfake-adapter: Dual-level adapter for deepfake detection," *International Journal of Computer Vision*, pp. 1–16, 2025.

[120] R. Gong, R. He, D. Zhang, A. K. Sangaiah, and M. J. Alenazi, "Robust face forgery detection integrating local texture and global texture information," *EURASIP Journal on Information Security*, vol. 2025, no. 1, p. 3, 2025.

[121] K. Tsigos, E. Apostolidis, and V. Mezaris, "Improving the perturbation-based explanation of deepfake detectors through the use of adversarially-generated samples," *arXiv preprint arXiv:2502.03957*, 2025.

[122] D. Wodajo, P. Lambert, G. V. Wallendael, S. Atnafu, and H. Mareen, "Improved deepfake video detection using convolutional vision transformer," *2024 IEEE Gaming, Entertainment, and Media Conference (GEM)*, pp. 1–6, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271115309

[123] B. Kaddar, S. A. Fezza, Z. Akhtar, W. Hamidouche, A. Hadid, and J. Serra-Sagristà, "Deepfake detection using spatiotemporal transformer," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 11, pp. 1–21, 2024.

[124] F. Siddiqui, J. Yang, S. Xiao, and M. Fahad, "Enhanced deepfake detection with densenet and cross-vit," *Expert Systems with Applications*, vol. 267, p. 126150, 2025.

[125] F. Mahmud, Y. Abdullah, M. Islam, and T. Aziz, "Unmasking deepfake faces from videos using an explainable cost-sensitive deep learning approach," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2023, pp. 1–6.

[126] T.-H. Shih, C.-Y. Yeh, and M.-S. Chen, "Does audio deepfake detection rely on artifacts?" *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12 446–12 450, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268524800

[127] X. Zhang, J. Yi, C. Wang, C. Zhang, S. Zeng, and J. Tao, "What to remember: Self-adaptive continual learning for audio deepfake detection," in *AAAI Conference on Artificial Intelligence*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266335673

[128] Y. Chen, J. Yi, C. Fan, J. Tao, Y. Ren, S. Zeng, C. Y. Zhang, X. Yan, H. Gu, J. Xue *et al.*, "Region-based optimization in continual learning for audio deepfake detection," *arXiv preprint arXiv:2412.11551*, 2024.

[129] T.-P. Doan, L. Nguyen-Vu, S. Jung, and K. Hong, "Bts-e: Audio deepfake detection using breathing-talking-silence encoder," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[130] N. Klein, T. Chen, H. Tak, R. Casal, and E. Khoury, "Source tracing of audio deepfake systems," *arXiv preprint arXiv:2407.08016*, 2024.

[131] H.-S. Shin, J.-S. Heo, J. ho Kim, C. Lim, W. Kim, and H.-J. Yu, "Hm-conformer: A conformer-based audio deepfake detection system with hierarchical pooling and multi-level classification token aggregation methods," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10 581–10 585, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:262012779

[132] Y. Chen, J. Yi, J. Xue, C. Wang, X. Zhang, S. Dong, S. Zeng, J. Tao, Z. Lv, and C. Fan, "Rawbmamba: End-to-end bidirectional state space model for audio deepfake detection," *ArXiv*, vol. abs/2406.06086, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270371253

[133] M. A. Rahman, Z. I. A. Hakim, N. H. Sarker, B. Paul, and S. A. Fattah, "Sonics: Synthetic or not - identifying counterfeit songs," in *International Conference on Learning Representations (ICLR)*, 2025.

[134] Y. Xie, H. Cheng, Y. Wang, and L. Ye, "Domain generalization via aggregation and separation for audio deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 344–358, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:264190166

[135] Y. Yang, H. Qin, H. Zhou, C. Wang, T. Guo, K. Han, and Y. Wang, "A robust audio deepfake detection system via multi-view feature," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 13 131–13 135.

[136] Y. Zhu, S. Koppisetti, T. Tran, and G. Bharaj, "Slim: Style-linguistics mismatch model for generalized audio deepfake detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 67 901–67 928, 2025.

[137] Q. Zhang, S. Wen, and T. Hu, "Audio deepfake detection with self-supervised xls-r and sls classifier," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 6765–6773.

