# A Survey of Deepfake and Related Techniques in Computer Vision and Artificial Intelligence

www.surveyx.cn

## Abstract

Deepfake technology, driven by Generative Adversarial Networks (GANs), poses significant societal challenges by enabling the creation of hyper-realistic synthetic media, thus threatening trust and security. This survey paper provides a comprehensive analysis of methodologies for creating and detecting deepfakes, focusing on deep learning architectures like CNNs and Transformers. It evaluates the limitations of current detectors, emphasizing the need for improved generalization across datasets and robust multimodal detectors. The paper categorizes deepfake technologies, exploring data augmentation, synthetic data, inverse graphics, and computer vision's role in enhancing detection models. It highlights the societal impact of deepfakes, urging interdisciplinary collaboration to develop effective forensic methods. The survey identifies challenges, such as the sophistication of deepfake generation outpacing detection efforts, and ethical implications, including privacy and misinformation concerns. Future research directions include enhancing model generalization, expanding datasets, and improving detection resilience against adversarial attacks. The paper underscores the necessity for innovative detection methods to counteract deepfake technologies' escalating sophistication, ensuring the integrity of digital media. By addressing these challenges, the survey aims to contribute to the advancement of reliable detection methods, safeguarding societal interests against the misuse of deepfake technologies.

## 1 Introduction

### 1.1 Significance and Societal Impact

Deepfake technology, driven by advanced generative models like Generative Adversarial Networks (GANs), poses significant challenges to societal trust and security through the creation of hyper-realistic synthetic media [1]. The capability of deepfakes to convincingly impersonate individuals in manipulated videos and synthesized voices raises critical security and privacy concerns [2]. Such synthetic media can mislead audiences by replacing one person's likeness with another, fostering misinformation and threatening personal reputations and public safety.

The societal implications of deepfakes are far-reaching, impacting political, social, financial, and legal stability, particularly through their potential malicious applications [3]. The rapid evolution of deep learning techniques has significantly advanced deepfake technologies, undermining public trust in online media [4]. As these methods progress, the reliability of online information diminishes, challenging media integrity [5]. The emergence of deepfake technologies, especially in facial manipulation, poses a substantial threat to digital society [6].

The rise of realistic synthetic content necessitates effective detection mechanisms to mitigate deepfake-related risks [7]. Concerns regarding video authenticity have intensified with advancements in synthetic video generation, highlighting the need for reliable methods to distinguish real from synthetic content [8]. In regions with underdeveloped technological infrastructure, residents are

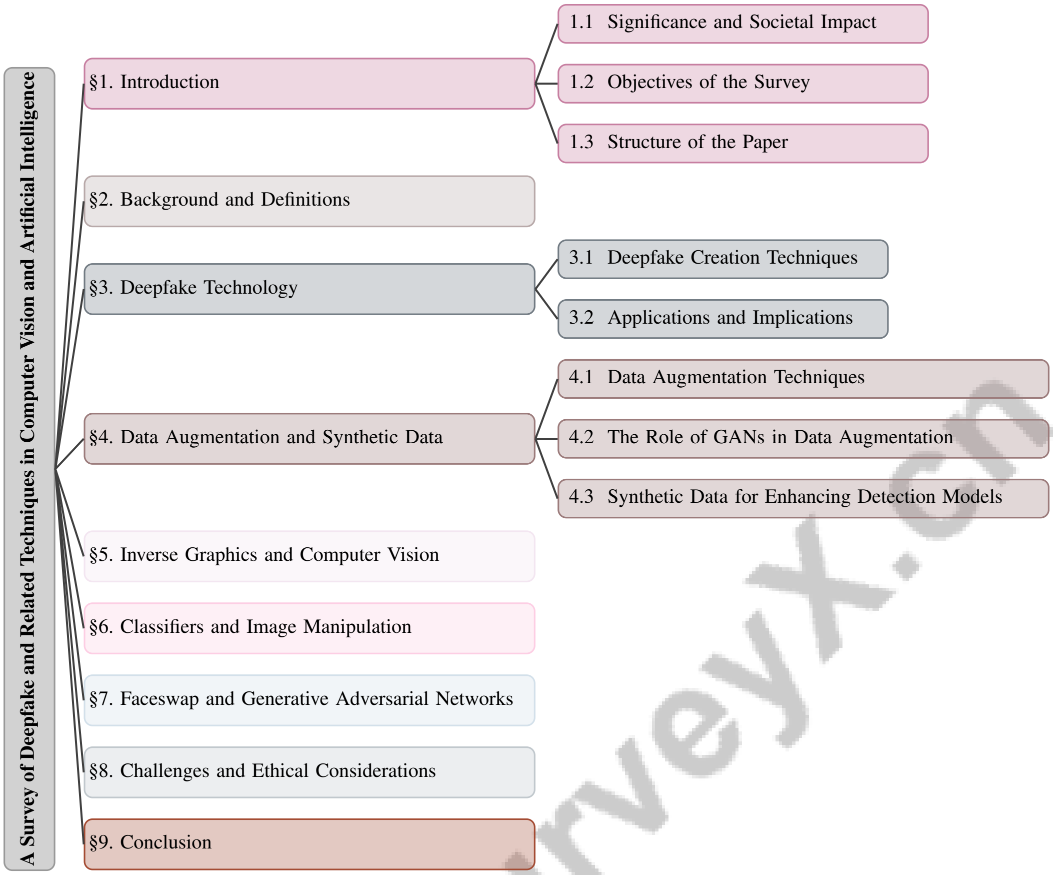

A Survey of Deepfake and Related Techniques in Computer Vision and Artificial Intelligence
§1. Introduction
1.1 Significance and Societal Impact
1.2 Objectives of the Survey
1.3 Structure of the Paper
§2. Background and Definitions
§3. Deepfake Technology
3.1 Deepfake Creation Techniques
3.2 Applications and Implications
§4. Data Augmentation and Synthetic Data
4.1 Data Augmentation Techniques
4.2 The Role of GANs in Data Augmentation
4.3 Synthetic Data for Enhancing Detection Models
§5. Inverse Graphics and Computer Vision
§6. Classifiers and Image Manipulation
§7. Faceswap and Generative Adversarial Networks
§8. Challenges and Ethical Considerations
§9. Conclusion

Figure 1: chapter structure

particularly vulnerable to deception due to limited access to high-quality video, exacerbating societal concerns about misinformation [9].

The difficulty in distinguishing authentic media from manipulated counterparts emphasizes the need for interdisciplinary collaboration to develop new forensic techniques for analyzing synthetic images and addressing the limitations of existing deepfake recognition methods. The creation and detection of deepfake media present significant social and criminal challenges [10]. As deepfake technology advances, its potential to erode public trust and stability remains a pressing issue, underscoring the need for continuous research and innovation to safeguard societal interests [11]. Recent studies also indicate that deep learning models may exhibit biases based on protected classes such as race and gender, further emphasizing the necessity for automated systems capable of combating misinformation, particularly with deepfake videos endangering communities [12]. Thus, the deepfake phenomenon, characterized by AI-generated multimedia content, highlights the urgent requirement for effective detection methods to differentiate real images from those produced by various AI architectures [13].

## 1.2 Objectives of the Survey

This survey aims to comprehensively analyze the methodologies used in the creation and detection of deepfakes, focusing on the rapidly evolving research landscape in this field [8]. It seeks to categorize and critically assess deepfake detection techniques leveraging advanced deep learning architectures, such as convolutional neural networks (CNNs) and Transformers, addressing the challenges in detecting AI-generated media that closely resembles authentic videos. By concentrating on deep learning-based approaches, the survey intentionally excludes non-deep learning methods to maintain a clear scope [14].

A crucial aspect of this survey involves evaluating the limitations of current CNN-based detectors, particularly their generalization capabilities across diverse datasets. This includes exploring innovative

AI-generated, for reference only.

approaches that integrate reconstruction and classification tasks to enhance detection performance [15]. Furthermore, the survey aims to establish benchmarks for future research, facilitating the development of more robust multimodal detectors that can address the increasing sophistication of deepfake videos and audio in digital media [12].

Additionally, the survey strives to provide a thorough understanding of the implications of deepfakes across various sectors, with the goal of developing more robust detection benchmarks [13]. It aims to contribute to the advancement of reliable detection methods that can effectively generalize to forgeries generated from unseen datasets and methodologies, thereby addressing the primary objective of achieving high accuracy in detecting deepfake videos and audio. By tackling the lack of consensus on definitions, performance metrics, datasets, and benchmarks, the survey identifies key challenges and offers recommendations for future research directions [8].

