Deepfake Dataset Privacy: A Comprehensive Survey and Framework for Safer Sharing*

A-Young Jeon, Yu-ran Jeon and Il-Gu Lee†

Sungshin Women's University, Seoul, 02844, Korea
{qiena11jeon, cseyrj, iglee19}@gmail.com

Abstract

Recent deepfake research has largely prioritized technical performance while paying insufficient attention to privacy in dataset construction and use. This study therefore surveys publicly available deepfake generation and detection datasets from a privacy perspective. Our analysis shows that of 22 generation datasets only 2 explicitly address privacy, and of 41 detection datasets only 8 incorporate privacy considerations. Most datasets were obtained via web crawling or platform-based collection and rely on post-hoc opt-out mechanisms. We identify three principal privacy risks: (i) degraded generalization due to dataset distribution bias, (ii) secondary and tertiary privacy infringements that arise during dataset derivation and re-processing, and (iii) structural risk resulting from the irreversibility of released data. To mitigate these harms, we propose a preemptive protection framework to be applied at the initial data-distribution stage. The proposed framework aims to reduce privacy risks during downstream dataset use and to promote safer, privacy-aware practices for constructing and sharing deepfake datasets.

Keywords: Deepfake Datasets, Privacy Protection, Data Governance

1 Introduction

The rapid advancement of generative models has transformed the synthetic media landscape, significantly enhancing the realism, diversity, and accessibility of deepfakes. Advances in generative AI technologies such as Generative Adversarial Networks (GANs)[1, 2] and diffusion-based architectures[3-5] have made it easy to create high-quality synthetic content that is difficult to distinguish from authentic media [6]. Deepfake generation using deep learning raises serious societal concerns due to its potential misuse for generating malicious content and its difficulty in distinguishing from real content[7, 8]. A representative manipulation technique, FaceSwap[9], replaces the target subject's face with the facial identity of a source individual while preserving the subject's expressions and head poses. Since faces are high-risk biometric information directly linked to personal identity, sophisticated deepfakes can create illusions of non-existent individuals or actions, potentially leading to political, social, financial, and legal consequences[10].

* Proceedings of the 9th International Conference on Mobile Internet Security (MobiSec'25), Article No. W7, December 16-18, 2025, Sapporo, Japan.

† Corresponding author

To address these risks, various deepfake detection methods have been proposed in the image and video domains. Image-based detection has primarily focused on spatial differences (local noise, boundary artifacts, fine-grained texture details, etc.) arising during the manipulation process, while video detection has centered on identifying inconsistencies between adjacent frames over time. Furthermore, extensive research has been conducted on inconsistencies between audio and visual modalities in deepfake videos, demonstrating that leveraging multiple modalities is more effective than relying solely on audio or visual features for deepfake video detection [6].

However, the generalization performance of current deepfake detection methods remains constrained by dataset distribution bias. Widely used public datasets (DFDC[11], FakeAVCeleb[12], FF++[13], etc.) often exhibit stark visual quality differences compared to real deepfake content circulating online, due to factors like low-resolution synthesis, noticeable boundaries and color mismatches, alignment errors, and limited conditions. To overcome these limitations, datasets reflecting high resolution, large scale, and the latest generation techniques are continuously proposed. However, most research focuses solely on improving detection performance while overlooking privacy issues during the data collection process. In particular, facial datasets are often collected from the internet via web crawling or platforms like YouTube. Some datasets mentioned that if you do not wish to be used your data included, you should request its deletion as a post-processing measure. Considering that deepfake detection research aims to protect individual rights, the potential for privacy infringement during the data collection stage must also be considered.

This study systematically investigates the current state of privacy considerations in the latest deepfake generation and detection datasets. Based on these findings, it proposes a framework introducing invisibility watermarking techniques as a preemptive protective measure. This approach seeks to achieve both the technological advancement of deepfake detection research and the protection of personal information simultaneously.