[138] H. Oiso, Y. Matsunaga, K. Kakizaki, and T. Miyagawa, "Prompt tuning for audio deepfake detection: Computationally efficient test-time domain adaptation with limited target dataset," *arXiv preprint arXiv:2410.09869*, 2024.

[139] G. Lee, J. Lee, M. Jung, J. Lee, K. Hong, S. Jung, and Y. Han, "Dual-channel deepfake audio detection: Leveraging direct and reverberant waveforms," *IEEE Access*, vol. 13, pp. 18 040–18 052, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:275836837

[140] H. Wu, J. Chen, R. Du, C. Wu, K. He, X. Shang, H. Ren, and G. Xu, "Clad: Robust audio deepfake detection against manipulation attacks with contrastive learning," *arXiv preprint arXiv:2404.15854*, 2024.

[141] Y. Li, M. Zhang, M. Ren, M. Ma, D. Wei, and H. Yang, "Cross-domain audio deepfake detection: Dataset and analysis," in *Conference on Empirical Methods in Natural Language Processing*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269004806

[142] Y. Xie, C. Xiong, X. Wang, Z. Wang, Y. Lu, X. Qi, R. Fu, Y. Liu, Z. Wen, J. Tao *et al.*, "Does current deepfake audio detection model effectively detect alm-based deepfake audio?" in *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2024, pp. 481–485.

[143] X. Li, K. Li, Y. Zheng, C. Yan, X. Ji, and W. Xu, "Safeear: Content privacy-preserving audio deepfake detection," in *Conference on Computer and Communications Security*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:272689640

[144] T. M. Wani, R. Gulzar, and I. Amerini, "Abc-capsnet: Attention based cascaded capsule network for audio deepfake detection," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2464–2472, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:272915892

[145] D. Combei, A. Stan, D. Oneata, and H. Cucu, "Wavlm model ensemble for audio deepfake detection," *arXiv preprint arXiv:2408.07414*, 2024.

[146] J. V. Tembhurne, M. M. Almin, and T. Diwan, "Mc-dnn: Fake news detection using multi-channel deep neural networks," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 18, no. 1, pp. 1–20, 2022.

[147] R. Amutha, "Detection on early dynamic rumor influence and propagation using biogeography-based optimization with deep learning approaches," *Multimedia Tools and Applications*, pp. 1–18, 2024.

[148] K. Wang, Y. Yang, and X. Wang, "Spectral clustering-guided news environments perception for fake news detection," *IEEE Access*, 2024.

[149] A. Uchendu, T. Le, and D. Lee, "Topformer: Topology-aware authorship attribution of deepfake texts with diverse writing styles," in *European Conference on Artificial Intelligence*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:262216942

[150] D. H. Lee and B. Jang, "Enhancing machine-generated text detection: Adversarial fine-tuning of pre-trained language models," *IEEE Access*, 2024.

[151] M. Q. Alnabhan and P. Branco, "Bertguard: Two-tiered multi-domain fake news detection with class imbalance mitigation," *Big Data and Cognitive Computing*, vol. 8, no. 8, p. 93, 2024.

[152] Y. Li, Q. Li, L. Cui, W. Bi, Z. Wang, L. Wang, L. Yang, S. Shi, and Y. Zhang, "MAGE: Machine-generated text detection in the wild," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 36–53. [Online]. Available: https://aclanthology.org/2024.acl-long.3

[153] J. Pu, Z. Sarwar, S. M. Abdullah, A. Rehman, Y. Kim, P. Bhattacharya, M. Javed, and B. Viswanath, "Deepfake text detection: Limitations and opportunities," in *Proc. of IEEE S&P*, 2023.

[154] Y. Zang, J. Shi, Y. Zhang, R. Yamamoto, J. Han, Y. Tang, S. Xu, W. Zhao, J. Guo, T. Toda, and Z. Duan, "Ctrsvdd: A benchmark dataset and baseline analysis for controlled singing voice deepfake detection," in *Interspeech 2024*, 2024, pp. 4783–4787.

[155] A. Weizman, Y. Ben-Shimol, and I. Lapidot, "Tandem spoofing-robust automatic speaker verification based on time-domain embeddings," *arXiv preprint arXiv:2412.17133*, 2024.