### 1.3 Structure of the Paper

The structure of this survey is designed to systematically explore deepfake technologies, emphasizing both creation and detection methodologies. The paper begins with the **Introduction**, which outlines the significance and societal impact of deepfakes, along with the survey's objectives. Following this, the **Background and Definitions** section provides foundational insights by defining essential concepts and tracing the historical evolution of deepfake and related technologies [16].

**Section 3: Deepfake Technology** examines the technological foundations of deepfakes, focusing on the role of GANs in their creation and the interplay between deepfake generation and detection methods, highlighting the ongoing technological arms race [17]. **Section 4: Data Augmentation and Synthetic Data** discusses the critical role of data augmentation and synthetic data in enhancing AI model performance, particularly in improving the robustness and generalizability of detection models across diverse datasets [18].

**Section 5: Inverse Graphics and Computer Vision** investigates the application of inverse graphics in inferring 3D models from images, contributing to advancements in the broader field of computer vision. In **Section 6: Classifiers and Image Manipulation**, the function of classifiers in deepfake detection is analyzed, along with techniques for identifying image manipulation. This section also considers the adaptability of various AI models for detecting deepfakes across multiple datasets, even under constraints such as limited computational resources [19].

**Section 7: Faceswap and Generative Adversarial Networks** focuses on the technology behind faceswap and the role of GANs in generating realistic synthetic data, including innovative frameworks for image synthesis [20]. **Section 8: Challenges and Ethical Considerations** addresses the multifaceted challenges and ethical implications associated with deepfake technologies, incorporating a multistakeholder perspective to examine the interactions among various actors in the misinformation landscape and detection technologies [21].

Finally, the **Conclusion** synthesizes key findings and reflects on future research opportunities, highlighting the necessity for innovative detection methods to counteract the escalating sophistication of deepfake technologies [22]. This comprehensive approach categorizes current methods into audio, text, video, and image deepfakes, providing a structured framework for understanding the complex landscape of deepfake media [23].The following sections are organized as shown in Figure 1.

## 2 Background and Definitions

### 2.1 Definitions and Core Concepts

Deepfakes, a sophisticated form of synthetic media, leverage neural network architectures such as Generative Adversarial Networks (GANs) and Diffusion Models (DMs) to produce hyper-realistic images and videos that often evade human detection. While these generative models offer creative opportunities in entertainment and cybersecurity, they also raise ethical and security concerns due to potential misuse. Consequently, there is a growing focus on developing robust detection techniques using machine learning methods, including Convolutional Neural Networks (CNNs), to identify subtle inconsistencies in manipulated media [24, 25, 26, 27, 28]. Techniques like face swapping, reenactment, and audio-driven animation highlight the dual-use nature of these technologies, facili-

AI-generated, for reference only.

tating both impersonation and misinformation. The complexity of deepfake creation and detection spans images, videos, and audio, posing challenges to current detection systems.

Data augmentation enhances AI model robustness and generalization by applying transformations to training datasets, crucial for improving model accuracy in detecting manipulated media across various scenarios and addressing data scarcity [15].

Inverse graphics, which reconstructs 3D models from 2D images, enhances machine perception by enabling computers to interpret visual data similarly to humans, vital for analyzing and detecting deepfakes [13].

Classifiers, essential in deepfake detection, categorize data into predefined classes, distinguishing between authentic and manipulated media. Developing sophisticated classification models capable of identifying deepfakes, even from unfamiliar algorithms, remains a significant challenge [12].

Computer vision, a subfield of artificial intelligence, enables machines to interpret visual data, crucial for analyzing and detecting deepfakes and image manipulation [10].

Faceswap technology exemplifies the dual-use nature of deepfake advancements, necessitating ethical considerations due to its potential for benign and malicious applications [14].

Generative Adversarial Networks (GANs), introduced in 2014, consist of a generator creating synthetic data and a discriminator evaluating its authenticity against real data. This architecture excels in image synthesis, enhancing data augmentation strategies and improving neural network training in low-data scenarios [29, 30]. The generator-discriminator collaboration produces realistic synthetic data, complicating the task of distinguishing real from synthetic media.

Image manipulation techniques alter images for specific purposes, posing risks to visual data integrity, especially where authenticity is critical [3]. The challenge lies in differentiating real from manipulated media, given the high realism of modern generated content [31].

### 2.2 Historical Development and Evolution

Deepfake technologies have evolved alongside artificial intelligence and digital media manipulation techniques. Initially emerging from basic image editing, they advanced with neural network architectures, particularly GANs, significantly enhancing synthetic media realism [29]. This progression has made face manipulation technologies more accessible and sophisticated, posing risks to media integrity [14].

As deepfake media realism increased, detection challenges intensified. Early detection relied on traditional image forensics, often inadequate against evolving deepfake capabilities [32]. The transition to advanced machine learning approaches marked a pivotal shift, enabling robust detection methodologies leveraging AI to identify manipulated content [25]. However, maintaining detection accuracy across diverse media types and contexts remains challenging [33].

Benchmarks and datasets have been critical in deepfake detection technology evolution. New benchmarks assess explanation methods' efficacy in identifying influential regions in manipulated images, enhancing deepfake detectors' decision-making capabilities [34]. Despite advancements, generalizing detection methods across media types, including audio and video, continues to be challenging [35].

The trajectory of deepfake technologies reflects a dynamic interplay between generative models' sophistication and detection techniques' advancement. Ongoing research aims to bridge gaps, enhancing detection methods' efficacy across media types and addressing challenges posed by increasing deepfake content realism [25]. Safeguarding digital media integrity remains a critical priority for researchers and practitioners.

In recent years, deepfake technology has garnered significant attention due to its profound implications across various sectors, including media, entertainment, and security. Understanding this technology requires a comprehensive examination of its hierarchical structure, which encompasses various creation techniques and applications. Figure 2 illustrates this structure by categorizing deepfake technology into distinct components: neural network architectures, detection methods, and relevant studies. The figure not only delineates these categories but also emphasizes the pressing need for robust detection strategies. As deepfake content becomes increasingly realistic, the implications

AI-generated, for reference only.

of these advancements necessitate a thorough exploration of methodologies designed to counteract potential misuse and ensure the integrity of visual media.