The contributions of this study are as follows:

- By investigating and analyzing datasets used in deepfake generation and detection, we analyzed the latest trends in datasets.
- We analyzed the limitations of previous deepfake research, where privacy concerns such as portrait rights and consent procedures were not sufficiently considered during dataset creation and utilization.
- To address privacy issues in deepfakes, we propose watermark-based deepfake generation and detection techniques, suggesting a new research direction for datasets that can satisfy personal information protection.

This paper is structured as follows. Section 2 describes deepfake technology, Section 3 analyzes trends in deepfake dataset generation and detection, Section 4 explains watermark-based deepfake technology to address privacy issues in deepfake datasets, and finally, Section 5 concludes.

2 Background

Research on deepfakes is structured around two core pillars: Generation and Detection. As the latest generation techniques advance, corresponding deepfake detection datasets have been being built more actively.

2.1 Deepfake Generation

Deepfake generation technology creates new synthetic content by combining conditional information (such as images, audio, or text) with target images or videos. Datasets for training generative models are primarily composed of real images/videos, and the main generation tasks are as follows.

- **Face Swapping**: A manipulation technique that replaces the identity information of the target face with that of the source person's face. During this process, ID-irrelevant attributes such as skin color or facial expressions retain the original characteristics of the target face. Initially, face swapping was achieved through video compositing techniques such as boundary blending and brightness adjustment, or by constructing facial parameters based on 3D Morphable Models (3DMM). However, recent approaches employ deep generative models such as GANs and diffusion models, enabling more natural and indistinguishable face swapping results.
- **Face Reenactment**: The core objective of face reenactment is to manipulate facial movements. This technique reproduces the target face's movements by extracting motion information—such as facial expressions, gaze, and head poses—from driving images or videos, while preserving the target's identity. Existing approaches can be broadly categorized into four main types: 3DMM-based modeling, landmark matching, face feature decoupling, and self-supervised learning**.** Among these, face feature decoupling and self-supervised learning methods achieve realistic face reproduction that preserves identity by separating identity and expression attributes.
- **Talking Face Generation**: A temporal extension task that synthesizes realistic talking video of a target person by generating lip movements, facial poses, expressions, emotions, and synchronized speech. Existing approaches can be categorized into several paradigms. Audio/Text Driven methods focus on aligning lip movements and facial dynamics with the corresponding audio or textual content. Multimodal Conditioning approaches incorporate additional modalities—such as emotional cues, visual context, or textual semantics—to enhance audiovisual coherence. Diffusion-based methods enable highly controllable and high-resolution video synthesis through fine-grained encoding and reconstruction processes. Furthermore, 3D model-based techniques leverage facial geometry and motion prior to achieve accurate and consistent head movements across frames.
- **Facial Attribute Editing**: A manipulates technique that modifies semantically important facial attributes (e.g., age, expression, skin tone) according to individual preferences or specific task requirements. The main challenge lies in achieving comprehensive editing, which involves effectively isolating different facial attributes while preserving unrelated attributes to ensure consistency in non-target facial information. Recent approaches explore text-based editing by introducing text-to-style mappings, enabling direct encoding of facial features into the latent space of models like StyleGAN and facilitating cross-modal alignment between text and facial representations. Moreover, diffusion-based frameworks have significantly enhanced controllability and image fidelity in fine-grained facial attribute manipulation.

2.2 Deepfake Detection

This dataset for training deepfake detection models includes both original videos and deepfake videos created using the latest generation techniques. Detection models perform binary classification to distinguish between these two classes or detect manipulated regions. Deepfake detection research primarily falls into three categories based on how manipulation traces are captured.

- **Spatial Analysis based Detection**: Detects spatial inconsistencies within a single frame that occur during the manipulation process. Key detection targets include local noise patterns, blending artifacts at facial boundaries, color and lighting inconsistencies, and fine-grained texture anomalies.
- **Temporal Analysis based Detection**: Analyzes temporal consistency between consecutive frames in a video. Since deepfakes are often generated frame-by-frame independently, unnatural discontinuities appear in adjacent frames regarding facial landmark positions, expression changes, and head pose transitions.
- **Multimodal Analysis based Detection**: Verifies consistency between audio and visual modalities. For talking face deepfakes, detection is possible through lip-sync mismatches between speech and mouth movements, or discrepancies between vocal characteristics and facial motions. Multimodal fusion models can effectively identify deepfakes that are difficult to detect using a single modality alone.