[156] Y. Zhou and S.-N. Lim, "Joint audio-visual deepfake detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 800–14 809.

[157] W. Liu, T. She, J. Liu, B. Li, D. Yao, Z. Liang, and R. Wang, "Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 91 131–91 155. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/a5a5b0ff87c59172a13342d428b1e033-Paper-Conference.pdf

[158] S. A. Shahzad, A. Hashmi, Y.-T. Peng, Y. Tsao, and H.-M. Wang, "Av-lip-sync+: Leveraging av-hubert to exploit multimodal inconsistency for video deepfake detection," *ArXiv*, vol. abs/2311.02733, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265033179

[159] Santosh, L. Lin, I. Amerini, X. Wang, and S. Hu, "Robust clip-based detector for exposing diffusion model-generated images," *2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–7, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269282710

[160] Y.-M. Chang, C. Yeh, W.-C. Chiu, and N. Yu, "Antifakeprompt: Prompt-tuned vision-language models are fake image detectors," *ArXiv*, vol. abs/2310.17419, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264490490

[161] X. Song, X. Guo, J. Zhang, Q. Li, L. Bai, X. Liu, G. Zhai, and X. Liu, "On learning multi-modal forgery representation for diffusion generated video detection," in *Proceeding of Thirty-eighth Conference on Neural Information Processing Systems*, Vancouver, Canada, December 2024.

[162] F. Laiti, B. Liberatori, T. D. Min, and E. Ricci, "Conditioned prompt-optimization for continual deepfake detection," *ArXiv*, vol. abs/2407.21554, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271571467

[163] Y. Li, X. Liu, X. Wang, B. S. Lee, S. Wang, A. Rocha, and W. Lin, "Fakebench: Probing explainable fake image detection via large multimodal models," *arXiv preprint arXiv:2404.13306*, 2024.

[164] Y. Zhang, B. Colman, A. Shahriyari, and G. Bharaj, "Common sense reasoning for deep fake detection," *ArXiv*, vol. abs/2402.00126, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267365533

[165] Y. Lai, Z. Yu, J. Yang, B. Li, X. Kang, and L. Shen, "Gm-df: Generalized multi-scenario deepfake detection," *arXiv preprint arXiv:2406.20078*, 2024.

[166] N. M. Foteinopoulou, E. Ghorbel, and D. Aouada, "A hitchhiker's guide to fine-grained face forgery detection using common sense reasoning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 2943–2976, 2024.

[167] R. Shao, T. Wu, and Z. Liu, "Detecting and grounding multi-modal media manipulation," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257952257

[168] S. A. Khan and D.-T. Dang-Nguyen, "Clipping the deception: Adapting vision-language models for universal deepfake detection," *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267759986

[169] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "De-fake: Detection and attribution of fake images generated by text-to-image generation models," in *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, 2023, pp. 3418–3432.

[170] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva, "Raising the Bar of AI-generated Image Detection with CLIP," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

[171] S. Jia, R. Lyu, K. Zhao, Y. Chen, Z. Yan, Y. Ju, C. Hu, X. Li, B. Wu, and S. Lyu, "Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4324–4333, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268554169

[172] M. N. Shah and A. Ganatra, "Feature integration-based residual deep learning model for fake news detection using multimodal data sources," in *2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*. IEEE, 2024, pp. 345–353.

[173] M. Keita, W. Hamidouche, H. Bougueffa Eutamene, A. Taleb-Ahmed, D. Camacho, and A. Hadid, "Bi-lora: A vision-language approach for synthetic image detection," *Expert Systems*, vol. 42, no. 2, p. e13829, 2025.

[174] M. Keita, W. Hamidouche, H. Bougueffa, A. Hadid, and A. Taleb-Ahmed, "Harnessing the power of large vision language models for synthetic image detection," *arXiv preprint arXiv:2404.02726*, 2024.

[175] H. Zou, M. Shen, Y. Hu, C. Chen, E. S. Chng, and D. Rajan, "Cross-modality and within-modality regularization for audio-visual deepfake detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 4900–4904.

[176] Y. Liang, M. Yu, G. Li, J. Jiang, B. Li, F. Yu, N. Zhang, X. Meng, and W. Huang, "Speechforensics: Audio-visual speech representation learning for face forgery detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 86 124–86 144, 2024.