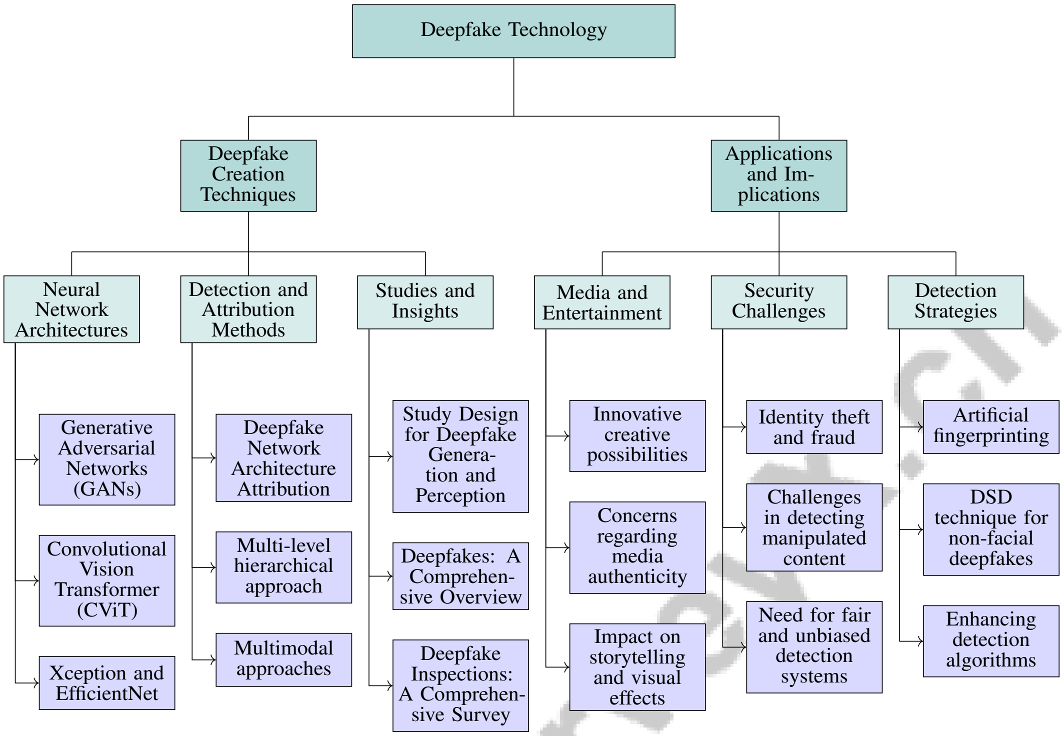

Deepfake Technology
Deepfake Creation Techniques
Applications and Implications
Neural Network Architectures
Detection and Attribution Methods
Studies and Insights
Media and Entertainment
Security Challenges
Detection Strategies
Generative Adversarial Networks (GANs)
Convolutional Vision Transformer (CViT)
Xception and EfficientNet
Deepfake Network Architecture Attribution
Multi-level hierarchical approach
Multimodal approaches
Study Design for Deepfake Generation and Perception
Deepfakes: A Comprehensive Overview
Deepfake Inspections: A Comprehensive Survey
Innovative creative possibilities
Concerns regarding media authenticity
Impact on storytelling and visual effects
Identity theft and fraud
Challenges in detecting manipulated content
Need for fair and unbiased detection systems
Artificial fingerprinting
DSD technique for non-facial deepfakes
Enhancing detection algorithms

Figure 2: This figure illustrates the hierarchical structure of deepfake technology, focusing on creation techniques and applications. It categorizes the technology into neural network architectures, detection methods, and studies, while also exploring its implications in media, entertainment, and security. The figure highlights detection strategies and the need for robust methodologies to counteract the realism of deepfake content.

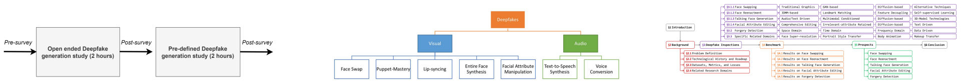
# 3 Deepfake Technology

## 3.1 Deepfake Creation Techniques

The evolution of deepfake technology is primarily driven by advanced neural network architectures, especially Generative Adversarial Networks (GANs). In this framework, a generator produces media while a discriminator evaluates its authenticity, enhancing the realism of the output through adversarial training [14]. This process has facilitated the creation of deepfakes that closely mimic genuine media, applicable in facial image synthesis and video manipulation [13]. GANs are particularly effective in generating high-quality deepfakes even from limited datasets, as demonstrated by specialized datasets like DeepFake MNIST+ that necessitate robust detection models for both low and high-quality video variants [15].

To address the challenges presented by advanced deepfake generation techniques, innovative strategies have been developed. Deepfake Network Architecture Attribution, for instance, aims to associate fake images with their generating architectures without relying on model weights, offering a new perspective for detection [36]. The Convolutional Vision Transformer (CViT), which combines Convolutional Neural Networks (CNN) and Vision Transformers (ViT), enhances detection by utilizing the strengths of both architectures [23]. Models like Xception and EfficientNet have proven effective in deepfake detection, showcasing the capability of sophisticated neural architectures to differentiate between authentic and manipulated media [15]. A multi-level hierarchical approach further boosts classification accuracy, enabling detailed analysis of images generated by various AI architectures [13].

Multimodal approaches, such as those combining audiovisual feature extraction with multi-task learning strategies, highlight the significance of integrating diverse data modalities to improve detection accuracy [8]. These methods emphasize the value of unique behavioral signatures alongside visual cues as effective detection mechanisms.



(a) Study Design for Deepfake Generation and Perception[37]

(b) Deepfakes: A Comprehensive Overview[38]

(c) Deepfake Inspections: A Comprehensive Survey[39]

Figure 3: Examples of Deepfake Creation Techniques

As shown in Figure 3, deepfake technology, which creates convincingly manipulated images or videos, has attracted significant research focus due to its implications for privacy, security, and misinformation. The studies illustrated provide insights into various aspects of deepfake creation. "Study Design for Deepfake Generation and Perception" explores individual perceptions of deepfake images through structured surveys, capturing psychological responses. "Deepfakes: A Comprehensive Overview" categorizes deepfakes into visual and audio forms, detailing methods like face swapping and lip-syncing, revealing the technology's diverse applications and potential for misuse. "Deepfake Inspections: A Comprehensive Survey" examines detection and editing techniques, underscoring the need to understand these methods to mitigate negative impacts. These studies collectively offer a thorough examination of deepfake creation techniques and the ongoing efforts to understand and counteract their effects [37, 38, 39].

### 3.2 Applications and Implications

Deepfake technologies are transforming various sectors, including media, entertainment, and security, by enabling the creation of hyper-realistic synthetic media. In media and entertainment, deepfakes offer innovative creative possibilities, such as placing actors in diverse roles or reviving deceased performers for new projects, thus enhancing storytelling and visual effects [14]. However, the widespread availability of deepfake creation tools raises significant concerns regarding media authenticity, as these technologies can be exploited to spread misinformation and manipulate public perception [40]. The societal risks associated with manipulated media necessitate the development of effective detection methods to safeguard digital content integrity [41].

In the realm of security, deepfakes pose substantial challenges, enabling convincing impersonations that can lead to identity theft, fraud, and the dissemination of false information. The rapid advancement of deepfake algorithms complicates the detection of manipulated content, as the sophistication of generation techniques often exceeds current detection capabilities [31]. Existing detection methods often struggle to identify subtle artifacts introduced during image generation, which are crucial for distinguishing real from synthetic media [42]. Moreover, performance disparities among different racial groups, with error rates differing by up to 10.7%, highlight the urgent need for fair and unbiased detection systems [12].

To improve the detectability of deepfakes, innovative strategies are being explored. Artificial fingerprinting, for example, shows promise by embedding transferable fingerprints within generative models, facilitating consistent detection across various models [42]. Additionally, the DSD technique aims to identify non-facial deepfakes, addressing a gap in existing methods that predominantly focus on facial features [41]. These advancements highlight the necessity of developing robust detection methodologies to counteract the increasing realism of deepfake content.

While deepfakes present exciting opportunities in media and entertainment, their security implications and effects on information authenticity necessitate ongoing research and innovation in detection technologies. The rise of increasingly realistic counterfeit media raises crucial social and security concerns. To effectively address these threats and uphold digital media integrity, it is vital to develop robust detection methods and establish comprehensive benchmarks. This involves enhancing detection algorithms through innovative deep learning techniques and systematically reviewing existing methodologies to identify their strengths and weaknesses. By focusing on these areas, researchers can devise more effective solutions to detect manipulated content across various platforms, ultimately protecting users from misinformation and the malicious use of digital media [28, 10, 27].

AI-generated, for reference only.

As illustrated in Figure 4, which highlights the applications, detection methods, and challenges associated with deepfake technologies, these advancements significantly impact media, security, and authenticity. The figure showcases the promising applications of deepfake technology across various domains, including its use in educational research to enhance survey design and data collection methodologies. The illustrated flowchart outlines key stages, including hypothesis development, meticulous survey design, and deepfake generation to simulate realistic scenarios for participants. This innovative approach enriches educational research by improving understanding of perceptions and behaviors. Furthermore, the performance of deepfake generation models like ProgressiveGAN, StyleGAN, and StyleGAN2 is critically evaluated through attribute-based comparisons, revealing ProgressiveGAN's superior performance in generating believable deepfakes. This highlights the nuanced capabilities of different models and their potential applications, from enhancing virtual learning environments to addressing ethical concerns in digital media [43, 44].

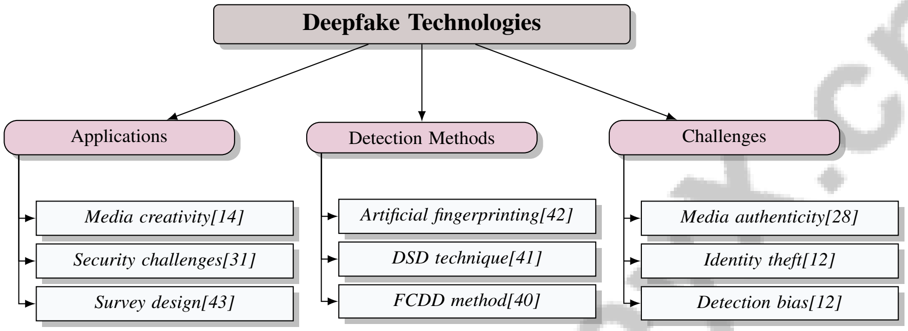


Figure 4: This figure illustrates the applications, detection methods, and challenges associated with deepfake technologies, highlighting their impact on media, security, and authenticity.