3 Datasets

This section analyzes the datasets used for deepfake generation and detection research from a privacy perspective. Over 60 related datasets have been released from 2008 to 2025. Initially, small-scale image-centric face recognition datasets like LFW[14] and CelebA[15] dominated. Recently, multimodal datasets combining hundreds of thousands of video clips with text, audio, and video (e.g., MM-Vox[16], CelebV-Text[17], DeepSpeak[18]) have been continuously released. Notably, datasets for training generative models tend to rely heavily on large-scale web crawling data to ensure high resolution and diversity.

To quantitatively assess the level of privacy consideration in deepfake face datasets, we systematically reviewed various materials for each dataset, including official documentation, research papers, GitHub repositories, and distribution websites. Based on this review, privacy considerations were evaluated according to four criteria: (i) whether participants were recruited with explicit consent, (ii) whether the data were self-generated through recordings of consenting individuals or collected from online platforms, (iii) whether the license type and usage restrictions were specified, and (iv) whether an opt-out or deletion mechanism was provided.

When no explicit mention of consent or privacy procedures was found in the available materials, datasets collected via web crawling or YouTube-based methods were regarded as non-consensual and, therefore, lacking privacy consideration. Conversely, datasets produced through studio recordings with hired actors were classified as privacy-considerate, as they were created under conditions of informed consent. Although some datasets stated that users could contact the creators to request data deletion, such post hoc opt-out mechanisms were considered insufficient for ensuring meaningful privacy protection.

3.1 Deepfake Generation Dataset

Deepfake generation datasets consist of source facial images or videos used to train generative models such as GANs and Diffusion Models. These datasets serve as foundational data for training deepfake generation models to perform various generative tasks (e.g., face swapping, reenactment, talking face generation, attribute editing). Table 1 summarizes generative datasets published from 2008 to 2024.

Dataset	Venue	Modality	Source	Privacy consideration
LFW [14]	WFRLI'08	Image	Internet	X
Multi-PIE [19]	BMVC'10	Image	Actors	O
VGGFace [20]	BMVC'15	Image	Internet	X
VGGFace2 [21]	FG'18	Image	Google Image Search	X
CelebA [15]	ICCV'15	Image	Internet	X
CelebA-HQ [22]	ICLR'18	Image	CelebA	X
LRS2 [23]	TPAMI'18	Audio-Video	BBC	X
LRS3 [24]	arXiv'18	Audio-Video	TED/TEDx(YouTube)	X
VoxCeleb1 [25]	Interspeech'17	Audio-Video	YouTube	X
VoxCeleb2 [26]	Interspeech'18	Audio-Video	YouTube	X
FFHQ [27]	CVPR'19	Image	Flickr(CC License)	X
MEAD [28]	ECCV'20	Audio-Video	Actors	O
CelebAMask-HQ [29]	CVPR'20	Image	CelebA-HQ	X
CelebAText-HQ [30]	MM'21	Image	CelebA-HQ	X
MM CelebA-HQ [31]	CVPR'21	Image	CelebA-HQ	X
HDTF [32]	CVPR'21	Audio-Video	YouTube	X
Talking Head-1KH [33]	CVPR'21	Audio-Video	YouTube	X
CelebV-HQ [34]	ECCV'22	Video	Internet	X
MM-Vox [16]	CVPR'22	Audio+Video+Text	VoxCeleb	X
CelebV-Text [17]	CVPR'23	Video+Text	CelebV-HQ	X
VGGFace2-HQ [35]	TPAMI'24	Image	VGGFace2	X
Arc2Face [36]	ECCV'24	Image	WebFace42M	X

Table 1 : Comparison of deepfake generation datasets

In the early stages, single image modalities were predominant. LFW[14] served as a benchmark for early face recognition and generation research using face images collected from the internet. VGGFace2[21] is a face recognition dataset built by the Oxford Visual Geometry Group (VGG), which collected face images of celebrities, public figures, actors, politicians, etc., via Google Image Search and enhanced quality through a manual review process.