[177] C. Feng, Z. Chen, and A. Owens, "Self-supervised video forensics by audio-visual anomaly detection," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 491–10 503, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:255416023

[178] Y. Zhang, W. Lin, and J. Xu, "Joint audio-visual attention with contrastive learning for more general deepfake detection," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 5, pp. 1–23, 2024.

[179] D. Cozzolino, M. Nießner, and L. Verdoliva, "Audio-visual person-of-interest deepfake detection," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 943–952, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248006166

[180] C. Koutlis and S. Papadopoulos, "Dimodif: Discourse modality-information differentiation for audio-visual deepfake detection and localization," *arXiv preprint arXiv:2411.10193*, 2024.

[181] X. Li, Z. Liu, C. Chen, L. Li, L. Guo, and D. Wang, "Zero-shot fake video detection by audio-visual consistency," *arXiv preprint arXiv:2406.07854*, 2024.

[182] S. Mo and P. Morgado, "Unveiling the power of audio-visual early fusion transformers with dense interactions through masked modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 186–27 196.

[183] J. Wu, Q. Yin, Z. Sheng, W. Lu, J. Huang, and B. Li, "Audio multi-view spoofing detection framework based on audio-text-emotion correlations," *IEEE Transactions on Information Forensics and Security*, 2024.

[184] J. Yoon, A. Panizo-LLedot, D. Camacho, and C. Choi, "Triple-modality interaction for deepfake detection on zero-shot identity," *Information Fusion*, vol. 109, p. 102424, 2024.

[185] B. Guo, H. Tai, G. Luo, and Y. Zhu, "Avsecure: An audio-visual watermarking framework for proactive deepfake detection," *2024 IEEE 14th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, pp. 1–4, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270639932

[186] A. Kharel, M. Paranjape, and A. Bera, "Df-transfusion: Multimodal deepfake detection via lip-audio cross-attention and facial self-attention," *arXiv preprint arXiv:2309.06511*, 2023.

[187] T. Oorloff, S. Koppisetti, N. Bonettini, D. Solanki, B. Colman, Y. Yacoob, A. Shahriyari, and G. Bharaj, "Avff: Audio-visual feature fusion for video deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 102–27 112.

[188] V. S. Katamneni and A. Rattani, "Mis-avoidd: Modality invariant and specific representation for audio-visual deepfake detection," *2023 International Conference on Machine Learning and Applications (ICMLA)*, pp. 1371–1378, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263608913

[189] A. A. Bekheet, G. Khoriba, and A. S. Ghoneim, "Development of a multimodal framework for deepfake detection: combining visual and audio analysis," in *Proceedings of the 10th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS 2024)*, 2024.

[190] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao, and K. Ren, "Avoid-df: Audio-visual joint learning for detecting deepfake," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2015–2029, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257797706

[191] R. Wang, D. Ye, L. Tang, Y. Zhang, and J. Deng, "Avt^2-dwf: Improving deepfake detection with audio-visual fusion and dynamic weighting strategies," *IEEE Signal Processing Letters*, vol. 31, pp. 1960–1964, 2024.

[192] K. Lee, Y. Zhang, and Z. Duarr, "A multi-stream fusion approach with one-class learning for audio-visual deepfake detection," in *2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2024, pp. 1–6.

[193] X. Liu, Y. Yu, X. Li, and Y. Zhao, "Mcl: Multimodal contrastive learning for deepfake detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, pp. 2803–2813, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:261629550

[194] A. L. Pellicer, Y. Li, and P. Angelov, "Pudd: towards robust multi-modal prototype-based deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3809–3817.

[195] S. Smeu, D.-A. Boldisor, D. Oneata, and E. Oneata, "Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning," *arXiv preprint arXiv:2412.00175*, 2024.

[196] T. Reiss, B. Cavia, and Y. Hoshen, "Detecting deepfakes without seeing any," *arXiv preprint arXiv:2311.01458*, 2023.

[197] Z. Wang, J. Bao, W. Zhou, W. Wang, and H. Li, "Altfreezing for more general video face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 4129–4138.