## 4 Data Augmentation and Synthetic Data

In the exploration of advanced methodologies for deepfake detection, it is essential to consider the role of data augmentation and synthetic data as foundational strategies. These approaches not only enhance the diversity of training datasets but also significantly improve the robustness and accuracy of detection models. The subsequent subsection will delve into specific data augmentation techniques that have proven effective in enriching the training process, thereby equipping models with the capability to generalize across varied deepfake manipulations and media types.

### 4.1 Data Augmentation Techniques

Data augmentation is a pivotal strategy in enhancing the performance and robustness of models, particularly in the context of deepfake detection. By applying various transformations to training datasets, data augmentation helps prevent overfitting and improve generalization capabilities, which is crucial given the complex and diverse nature of deepfake media [45]. Techniques such as cropping, flips, and noise addition are commonly used to diversify the training data, thereby enhancing the model's ability to generalize to unseen data [46].

One innovative approach is the Face-Cutout technique, which leverages facial landmark positions to dynamically remove regions of an image. This method forces models to learn from incomplete data, thereby enhancing their robustness against manipulations [47]. Additionally, domain-independent augmentation techniques like CutMix and MixUp have been integrated into models to further bolster their generalization capabilities. CutMix involves replacing parts of an image with patches from another image, while MixUp creates new training samples by blending pairs of images and their labels [48].

The application of data augmentation in face data has been extensively surveyed, highlighting both traditional methods and advanced approaches leveraging deep learning techniques such as GANs. GAN-based data augmentation, such as DAGAN, uses a conditional GAN framework to produce augmented samples that maintain class characteristics while introducing variability, thus enriching the training dataset [30].

In practical applications, data augmentation has been effectively utilized in large datasets comprising both real and fake faces, significantly enhancing the training process for detection models. This approach ensures that models are exposed to a wide variety of facial configurations and manipulations, thereby improving their detection accuracy [49]. Furthermore, the integration of data augmentation techniques with advanced neural networks, such as XceptionNet and EfficientNet-B4, underscores the importance of leveraging state-of-the-art architectures to enhance model training [50].

Moreover, the use of robust datasets, such as FaceForensics++, combined with specific hyperparameters and data augmentation techniques, has been shown to significantly improve model performance [34]. This highlights the critical role of well-curated datasets in training effective detection models.

As illustrated in Figure 5, this figure categorizes various data augmentation techniques used in deepfake detection, emphasizing traditional methods such as cropping and noise addition, innovative approaches like Face-Cutout and MixUp, and GAN-based techniques exemplified by DAGAN. These strategies are essential for enhancing model robustness and generalization in detecting manipulated media. "Overall, data augmentation is essential for enhancing the robustness of deepfake detection systems, as it effectively addresses the challenges posed by real-world image distortions and processing variations, thereby improving the generalization ability of these systems in identifying manipulated content" [28, 51, 52, 53]. By employing a combination of traditional and cutting-edge techniques, researchers continue to advance the efficacy of detection models, ensuring they are well-equipped to handle the evolving challenges posed by deepfake technologies.

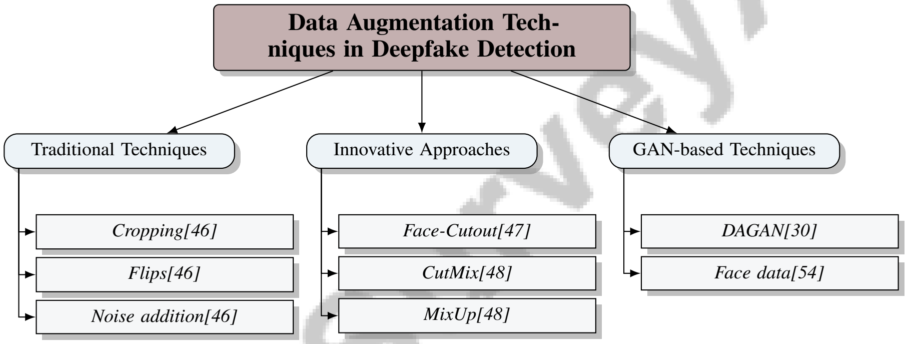


Figure 5: This figure illustrates the categorization of data augmentation techniques used in deepfake detection, highlighting traditional methods such as cropping and noise addition, innovative approaches like Face-Cutout and MixUp, and GAN-based techniques exemplified by DAGAN. These strategies enhance model robustness and generalization in detecting manipulated media.

### 4.2 The Role of GANs in Data Augmentation

Generative Adversarial Networks (GANs) play a pivotal role in the realm of data augmentation, significantly enhancing the robustness and performance of deepfake detection models by generating high-quality synthetic data that closely mirrors real-world distributions. This capability addresses critical challenges related to data scarcity and diversity, which are essential for training models that can generalize effectively across various media types and architectures [41]. The ability of GANs to produce realistic face images enriches training datasets and improves model performance, particularly in the detection of manipulated media [55].

A significant advantage of utilizing GANs in data augmentation is their capacity to create architecture-dependent features that remain consistent across different seeds and fine-tuning processes. This consistency is crucial for enhancing the understanding of GAN-generated data and increasing the efficacy of detection models [11]. Furthermore, GANs facilitate the learning of subtle features through a fine-grained approach, inspired by visual classification methods, which is particularly beneficial for detecting intricate manipulations in deepfake media [1].

Moreover, GANs are instrumental in generating synthetic artifacts, such as those utilized in frameworks that reduce reliance on real DeepFake images by providing negative training examples through simulated artifacts [56]. This approach underscores the importance of GANs in creating diverse datasets that enhance the generalization capabilities of detection models. The integration of adversarial examples, such as pixel-wise Gaussian blurring, further bolsters the adaptability of detection models to new generative techniques without forgetting previously learned information [11].

Innovative methods, such as the adaptation of existing explainability techniques, exemplify the role of GANs in data augmentation by improving the understanding of deepfake detection predictions and enhancing model training [57]. Additionally, the use of Fourier domain analysis to identify anomalies in deepfake images highlights the potential of GANs in advancing data augmentation strategies [55].

Overall, GANs are essential in advancing data augmentation strategies by providing high-quality synthetic data that enhances the training of detection models. The ability to generate diverse and realistic datasets is essential for effectively combating the challenges presented by advanced deepfake technologies, as it enables the development of robust machine learning models that can accurately detect manipulated media across a wide array of contexts, including in-the-wild images and videos. This necessity arises from the increasing sophistication of deepfake generation techniques, such as Generative Adversarial Networks (GANs) and Diffusion Models, which produce highly convincing synthetic content. Consequently, ongoing research in deepfake detection focuses on identifying subtle inconsistencies and artifacts within this media, ensuring that detection algorithms remain effective even against novel and unseen deepfake generators. [25, 26, 27, 58, 10]

### 4.3 Synthetic Data for Enhancing Detection Models

The strategic use of synthetic data is crucial in enhancing the robustness and efficacy of deepfake detection models, offering diverse datasets that enable models to generalize across various manipulation techniques and media types. Datasets such as those derived from well-established sources, including FaceForensics++, Google DFD, Celeb-DF, Deeper Forensics, and the Facebook Deepfake Detection Challenge, provide a comprehensive array of real and fake face images, facilitating the development of more accurate detection models . These datasets incorporate images generated by various GANs and Diffusion models, offering a diverse range of synthetic content for comprehensive evaluation [7].

Generative Adversarial Networks (GANs) play a pivotal role in synthetic data generation, significantly enriching training datasets by producing high-quality synthetic data that closely mirrors real-world distributions. This capability addresses challenges related to data scarcity and diversity, which are essential for training models that can effectively generalize across different media types [13]. The integration of synthetic datasets comprising 42,500 synthetic images and 40,500 real images, sourced from CelebA, FFHQ, and ImageNet, underscores the significance of synthetic data in improving model performance [13].