CelebA[15] collected 202,599 celebrity face images of 10,177 individuals from the internet, providing 40 attribute labels per image (e.g., wearing glasses, beard presence, hair color, smiling). This rich annotation became the foundation for facial feature analysis and conditional face generation research, and it is currently the most widely used foundational dataset in deepfake research.

Multimodal datasets, combining audio and visual data, began to be constructed starting in 2017. LRS2[23] established the foundation for lip-sync research using audio-video data collected from BBC broadcast content. Additionally, LRS3[24] collects TED and TEDx videos, providing a dataset essential for large-scale lip-reading, audio-video synchronization, and speech-video generation research. VoxCeleb1/2[25, 26] is a large-scale speaker recognition dataset based on YouTube interview videos of celebrities. VoxCeleb2, with approximately 2,000 hours of data and improved racial diversity, has become a key benchmark for deepfake research.

FFHQ (Flickr-Faces-HQ)[27] is a high-quality facial image dataset created by NVIDIA for training GANs. It consists of 70,000 high-resolution (1024×1024) facial images collected from Flickr-CC. It features high diversity in age, ethnicity, and background, and its automated pipeline systematically

handles face alignment, quality filtering, and duplicate removal, establishing it as a standard for GAN and Diffusion training.

CelebAMask-HQ [29] adds precise mask annotations for eyes, nose, mouth, hair, and skin to CelebA-HQ face images, enabling research on local region editing and forgery. Additionally, various datasets such as HDTF [32], Talking Head-1KH [33], CelebV-HQ [34], MM-Vox [16], CelebV-Text [17], and Arc2Face [36] have been released, covering high-resolution talking heads, video+text alignment, and ID-conditional generation.

From a privacy perspective, Table 1 shows that among the 22 major generative datasets released from 2008 to 2024, only two datasets protected privacy. Notably, Multi-PIE [19] and MEAD [28] built consent-based datasets by hiring actors in a controlled studio environment and capturing diverse poses, expressions, lighting, and other conditions. In contrast, most datasets were collected without consent via YouTube, the internet, and web crawling. Repurposing existing datasets further amplified the privacy issues of the original data through secondary and tertiary reuses. Notably, CelebA-based derivative datasets represent a single instance of unauthorized collection being reused long-term across multiple studies.

3.2 Deepfake Detection Dataset

The deepfake detection dataset was designed to include both manipulated content and original content in a balanced manner, enabling classification and learning of manipulation locations.

Initially, research focused primarily on face recognition, attribute classification, and face generation, with most datasets containing only images of real people. However, as technology advanced, datasets incorporating manipulated images became widespread. Recently, there has been a growing trend toward multimodal datasets encompassing audio-video mismatches and text-conditioned synthesis. Table 2 summarizes detection datasets released between 2017 and 2025.

Deepfake detection research began in earnest in 2017. UADFV[37] demonstrated the feasibility of head pose mismatch detection using 49 original and 49 deepfake videos collected from YouTube, while Deepfake-TIMIT[38] provided 320 videos created by applying early synthesis techniques to VidTIMIT.

Large-scale benchmark dataset construction began in 2019. FaceForensics++[13] is widely used as a benchmark today. It generates a total of 5,000 videos by applying four manipulation techniques—Deepfakes, Face2Face, FaceSwap, and NeuralTextures—to 1,000 original videos collected from YouTube, and provides three levels of compression: raw, c23, and c40. The DFDC (Deepfake Detection Challenge)[11], led by Facebook (now Meta), provides 104,500 videos. This dataset was created by applying eight types of synthesis to 23,654 original videos recorded by 3,426 actors via self-recording. This was the largest deepfake detection dataset at the time, aiming to evaluate generalization performance by including participants of diverse ages, backgrounds, races, and ages.