[198] Y. Wang, K. Yu, C. Chen, X. Hu, and S. Peng, "Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7278–7287, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259226281

[199] C. Zhou, F. W. Li, C. Song, D. Zheng, and B. Yang, "3d data augmentation and dual-branch model for robust face forgery detection," *Graphical Models*, vol. 138, p. 101255, 2025.

[200] D. Nguyen, M. Astrid, A. Kacem, E. Ghorbel, and D. Aouada, "Vulnerability-aware spatio-temporal learning for generalizable and interpretable deepfake video detection," *arXiv preprint arXiv:2501.01184*, 2025.

[201] C. Peng, Z. Miao, D. Liu, N. Wang, R. Hu, and X. Gao, "Where deepfakes gaze at? spatial-temporal gaze inconsistency analysis for video face forgery detection," *IEEE Transactions on Information Forensics and Security*, 2024.

[202] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, and L. Verdoliva, "Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 20 606–20 615.

[203] R. Bai, "Image manipulation detection and localization using multi-scale contrastive learning," *Applied Soft Computing*, p. 111914, 2024.

[204] K. Triaridis and V. Mezaris, "Exploring multi-modal fusion for image manipulation detection and localization," in *Proc. 30th Int. Conf. on MultiMedia Modeling (MMM 2024)*, Jan.-Feb. 2024.

[205] D. Dagar and D. K. Vishwakarma, "A noise and edge extraction-based dual-branch method for shallowfake and deepfake localization," *Signal, Image and Video Processing*, vol. 19, no. 3, p. 198, 2025.

[206] E. D. Cannas, S. Mandelli, N. Popovic, A. Alkhateeb, A. Gnutti, P. Bestagini, and S. Tubaro, "Is jpeg ai going to change image forensics?" *arXiv preprint arXiv:2412.03261*, 2024.

[207] X. Guo, X. Liu, Z. Ren, S. Grosz, I. Masi, and X. Liu, "Hierarchical fine-grained image forgery detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3155–3165.

[208] X. Ma, B. Du, Z. Jiang, A. Y. A. Hammadi, and J. Zhou, "Iml-vit: Benchmarking image manipulation localization by vision transformer," *arXiv preprint arXiv:2307.14863*, 2023.

[209] L. Su, X. Ma, X. Zhu, C. Niu, Z. Lei, and J.-Z. Zhou, "Can we get rid of handcrafted feature extractors? sparsevit: Nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer," *arXiv preprint arXiv:2412.14598*, 2024.

[210] K. Guo, G. Cao, Z. Lou, X. Huang, and J. Liu, "A lightweight and effective image tampering localization network with vision mamba," *arXiv preprint arXiv:2502.09941*, 2025.

[211] J. Zhu, D. Li, X. Fu, G. Yang, J. Huang, A. Liu, and Z.-J. Zha, "Learning discriminative noise guidance for image forgery detection and localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7739–7747.

[212] Z. Lou, G. Cao, K. Guo, L. Yu, and S. Weng, "Exploring multi-view pixel contrast for general and robust image forgery localization," *IEEE Transactions on Information Forensics and Security*, 2025.

[213] D.-C. Tantaru, E. Oneata, and D. Oneata, "Weakly-supervised deepfake localization in diffusion-generated images," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 6258–6268.

[214] S.-L. Lee, M. Kang, and J.-U. Hou, "Localization of diffusion model-based inpainting through the inter-intra similarity of frequency features," *Image and Vision Computing*, p. 105138, 2024.

[215] Y. Yao, T. Han, S. Jia, and S. Lyu, "Dense feature interaction network for image inpainting localization," *IEEE Transactions on Information Forensics and Security*, 2025.

[216] X. Zhang, R. Li, J. Yu, Y. song Xu, W. Li, and J. Zhang, "Editguard: Versatile image watermarking for tamper localization and copyright protection," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 964–11 974, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266210194

[217] X. Zhang, Z. Tang, Z. Xu, R. Li, Y. song Xu, B. Chen, F. Gao, and J. Zhang, "Omniguard: Hybrid manipulation localization via augmented versatile deep image watermarking," *ArXiv*, vol. abs/2412.01615, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:274436548

[218] N. Bai, X. Wang, R. Han, J. Hou, Y. Wang, and S. Pang, "Pim-net: Progressive inconsistency mining network for image manipulation localization," *Pattern Recognition*, vol. 159, p. 111136, 2025.