Innovative approaches such as DNA-Det, which performs architecture-level attribution, highlight the potential of synthetic data in real-world scenarios where model-level attribution is less applicable [36]. Additionally, the exploitation of unique characteristics in the latent space associated with different models allows for accurate discrimination between images generated by various GAN architectures [59].

The inclusion of manipulated videos with corresponding synthesized audios, specifically designed to be lip-synced, further enhances the comprehensive testing of detection methods across multiple modalities [2]. Furthermore, benchmarks introducing video corruption pipelines simulate various degradation methods, providing a more realistic assessment of deepfake detection models [9].

Overall, the use of synthetic data is integral to advancing deepfake detection technologies, ensuring models remain effective in identifying manipulations across diverse scenarios and media types. This approach effectively meets the pressing demand for enhanced detection techniques and the establishment of extensive datasets for validation, thereby facilitating future research aimed at improving the generalization and robustness of models. By investigating the specific artifacts that distinguish fake images and employing advanced methodologies such as patch-based classifiers and reverse engineering of generative models, this strategy not only addresses the current challenges posed by sophisticated image synthesis algorithms but also lays the groundwork for developing more resilient detection systems that can adapt to new and unseen manipulations. [60, 61, 10]

## 5 Inverse Graphics and Computer Vision

Inverse graphics is a foundational concept in computer vision, enabling machines to interpret visual data by reconstructing three-dimensional (3D) structures from two-dimensional (2D) images. This section delves into the principles of inverse graphics and its pivotal role in image manipulation detection, which is elaborated in the following subsection.

### 5.1 Concept of Inverse Graphics

Inverse graphics facilitates the inference of 3D structures from 2D images, mimicking the human visual system's depth perception capabilities. This computational process is crucial for enabling machines to deduce the physical properties underlying visual inputs, effectively reversing the conventional graphics pipeline that transforms 3D models into 2D representations [29]. Incorporating inverse graphics into machine learning frameworks has significantly advanced the interpretation of visual data, achieving a level of sophistication akin to human perception. This advancement is particularly relevant in image tampering detection, where reconstructing original scenes from altered images is critical. Detection models in this domain are assessed using metrics that evaluate their effectiveness in identifying manipulations [62].

As illustrated in Figure 6, the figure encapsulates the key concepts of inverse graphics, highlighting the processes involved in 3D structure inference, the pivotal role of Generative Adversarial Networks (GANs), and their applications in image tampering detection. GANs have propelled inverse graphics forward, enabling the creation of high-quality synthetic images that closely resemble real ones. Their applications in text-to-image synthesis and image-to-image translation highlight GANs' potential to generate realistic visual content, thereby enhancing inverse graphics' capacity to reconstruct and interpret complex visual environments [29]. This ongoing research continues to bridge the gap between 2D observations and 3D reality.

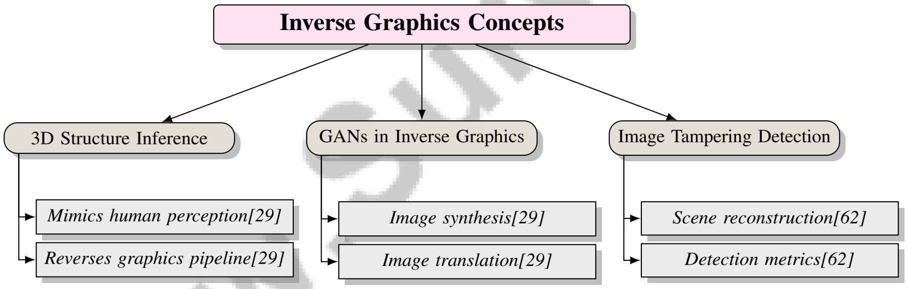


Figure 6: This figure illustrates the key concepts of inverse graphics, focusing on 3D structure inference, the role of GANs, and applications in image tampering detection.

### 5.2 Application of Inverse Graphics in Detecting Manipulations

Inverse graphics plays a crucial role in detecting image manipulations by reconstructing 3D structures from 2D images, enabling the identification of inconsistencies such as unnatural lighting or spatial discrepancies that arise during manipulation [63]. These techniques can reveal alterations not immediately apparent in 2D representations, thereby bolstering the robustness of manipulation detection systems.

The application of synthetic data generated through inverse graphics enhances the accuracy and reliability of biometric systems, as evidenced by discussions on improving biometric technologies [63]. This approach facilitates the creation of high-quality synthetic biometric samples for training, equipping models to detect manipulations across various scenarios and media types. By simulating diverse manipulation techniques and their effects on perceived 3D structures, inverse graphics enhances models' detection capabilities against deepfakes and other altered media.

Integrating inverse graphics into manipulation detection highlights its essential role in computer vision, providing advanced methodologies for identifying subtle artifacts in increasingly sophisticated image and video alterations. This approach enhances detection systems' generalization across diverse datasets and model architectures, addressing challenges posed by rapidly evolving image synthesis technologies, including deepfakes. Researchers are developing robust solutions utilizing innovative frameworks such as graph neural networks and advanced classification strategies, significantly improving the accuracy and reliability of fake image and video detection, thus contributing to the preservation of information integrity and security in the digital landscape [61, 64].

## 6 Classifiers and Image Manipulation

### 6.1 Role of Classifiers in Deepfake Detection

Classifiers are pivotal in deepfake detection, leveraging advanced machine learning to differentiate between genuine and altered media. The CViT model exemplifies this with a 91.5% accuracy and 0.91 AUC on the DFDC dataset, showcasing how sophisticated classifier architectures can enhance detection by identifying intricate visual data patterns [65]. Multimodal analysis, integrating audio and visual features, further improves detection accuracy, as demonstrated by Muppalla et al., by providing a comprehensive media understanding [66].

Innovative approaches like the FCDD utilize additional classifiers to detect manipulations, highlighting their versatility [40]. However, adversarial threats such as the AVA attack, which bypasses commercial detectors over 95% of the time, underscore the need for continuous refinement of classifier models to counter sophisticated evasion techniques [67]. Guarnera et al. reveal that generative architectures leave unique fingerprints, aiding classifiers in distinguishing between single and multiple manipulations, thus enhancing detection robustness [56].

The evolution of classifiers in deepfake detection reflects the integration of diverse methodologies and adaptation to challenges posed by advanced generative techniques, essential for maintaining digital media integrity and ensuring reliable detection across scenarios [20].

### 6.2 Techniques for Image Manipulation Detection

Sophisticated techniques are crucial for detecting image manipulations within the broader context of deepfake detection. AdvHeat, for instance, surpasses traditional adversarial attacks by exploiting vulnerabilities across detectors, enhancing transferability and robustness [68]. The integration of attention layers and siamese training strategies into CNNs, as discussed by Bonettini et al., advances the focus on informative features, improving class generalization [69].

FDFtNet employs fine-tuning of pretrained models with a self-attention module, providing a robust framework for identifying manipulated media [70]. Khoo et al. address the challenges of attributing synthetic images generated by various models, crucial for understanding manipulated content origins and improving detection frameworks [71]. Lu et al. highlight the impact of video processing on detection accuracy, revealing notable performance degradation from realistic processing operations, indicating a need for enhanced detection robustness [72, 4].

Delmas et al. propose a frugal deepfake detection method with lower resource requirements that maintains effective performance with limited training data, valuable in resource-constrained environments [73]. The evolution of detection techniques is significantly influenced by adversarial methods, attention mechanisms, and resource-efficient frameworks, addressing challenges posed by sophisticated forgery techniques like Deepfakes and Face2Face. Recent research underscores the effectiveness of ViT-based approaches, which excel in detecting manipulated images across scenarios, including those with unknown post-processing. Adaptive manipulation trace extraction networks and ensemble modeling techniques further enhance detection accuracy and generalization, underscoring the importance of these advancements in maintaining digital media integrity [74, 75, 61, 76, 77].

### 6.3 Advanced Classifier Architectures

Advanced classifier architectures are vital for enhancing deepfake detection capabilities amid increasingly sophisticated synthetic media. Innovations in machine learning have led to architectural enhancements that significantly boost classifier accuracy and reliability. The Convolutional Vision Transformer (CViT) combines CNN and ViT strengths, achieving superior performance in deepfake detection tasks [23]. This hybrid architecture leverages CNNs' spatial hierarchies and ViTs' global context awareness, enhancing feature extraction and generalization across datasets.