Dataset	Venue	Modality	Source	Privacy consideration
SwapMe and FaceSwap [39]	CVPRW'17	Image	Self-Collection	X
Deepfake-TIMIT [38]	arXiv'18	Video	VidTIMIT Dataset	O
UADFV [37]	ICASSP'19	Video	YouTube	X
FaceForensics++ [13]	ICCV'19	Video	YouTube	X
DFD (Google) [40]	-	Video	Self-Recording	O
FakeSpotter [41]	arXiv'19	Image	CelebA, FFHQ, FF++, DFDC, Celeb-DF	X
Celeb-DF (v1) [10]	CVPR'20	Video	YouTube	X
Celeb-DF (v2) [10]	CVPR'20	Video	YouTube	X

DFFD [42]	CVPR'20	Video/Image	FFHQ, CelebA, FaceForensics++	X
APFDD [43]	IJCNN'20	Image	CelebA	X
Deeper Forensic (-1.0) [44]	CVPR'20	Video	Self-Recording	O
DFDC-P(Priview) [45]	arXiv'19	Audio-Video	Self-Recording	O
DFDC (Facebook) [11]	arXiv'20	Audio-Video	Self-Recording	O
Wild Deepfake [46]	MM'20	Video	Internet	X
Deepfake MNIST+ [47]	ICCVW'21	Video	VoxCeleb1, ADFES	X
ForgeryNet [48]	CVPR'21	Video	CREMA-D, RAVDESS, VoxCeleb2, AVSpeech	X
DF-W [49]	arXiv'21	Video	YouTube, Bilibili	X
KoDF [50]	ICCV'21	Video	Self-Recording	O
FakeAVCeleb [12]	arXiv'21	Audio-Video	VoxCeleb2	X
FFIW-10K [51]	CVPR'21	Video	YouTube	X
DFGC [52]	IJCB'21	Image	Celeb-DF v2	X
OpenForensics [53]	ICCV'21	Image	Google Open Images	X
LAV-DF [54]	DICTA'22	Audio-Video	VoxCeleb2	X
DF^3 [55] (DeepFakeFaceForensics)	TMM'23	Image	YouTube	X
Deephy [56]	IJCB'22	Video	YouTube	X
DeepFakeFace [57]	arXiv'23	Image	IMDB-WIKI	X
DF-Platter [58]	CVPR'23	Video	Youtube, Celebrity	X
DGM4 [59]	CVPR'23	Text+Image	VisualNews	X
AV-Deepfake1M [60]	MM'24	Audio-Video	Voxceleb2	X
DiFF [61]	MM'24	Image	VoxCeleb2, CelebA	X
JDB-Face [62]	arXiv'24	Image	JourneyDB	X
DFDB-Face [62]	arXiv'24	Image	DiffusionDB	X
DF40 [63]	NeurIPS'24	Video/Image	FF++, Celeb-DF, CelebA	X
PolyGlotFake [64]	arXiv'24	Audio-Video	YouTube, Internet	X
DeepSpeak v1.0 [18]	arXiv'24	Audio-Video	Self-Recording	O
DeepSpeak v2.0 [18]	arXiv'24	Audio-Video	Self-Recording	O
DeepFake-eval-2024 [65]	arXiv'25	Audio-Video	X, direct upload, TikTok, Instagram, Youtube	X
TalkingHeadBench [66]	arXiv'25	Audio-Video	FFHQ, CelebV-HQ	X
VLF [67]	arXiv'25	Text+Image	CelebAMask-HQ, FF++, FFHQ, CelebA-HQ	X
Celeb-DF++ [68]	arXiv'25	Audio-Video	Celeb-DF	X
HiDF [6]	KDD'25	Video/Image	Image - CelebA-HQ, FFHQ Video - YouTube, FakeAVCeleb	X

Table 2 : Comparison of deepfake detection datasets

Subsequently released datasets began to evolve by combining existing datasets. DFFD[42] integrated FFHQ, CelebA, and FF++, containing a total of 58,703 images, 240 videos, and various generation techniques. FakeSpotter[41] was constructed by integrating CelebA, FFHQ, FF++, DFDC, and Celeb-DF to enable focused evaluation of cross-dataset generalization.