[219] R. Han, X. Wang, N. Bai, Y. Wang, J. Hou, and J. Xue, "Hdf-net: Capturing homogeny difference features to localize the tampered image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[220] J. Zhou, X. Ma, X. Du, A. Y. A. Hammadi, and W. Feng, "Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22 289–22 299, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:262832760

[221] S. Smeu, E. Oneata, and D. Oneata, "Declip: Decoding clip representations for deepfake localization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[222] J. Wang, Z. Wu, J. Chen, X. Han, A. Shrivastava, S.-N. Lim, and Y.-G. Jiang, "Objectformer for image manipulation detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2364–2373.

[223] Z. Xu, X. Zhang, R. Li, Z. Tang, Q. Huang, and J. Zhang, "Fakeshield: Explainable image forgery detection and localization via multi-modal large language models," *arXiv preprint arXiv:2410.02761*, 2024.

[224] D. Li, J. Zhu, X. Fu, X. Guo, Y. Liu, G. Yang, J. Liu, and Z.-J. Zha, "Noise-assisted prompt learning for image forgery detection and localization," in *European Conference on Computer Vision*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:274024057

[225] D. Luo, Y. Liu, R. Yang, X. Liu, J. Zeng, Y. Zhou, and X. Bai, "Toward real text manipulation detection: New dataset and new solution," *Pattern Recognition*, vol. 157, p. 110828, 2025.

[226] Z. Huang, B. Xia, Z. Lin, Z. Mou, W. Yang, and J. Jia, "Ffaa: Multimodal large language model based explainable open-world face forgery analysis assistant," *arXiv preprint arXiv:2408.10072*, 2024.

[227] C. Qu, Y. Zhong, C. Liu, G. Xu, D. Peng, F. Guo, and L. Jin, "Towards modern image manipulation localization: A large-scale dataset and novel methods," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 781–10 790, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271245780

[228] X. Zhu, X. Ma, L. Su, Z. Jiang, B. Du, X. Wang, Z. Lei, W. Feng, C.-M. Pun, and J. Zhou, "Mesoscopic insights: Orchestrating multi-scale & hybrid architecture for image manipulation localization," *arXiv preprint arXiv:2412.13753*, 2024.

[229] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, and J. Tang, "Istvt: interpretable spatial-temporal video transformer for deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1335–1348, 2023.

[230] M. Lin, G. Cao, Z. Lou, and C. Zhang, "Spatio-temporal co-attention fusion network for video splicing localization," *Journal of Electronic Imaging*, vol. 33, no. 3, pp. 033 027–033 027, 2024.

[231] J. Hu, X. Liao, D. Gao, S. Tsutsui, Q. Wang, Z. Qin, and M. Z. Shou, "Delocate: Detection and localization for deepfake videos with randomly-located tampered traces," in *International Joint Conference on Artificial Intelligence*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267199866

[232] J. Xu, X. Liu, W. Lin, W. Shang, and Y. Wang, "Localization and detection of deepfake videos based on self-blending method," *Scientific Reports*, vol. 15, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:276058766

[233] Y. Lai, Z. Luo, and Z. Yu, "Detect any deepfakes: Segment anything meets face forgery detection and localization," in *Chinese Conference on Biometric Recognition*. Springer, 2023, pp. 180–190.

[234] S. Saha, R. Perera, S. Seneviratne, T. Malepathirana, S. Rasnayaka, D. Geethika, T. Sim, and S. Halgamuge, "Undercover deepfakes: detecting fake segments in videos," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 415–425.

[235] C. Shuai, J. Zhong, S. Wu, F. Lin, Z. Wang, Z. Ba, Z. Liu, L. Cavallaro, and K. Ren, "Locate and verify: A two-stream network for improved deepfake detection," *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:262064237

[236] J. Jung, S. Lee, J. Kang, and Y. Na, "Www: Where, which and whatever enhancing interpretability in multimodal deepfake detection," *ArXiv*, vol. abs/2408.02954, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271720087

[237] R. K. Bharadwaj, H. Gani, M. Naseer, F. S. Khan, and S. H. Khan, "Vane-bench: Video anomaly evaluation benchmark for conversational lmms," *ArXiv*, vol. abs/2406.10326, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270559848

[238] Z. Cai, S. Ghosh, A. Dhall, T. Gedeon, K. Stefanov, and M. Hayat, "Glitch in the matrix: A large scale benchmark for content driven audio–visual forgery detection and localization," *Computer Vision and Image Understanding*, vol. 236, p. 103818, 2023.