Attention mechanisms within CNNs allow models to focus on critical image regions likely to contain manipulation artifacts, improving interpretability and robustness [69]. Self-supervised learning techniques have shown promise in enhancing model performance by utilizing unlabeled data to learn meaningful representations, improving detection accuracy in challenging scenarios [70]. This is particularly beneficial in deepfake detection, where labeled data may be scarce.

Lightweight and efficient architectures, like those proposed by Delmas et al., offer a frugal approach to deepfake detection suitable for resource-constrained environments [73]. These architectures prioritize computational efficiency while maintaining high detection accuracy, ideal for real-time applications.

The exploration of advanced classifier architectures remains a vital research area, fostering innovations that enhance both precision and speed of deepfake detection systems. As sophisticated deep learning models produce highly convincing counterfeit media, robust computational models are urgently needed to identify and mitigate risks associated with manipulated content. Recent studies highlight challenges posed by limited computing resources and the necessity for diverse deep learning techniques to enhance detection efficiency. Continuous research into various methodologies is crucial for developing effective deepfake detection systems to safeguard information integrity and counter misinformation [28, 27, 78]. By integrating state-of-the-art machine learning techniques and architectural innovations, researchers aim to create robust classifiers capable of addressing the evolving challenges posed by deepfake technologies.

## 7 Faceswap and Generative Adversarial Networks

The intersection of faceswap technology and Generative Adversarial Networks (GANs) is notable for its creative potential and ethical considerations. GANs are instrumental in generating high-quality images, significantly enhancing the realism of synthetic media in faceswap applications. This section examines the frameworks and methodologies of GAN-based image synthesis, highlighting their influence on faceswap technologies and the challenges related to media authenticity and detection.

### 7.1 GAN-based Frameworks in Image Synthesis

Generative Adversarial Networks (GANs) have revolutionized image synthesis by enabling the creation of highly realistic visuals. The GAN architecture consists of two neural networks: a generator that creates images and a discriminator that evaluates their authenticity, refining the generator's output through an adversarial process [13]. Notable models like StyleGAN and ProGAN exemplify facial image synthesis, producing high-fidelity images with intricate details through techniques such as progressive growing, which enhances image resolution incrementally [56, 36].

GANs have also advanced image-to-image translation frameworks, enabling domain transformations while maintaining essential content. Applications include converting sketches to photorealistic images and altering daytime scenes to nighttime visuals, with CycleGAN emerging as a powerful tool for unsupervised image-to-image translation [57, 55]. In deepfake media, GANs facilitate the creation of synthetic videos and images closely resembling genuine content, necessitating sophisticated detection methods to prevent misuse [13]. Their role in generating training data for detection models underscores GANs' dual impact on advancing and challenging AI capabilities in media synthesis and security [1].

Since their inception, GANs have been pivotal in image synthesis research, with applications across computer vision and natural language processing. While they enrich creative content generation, the potential misuse of GAN technology for deepfakes raises significant ethical and privacy concerns. Recent advancements in detection algorithms, such as those employing Discrete Cosine Transform (DCT) to identify GAN-specific frequencies, are crucial for mitigating the risks associated with malicious applications of this powerful AI technology [29, 79].

### 7.2 Wavelet Knowledge Distillation in Faceswap

Wavelet knowledge distillation in faceswap technology is an innovative approach where a smaller student network is trained using outputs from a larger teacher network, focusing on high-frequency details through wavelet analysis [80]. This technique employs wavelet transforms to decompose images into various frequency components, allowing the student network to efficiently learn by emphasizing critical features often lost during downscaling. Integrating wavelet analysis helps the student network retain more detail, crucial for maintaining the realism and authenticity of swapped faces. This approach not only improves faceswap algorithm performance but also enhances detection models by increasing their ability to identify subtle artifacts introduced during the faceswap process. Wavelet knowledge distillation thus represents a significant advancement, providing a robust framework for both the creation and detection of faceswap media.

## 8 Challenges and Ethical Considerations

### 8.1 Detection Methods and Challenges

| Benchmark | Size | Domain | Task Format | Metric |
|---|---|---|---|---|
| FFpp[4] | 5,000 | Media Forensics | Binary Classification | AUC |
| MDDR[13] | 83,000 | Multimedia Forensics | Classification | Accuracy |
| FakeAVCeleb[15] | 1,200 | Media Forensics | Deepfake Detection | F1-score, Accuracy |
| CT-GAN[62] | 2,384 | Medical Imaging | Image Classification | Accuracy, AUC |
| DFSI[81] | 30,000 | Image Classification | Image Classification | Accuracy, F1-score |
| FaceForensics++[76] | 1,800,000 | Facial Manipulation Detection | Binary Classification | Accuracy, F1-score |
| DDDB[82] | 65,556 | Art Detection | Image Classification | AA, mAP |
| DFD[53] | 104,000 | Video Forensics | Face Manipulation Detection | AUC |

Table 1: This table provides a comprehensive overview of various benchmarks used in the evaluation of deepfake detection methods. It includes details on the size, domain, task format, and performance metrics for each benchmark, highlighting the diversity and scope of datasets utilized in media forensics and related fields.

The rapid evolution of generative technologies presents significant challenges in detecting deepfakes, as these technologies produce increasingly realistic media. A major obstacle is the generalization of detection methods, which often overfit specific datasets and perform poorly on new data [14]. Many techniques focus on artifacts specific to certain manipulations, limiting their applicability to novel methods [36]. The transferability of models is crucial, as many struggle with different datasets and real-world applications [23].

Current benchmarks often use test data from the same distribution as training data, which can lead to overestimated performance and fail to reflect real-world complexities [4]. Table 1 presents a detailed comparison of representative benchmarks that are critical for assessing the effectiveness of deepfake detection techniques across different domains and task formats. Guarnera et al.'s framework contributes significantly by providing a multi-level classification approach for evaluating detection methods [13]. Innovative strategies are needed to address these challenges, such as Khalid et al.'s evaluation of benchmarks that highlight the scarcity of publicly available datasets, restricting research community access [15]. Developing cross-dataset detection methods is essential to enhance model generalization and resilience against adversarial attacks.

Despite progress, deepfake generation sophistication often surpasses detection capabilities. Future research should focus on adaptable detection methods to handle images from unseen or privately trained models, addressing significant gaps in current capabilities [36].

### 8.2 Ethical Implications and Societal Impact

Deepfake technologies pose profound ethical and societal challenges, raising issues of privacy, consent, and misuse in creating misleading media. The ability to produce non-consensual content presents ethical concerns, particularly with hyper-realistic media that can facilitate identity theft and misinformation [1]. These technologies threaten media integrity and authenticity, underscoring the need for forensic models to detect GAN-generated images [83].

Beyond individual privacy, deepfakes could undermine trust in digital media. The ongoing arms race between generation and detection technologies complicates the ethical landscape, as generative model advancements often outpace detection method development [61]. This necessitates adaptive detection methods to manage the sophistication of deepfake technologies [20].

The misuse of deepfake technology for avatar personalization and digital representations raises authenticity concerns, requiring ethical guidelines and standards [58]. Integrating audio and visual cues in detection methods is vital for categorizing deepfake samples based on characteristics to address ethical challenges [66]. However, the risk of datasets being misused for malicious purposes remains a critical limitation, requiring continuous updates to keep pace with evolving technologies [2].

Addressing fairness in AI systems is essential, as biased detection can have significant societal consequences for underrepresented groups [12]. Understanding detection method vulnerabilities in low-quality contexts is crucial for developing solutions against disinformation [9]. Additionally, variability in user preferences for explanation methods suggests a combination of techniques may be necessary for effective forensic investigation [57].

# 9 Conclusion

## 9.1 Future Directions and Research Opportunities

Advancements in deepfake technology necessitate a focused approach to enhance detection capabilities amidst increasingly sophisticated generative models. It is imperative to improve the generalization of detection systems across diverse deepfake creation methods and bolster their resilience against adversarial manipulations. This involves refining detection models for more precise visual feature interpretation and developing cost-effective strategies. Expanding dataset diversity remains a promising area, requiring the inclusion of varied generative models and modalities to enhance detection accuracy. Incorporating temporal data and intelligent voting mechanisms is anticipated to improve performance further. Additionally, broadening datasets to encompass a spectrum of deepfake techniques and evaluation metrics is crucial for comprehensive robustness assessments.