[239] Z. Cai, A. Dhall, S. Ghosh, M. Hayat, D. Kollias, K. Stefanov, and U. Tariq, "1m-deepfakes detection challenge," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 11 355–11 359.

[240] Z. Cai, S. Ghosh, A. P. Adatia, M. Hayat, A. Dhall, T. Gedeon, and K. Stefanov, "Av-deepfake1m: A large-scale llm-driven audio-visual deepfake dataset," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 7414–7423.

[241] V. S. Katamneni and A. Rattani, "Contextual cross-modal attention for audio-visual deepfake detection and localization," *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–11, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271710159

[242] ——, "Multi-modal deepfake detection using attention-based fusion framework," in *The IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2024.

[243] X. Zhang, Y. song Xu, R. Li, J. Yu, W. Li, Z. Xu, and J. Zhang, "V2a-mark: Versatile deep visual-audio watermarking for manipulation localization and copyright protection," *ArXiv*, vol. abs/2404.16824, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269362565

[244] R. Zhang, H. Wang, M. Du, H. Liu, Y. Zhou, and Q. Zeng, "Ummaformer: A universal multimodal-adaptive transformer framework for temporal forgery localization," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 8749–8759.

[245] S. Zeng, J. Yi, J. Tao, J. He, Z. Lian, S. Liang, C. Zhang, Y. Chen, and X. Zhang, "Adversarial training and gradient optimization for partially deepfake audio localization," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[246] J. Wu, W. Lu, X. Luo, R. Yang, Q. Wang, and X. Cao, "Coarse-to-fine proposal refinement framework for audio temporal forgery detection and localization," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 7395–7403.

[247] H.-T. Luong, H. Li, L. Zhang, K. A. Lee, and E. S. Chng, "Llamapartialspoof: An llm-driven fake speech dataset simulating disinformation generation," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[248] R. S. Roman, P. Fernandez, A. Défossez, T. Furon, T. Tran, and H. Elsahar, "Proactive detection of voice cloning with localized watermarking," *arXiv preprint arXiv:2401.17264*, 2024.

[249] A. Mehta, B. McArthur, N. Kolloju, and Z. Tu, "Hfmf: Hierarchical fusion meets multi-stream models for deepfake detection," *arXiv preprint arXiv:2501.05631*, 2025.

[250] C. Miao, Q. Chu, Z. Tan, Z. Jin, T. Gong, W. Zhuang, Y. Wu, B. Liu, H. Hu, and N. Yu, "Multi-spectral class center network for face manipulation detection and localization," *arXiv preprint arXiv:2305.10794*, 2023.

[251] Y. Hou, Q. Guo, Y. Huang, X. Xie, L. Ma, and J. Zhao, "Evading deepfake detectors via adversarial statistical consistency," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 12 271–12 280.

[252] Y. Yang, "Adversarial attack against images classification based on generative adversarial networks," *arXiv preprint arXiv:2412.16662*, 2024.

[253] J. Liu, M. Zhang, J. Ke, and L. Wang, "Advshadow: Evading deepfake detection via adversarial shadow attack," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 4640–4644.

[254] Z. Chen, Z. Wang, J.-j. Huang, W. Zhao, X. Liu, and D. Guan, "Imperceptible adversarial attack via invertible neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 414–424.

[255] K. Tsigos, E. Apostolidis, S. Baxevanakis, S. Papadopoulos, and V. Mezaris, "Towards quantitative evaluation of explainable ai methods for deepfake detection," in *Proceedings of the 3rd ACM international workshop on multimedia AI against disinformation*, 2024, pp. 37–45.

[256] D. A. Coccomini, R. Caldelli, G. Amato, F. Falchi, and C. Gennaro, "Adversarial magnification to deceive deepfake detection through super resolution," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2023, pp. 491–501.