Enhancing the adaptability of detection methods to novel video generators and exploring additional feature sets to boost accuracy are vital. This includes refining evaluation metrics and expanding datasets to cover a wider array of manipulation techniques. Future research should also focus on the effects of image degradation on detection efficacy, delving into sophisticated models to address these challenges. Advanced fusion techniques could augment model explainability and real-time detection capabilities. Efforts to improve generalization against user-customized models and enhance adversarial resilience through innovative training strategies are essential. Developing detectors capable of handling out-of-distribution content and exploring emerging trends in generative modeling are necessary for continued progress.

In the realm of audio deepfake detection, the development of specialized models and the enhancement of multimodal approaches are critical, alongside the establishment of benchmarks for AI-generated media detection. These initiatives will foster a comprehensive understanding of deepfake technologies, ensuring responsible use and the protection of digital media integrity. Continuous research and innovation are vital to mitigate societal impacts, ensuring that detection technologies evolve alongside generative advancements. Investigating memory-based strategies to enhance the robustness of continual learning techniques in deepfake detection presents a promising avenue. The application of frameworks like AntiDeepFake to video analysis, expanding beyond audio recognition, offers another vital exploration area. Establishing standardized definitions, addressing ethical considerations, and expanding datasets to cover a broader range of applications are essential steps forward.

Enhancing multi-supervision modules, refining super-resolution techniques, and improving sequential predictions will further boost detection and recovery performance. Improving the transferability of generated adversarial images across different model architectures and examining the relationship between noise levels, face attributes, and adversarial strength are critical. Expanding datasets to include diverse manipulation types and enhancing model robustness against various techniques are vital. Investigating positive applications of deepfake technology and developing ethical frameworks for its use in fields like entertainment, education, and healthcare are also essential. Enhancing dataset diversity and implementing adversarial patches to improve model robustness are crucial steps. Optimizing the DeepFake detection pipeline and deploying it as a web service for broader accessibility are future directions worth pursuing. Exploring the integration of physiological signals and alternative deep learning architectures to enhance fake video detection robustness is also essential.

Future research should refine methods like FCDD to improve detection accuracy in challenging scenarios and explore additional identity manipulation cues. Expanding datasets for non-facial videos and enhancing the generalization capabilities of detectors across different video content are critical areas of focus. Enhancing detection methods for diverse visual content and integrating fact verification techniques are paramount. Expanding datasets, improving methodologies, and investigating adversarial attacks against detection systems remain vital. Integrating multiple explanation methods to enhance detection performance and exploring user-centric evaluation metrics are essential for future research. Additionally, expanding datasets to include diverse actions and exploring advanced detection techniques are critical. Investigating deeper behavioral analysis techniques and broadening datasets to encompass various subjects and contexts are important research directions. Developing robust detection methods that withstand a wider range of distortions will further advance the field. Enhancing method robustness against complex architectures and improving performance in real-world conditions are crucial. Considering ethical implications of deploying these technologies in vulnerable populations is also vital.

Future research should concentrate on enhancing model transferability, developing universal datasets for training and testing, and creating accessible detection tools for the public. Efforts should focus on improving the generalization capabilities of detection systems, exploring fusion techniques, and establishing new benchmarks for evaluating manipulation detection performance. Developing novel multimodal detection methods that leverage unique dataset characteristics is essential for improving detection accuracy. Furthermore, exploring racially aware methods for data generation and additional fairness metrics in deepfake detection are crucial areas for future research. Lastly, investigating benchmark robustness in real-world scenarios and identifying analytical traces unique to different generative models could provide valuable insights for advancing detection methodologies.

## References

[1] Chaofei Yang, Lei Ding, Yiran Chen, and Hai Li. Defending against gan-based deepfake attacks via transformation-aware adversarial faces, 2020.

[2] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset, 2022.

[3] Vrizlynn L. L. Thing. Deepfake detection with deep learning: Convolutional neural networks versus transformers, 2023.

[4] Yuhang Lu and Touradj Ebrahimi. Assessment framework for deepfake detection in real-world situations, 2023.

[5] Xiaoyu Cao and Neil Zhenqiang Gong. Understanding the security of deepfake detection, 2021.

[6] Jiajun Huang, Xueyu Wang, Bo Du, Pei Du, and Chang Xu. Deepfake mnist+: A deepfake facial animation dataset, 2021.

[7] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection, 2024.

[8] Qiaomu Miao, Sinhwa Kang, Stacy Marsella, Steve DiPaola, Chao Wang, and Ari Shapiro. Study of detecting behavioral signatures within deepfake videos, 2024.

[9] Yang A. Chuming, Daniel J. Wu, and Ken Hong. Practical deepfake detection: Vulnerabilities in global contexts, 2022.

[10] Trung-Nghia Le, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Robust deepfake on unrestricted media: Generation and detection, 2022.

[11] Chia-Mu Yu, Ching-Tang Chang, and Yen-Wu Ti. Detecting deepfake-forged contents with separable convolutional neural network and image segmentation, 2019.

[12] Loc Trinh and Yan Liu. An examination of fairness of ai models for deepfake detection, 2021.

[13] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models, 2023.

[14] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection, 2020.

[15] Hasam Khalid, Minha Kim, Shahroz Tariq, and Simon S. Woo. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors, 2021.

[16] Bahar Uddin Mahmud and Afsana Sharmin. Deep insights of deepfake technology : A review, 2023.

[17] Zhi Wang, Yiwen Guo, and Wangmeng Zuo. Deepfake forensics via an adversarial game, 2022.

[18] Liviu-Daniel Ştefan, Dan-Cristian Stanciu, Mihai Dogariu, Mihai Gabriel Constantin, Andrei Cosmin Jitaru, and Bogdan Ionescu. Deepfake sentry: Harnessing ensemble intelligence for resilient detection and generalisation, 2024.

[19] Piotr Kawa and Piotr Syga. A note on deepfake detection with low-resources, 2020.

[20] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, and Yang Liu. Countering malicious deepfakes: Survey, battleground, and horizon, 2022.

[21] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts, 2019.

[22] Wasim Ahmad, Imad Ali, Adil Shahzad, Ammarah Hashmi, and Faisal Ghaffar. Resvit: A framework for deepfake videos detection. *International journal of electrical and computer engineering systems*, 13(9):807–813, 2022.

[23] Jacob Mallet, Rushit Dave, Naeem Seliya, and Mounika Vanamala. Using deep learning to detecting deepfakes, 2022.

[24] Aniruddha Tiwari, Rushit Dave, and Mounika Vanamala. Leveraging deep learning approaches for deepfake detection: A review, 2023.

[25] Florinel-Alin Croitoru, Andrei-Iulian Hiji, Vlad Hondru, Nicolae Catalin Ristea, Paul Irofti, Marius Popescu, Cristian Rusu, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Deepfake media generation and detection in the generative ai era: A survey and outlook, 2024.

[26] Irene Amerini, Mauro Barni, Sebastiano Battiato, Paolo Bestagini, Giulia Boato, Tania Sari Bonaventura, Vittoria Bruni, Roberto Caldelli, Francesco De Natale, Rocco De Nicola, Luca Guarnera, Sara Mandelli, Gian Luca Marcialis, Marco Micheletto, Andrea Montibeller, Giulia Orru', Alessandro Ortis, Pericle Perazzo, Giovanni Puglisi, Davide Salvi, Stefano Tubaro, Claudia Melis Tonti, Massimo Villari, and Domenico Vitulano. Deepfake media forensics: State of the art and challenges ahead, 2024.

[27] Sm Zobaed, Md Fazle Rabby, Md Istiaq Hossain, Ekram Hossain, Sazib Hasan, Asif Karim, and Khan Md. Hasib. Deepfakes: Detecting forged and synthetic media content using machine learning, 2021.

[28] Leandro A. Passos, Danilo Jodas, Kelton A. P. da Costa, Luis A. Souza Júnior, Douglas Rodrigues, Javier Del Ser, David Camacho, and João Paulo Papa. A review of deep learning-based approaches for deepfake content detection, 2024.

[29] He Huang, Philip S. Yu, and Changhu Wang. An introduction to image synthesis with generative adversarial nets, 2018.

[30] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks, 2018.

[31] Ammarah Hashmi, Sahibzada Adil Shahzad, Chia-Wen Lin, Yu Tsao, and Hsin-Min Wang. Unmasking illusions: Understanding human perception of audiovisual deepfakes, 2024.

[32] Naciye Celebi, Qingzhong Liu, and Muhammed Karatoprak. A survey of deep fake detection for trial courts, 2022.

[33] Enes Altuncu, Virginia N. L. Franqueira, and Shujun Li. Deepfake: Definitions, performance metrics and standards, datasets and benchmarks, and a meta-review, 2022.

[34] Konstantinos Tsigos, Evlampios Apostolidis, Spyridon Baxevanakis, Symeon Papadopoulos, and Vasileios Mezaris. Towards quantitative evaluation of explainable ai methods for deepfake detection, 2024.

[35] Joel Frank, Franziska Herbert, Jonas Ricker, Lea Schönherr, Thorsten Eisenhofer, Asja Fischer, Markus Dürmuth, and Thorsten Holz. A representative study on human detection of artificially generated media across countries, 2023.

[36] Tianyun Yang, Ziyao Huang, Juan Cao, Lei Li, and Xirong Li. Deepfake network architecture attribution, 2022.

[37] Pulak Mehta, Gauri Jagatap, Kevin Gallagher, Brian Timmerman, Progga Deb, Siddharth Garg, Rachel Greenstadt, and Brendan Dolan-Gavitt. Can deepfakes be created by novice users?, 2023.

[38] Momina Masood, Marriam Nawaz, Khalid Mahmood Malik, Ali Javed, and Aun Irtaza. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward, 2021.

[39] Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey, 2024.

[40] Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. Deepfake detection based on the discrepancy between the face and its context, 2020.

AI-generated, for reference only.

[41] Omran Alamayreh and Mauro Barni. Detection of gan-synthesized street videos, 2021.

[42] Run Wang, Felix Juefei-Xu, Meng Luo, Yang Liu, and Lina Wang. Faketagger: Robust safeguards against deepfake dissemination via provenance tracking, 2021.

[43] Gabrielle Watson, Zahra Khanjani, and Vandana P. Janeja. Audio deepfake perceptions in college going populations, 2021.

[44] Baiwu Zhang, Jin Peng Zhou, Ilia Shumailov, and Nicolas Papernot. On attribution of deepfakes, 2021.

[45] Hannah Lee, Changyeon Lee, Kevin Farhat, Lin Qiu, Steve Geluso, Aerin Kim, and Oren Etzioni. The tug-of-war between deepfake generation and detection, 2024.

[46] Sohail Ahmed Khan, Alessandro Artusi, and Hang Dai. Adversarially robust deepfake media detection using fused convolutional neural network predictions, 2021.

[47] Sowmen Das, Selim Seferbekov, Arup Datta, Md. Saiful Islam, and Md. Ruhul Amin. Towards solving the deepfake problem : An analysis on improving deepfake detection using dynamic face augmentation, 2021.

[48] Sudarshana Kerenalli, Vamsidhar Yendapalli, and Mylarareddy Chinnaiah. Classification of deepfake images using a novel explanatory hybrid model. *CommIT (Communication and Information Technology) Journal*, 17(2):151–168, 2023.

[49] Jacob mallet, Laura Pryor, Rushit Dave, and Mounika Vanamala. Deepfake detection analyzing hybrid dataset utilizing cnn and svm, 2023.

[50] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Hybrid transformer network for deepfake detection, 2022.

[51] Yuhang Lu, Ruizhi Luo, and Touradj Ebrahimi. A novel framework for assessment of learning-based detectors in realistic conditions with application to deepfake detection, 2022.

[52] Yuhang Lu and Touradj Ebrahimi. A new approach to improve learning-based deepfake detection in realistic conditions, 2022.

[53] Luca Bondi, Edoardo Daniele Cannas, Paolo Bestagini, and Stefano Tubaro. Training strategies and data augmentations in cnn-based deepfake video detection, 2020.

[54] Xiang Wang, Kai Wang, and Shiguo Lian. A survey on face data augmentation, 2019.

[55] Luca Guarnera, Oliver Giudice, Cristina Nastasi, and Sebastiano Battiato. Preliminary forensics analysis of deepfake images, 2020.

[56] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake style transfer mixture: a first forensic ballistics study on synthetic images, 2022.

[57] Samuele Pino, Mark James Carman, and Paolo Bestagini. What's wrong with this video? comparing explainers for deepfake detection, 2021.

[58] Nikolaos Misirlis and Harris Bin Munawar. From deepfake to deep useful: risks and opportunities through a systematic literature review, 2023.

[59] Luca Guarnera, Oliver Giudice, Matthias Niessner, and Sebastiano Battiato. On the exploitation of deepfake model recognition, 2022.

[60] Vishal Asnani, Xi Yin, Tal Hassner, and Xiaoming Liu. Reverse engineering of generative models: Inferring model hyperparameters from generated images, 2023.

[61] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XXVI 16*, pages 103–120. Springer, 2020.

[62] Siddharth Solaiyappan and Yuxin Wen. Machine learning based medical image deepfake detection: A comparative study, 2022.

[63] John Jenkins and Kaushik Roy. Exploring deep convolutional generative adversarial networks (dcgan) in biometric systems: a survey study. *Discover Artificial Intelligence*, 4(1):42, 2024.

[64] Mostafa M El-Gayar, Mohamed Abouhawwash, Sameh S Askar, and Sara Sweidan. A novel approach for detecting deep fake videos using graph neural network. *Journal of Big Data*, 11(1):22, 2024.

[65] Deressa Wodajo and Solomon Atnafu. Deepfake video detection using convolutional vision transformer, 2021.

[66] Sneha Muppalla, Shan Jia, and Siwei Lyu. Integrating audio-visual features for multimodal deepfake detection, 2023.

[67] Xiangtao Meng, Li Wang, Shanqing Guo, Lei Ju, and Qingchuan Zhao. Ava: Inconspicuous attribute variation-based adversarial attack bypassing deepfake detection, 2023.

[68] Weijie Wang, Zhengyu Zhao, Nicu Sebe, and Bruno Lepri. Turn fake into real: Adversarial head turn attacks against deepfake detection, 2023.

[69] Nicolò Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. Video face manipulation detection through ensemble of cnns, 2020.

[70] Hyeonseong Jeon, Youngoh Bang, and Simon S. Woo. Fdftnet: Facing off fake images using fake detection fine-tuning network, 2020.

[71] Brandon B. G. Khoo, Chern Hong Lim, and Raphael C. W. Phan. Transferable class-modelling for decentralized source attribution of gan-generated images, 2022.

[72] Yuhang Lu and Touradj Ebrahimi. Impact of video processing operations in deepfake detection, 2023.

[73] Matthieu Delmas and Renaud Seguier. Latentforensics: Towards frugal deepfake detection in the stylegan latent space, 2024.

[74] Sifat Muhammad Abdullah, Aravind Cheruvu, Shravya Kanchi, Taejoong Chung, Peng Gao, Murtuza Jadliwala, and Bimal Viswanath. An analysis of recent advances in deepfake image detection in an evolving threat landscape, 2024.

[75] Zhikan Wang, Zhongyao Cheng, Jiajie Xiong, Xun Xu, Tianrui Li, Bharadwaj Veeravalli, and Xulei Yang. A timely survey on vision transformer for deepfake detection, 2024.

[76] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images, 2019.

[77] Zhiqing Guo, Gaobo Yang, Jiyou Chen, and Xingming Sun. Fake face detection via adaptive manipulation traces extraction network, 2020.

[78] Paloma Cantero-Arjona and Alfonso Sánchez-Macián. Deepfake detection and the impact of limited computing capabilities, 2024.

[79] Oliver Giudice, Luca Guarnera, and Sebastiano Battiato. Fighting deepfakes by detecting gan dct anomalies, 2021.

[80] Alex Norlin. The effectiveness of knowledge distillation methods for real-time deepfake solutions, 2024.

[81] Shahzeb Naeem, Ramzi Al-Sharawi, Muhammad Riyyan Khan, Usman Tariq, Abhinav Dhall, and Hasan Al-Nashash. Real, fake and synthetic faces – does the coin have three sides?, 2024.

[82] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. Benchmarking deepart detection, 2023.

[83] Dongze Li, Wei Wang, Hongxing Fan, and Jing Dong. Exploring adversarial fake images on face manifold, 2021.

**Disclaimer:**

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.