[257] M. Saberi, V. S. Sadasivan, K. Rezaei, A. Kumar, A. Chegini, W. Wang, and S. Feizi, "Robustness of ai-image detectors: Fundamental limits and practical attacks," *arXiv preprint arXiv:2310.00076*, 2023.

[258] C. Galdi, M. Panariello, M. Todisco, and N. Evans, "2d-malafide: Adversarial attacks against face deepfake detection systems," in *2024 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2024, pp. 1–7.

[259] Y. Diao, N. Zhai, C. Miao, X. Yang, and M. Wang, "Vulnerabilities in ai-generated image detection: The challenge of adversarial attacks," *ArXiv*, vol. abs/2407.20836, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271544461

[260] X. Meng, L. Wang, S. Guo, L. Ju, and Q. Zhao, "Ava: Inconspicuous attribute variation-based adversarial attack bypassing deepfake detection," in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2024, pp. 74–90.

[261] Z. Zhou, K. Sun, Z. Chen, H. Kuang, X. Sun, and R. Ji, "Stealthdiffusion: Towards evading diffusion forensic detection through diffusion model," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 3627–3636.

[262] Y. Sun, L. Yu, H. Xie, J. Li, and Y. Zhang, "Diffam: Diffusion-based adversarial makeup transfer for facial privacy protection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 24 584–24 594.

[263] Q. Guo, S. Pang, X. Jia, Y. Liu, and Q. Guo, "Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models," *IEEE Transactions on Information Forensics and Security*, 2024.

[264] J. Liu, C. P. Lau, and R. Chellappa, "Diffprotect: Generate adversarial examples with diffusion models for facial privacy protection," *arXiv preprint arXiv:2305.13625*, 2023.

[265] A. Kassis, U. Hengartner, and Y. Yu, "Unlocking the potential of adaptive attacks on diffusion-based purification," *arXiv preprint arXiv:2411.16598*, 2024.

[266] P. Liu, Q. Tao, and J. Zhou, "Robust deepfake detection by addressing generalization and trustworthiness challenges: A short survey," in *Proceedings of the 1st ACM Multimedia Workshop on Multi-modal Misinformation Governance in the Era of Foundation Models*, 2024, pp. 3–11.

[267] A. Kassis and U. Hengartner, "Unmarker: A universal attack on defensive image watermarking," in *2025 IEEE Symposium on Security and Privacy (SP)*, vol. 2, no. 6, 2025, p. 8.

[268] X. Wu, X. Liao, B. Ou, Y. Liu, and Z. Qin, "Are watermarks bugs for deepfake detectors? rethinking proactive forensics," *arXiv preprint arXiv:2404.17867*, 2024.

[269] P. Kawa, M. Plata, and P. Syga, "Defense against adversarial attacks on audio deepfake detection," *arXiv preprint arXiv:2212.14597*, 2022.

[270] M. U. Farooq, A. Khan, K. Uddin, and K. M. Malik, "Transferable adversarial attacks on audio deepfake detection," *arXiv preprint arXiv:2501.11902*, 2025.

[271] Z. Wang, Y. Zhao, H. Huang, J. Liu, A. Yin, L. Tang, L. Li, Y. Wang, Z. Zhang, and Z. Zhao, "Connecting multi-modal contrastive representations," *Advances in Neural Information Processing Systems*, vol. 36, pp. 22 099–22 114, 2023.

[272] Z. Zhang, Z. Wang, L. Liu, R. Huang, X. Cheng, Z. Ye, H. Liu, H. Huang, Y. Zhao, T. Jin *et al.*, "Extending multi-modal contrastive representations," *Advances in Neural Information Processing Systems*, vol. 37, pp. 91 880–91 903, 2024.

[273] T. Nguyen, N. Khan, and I. Khalil, "Capsfake: A multimodal capsule network for detecting instruction-guided deepfakes," *arXiv preprint arXiv:2504.19212*, 2025.

[274] Y. Cheng, Y. Li, J. He, and R. Feng, "Mixtures of experts for audio-visual learning," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 219–243. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/009729d26288b9a8826023692a876107-Paper-Conference.pdf

[275] B. Pinhasov, R. Lapid, R. Ohayon, M. Sipper, and Y. Aperstein, "Xai-based detection of adversarial attacks on deepfake detectors," *arXiv preprint arXiv:2403.02955*, 2024.