Furthermore, specialized large-scale datasets targeting specific manipulation types and multi-task learning began to emerge. ForgeryNet[48] integrates CREMA-D, VoxCeleb2, and AVSpeech (or similar AV corpora) data, containing approximately 2.9M images and 221,247 videos, and includes 15 manipulation types. This rigorously tests various manipulation methods and domain generalization problems. FakeAVCeleb[12] provides a multimodal detection benchmark based on VoxCeleb2, encompassing both audio and video manipulations (lip-syncing, voice morphing, face swapping). FFIW-10K[51] complements this diversity in real-world settings by collecting over 10,000 field-based videos and data from over 2,000 individuals from YouTube.

In response to advances in audiovisual synthesis technology, datasets targeting AV mismatch manipulation are also gradually expanding. LAV-DF[54] provides a large-scale audiovisual forgery sample set based on VoxCeleb2, including lip-syncing, voice conversion, and face swapping. Furthermore, with the rise of diffusion-based generation techniques, DeepFakeFace[57] incorporates diffusion-based synthetic images using IMDB-WIKI as source data, while DFDB-Face[62] utilizes DiffusionDB to construct a large-scale diffusion forgery benchmark. DGM4[59] utilizes VisualNews to provide news domain-specific text–image forgery data, expanding the scope of text-condition-based synthesis detection.

According to Table 2, among the 41 major detection datasets released from 2018 to 2025, only eight datasets considered privacy protection measures.

- Deepfake-TIMIT [38]: Uses VidTIMIT Dataset (pre-consent data)
- DFD (Google) [40]: Hires professional actors
- DeeperForensics-1.0 [44]: Hires 100 actors
- DFDC-P(Preview)[45]/DFDC(Facebook)[11]: Actor hiring and participant self-recording
- KoDF [50]: Korean participant self-recording
- DeepSpeak v1.0/DeepSpeak v2.0 [18]: Participant consent-based recording

Most datasets collect data without individual consent, and the repurposing of previously collected unauthorized datasets results in a situation of double privacy infringement. Large-scale projects like the Deepfake Detection Challenge and Deep Forensics Challenge, conducted from 2019 to 2020, showed a temporary increase in privacy considerations by directly filming their datasets. However, since 2021, most deepfake datasets have relied on recycling existing datasets or unauthorized collection methods to obtain data, further exacerbating privacy issues.

A clear example of this can be founded on the Celeb-DF GitHub page, which explicitly notes that individuals concerned about their identity being included in the dataset can request the removal of corresponding information. However, this post-hoc opt-out policy is inherently limited: it cannot recall copies already redistributed to third parties. It also creates a structural risk in which the potential for privacy infringement and misuse continually increases through data derivation. Once data is made public, it cannot be recalled, and derivative outputs created from it can continue to circulate without authorization. The unauthorized use of celebrities' images is not justified merely by their fame, and the mere fact that data appear on a publicly accessible website does not negate privacy obligations associated with their collection and distribution.

Moreover, while attack techniques are becoming increasingly sophisticated over time, defense technologies inevitably experience delays in development when responding to new manipulation types. These technical limitations make it difficult to respond to deepfake attacks in real time. In particular, subsequent opt-out methods related to datasets reveal several limitations. Publicly distributed copies of data are effectively impossible to recall. Individuals often struggle to identify the channels through which their images have been disseminated. Moreover, the persistent privacy violations arising from the production of derivative data are fundamentally difficult to resolve. This situation underscores the imperative need for ethical discussions and strengthened protective frameworks in response to the advancement of deepfake technology.

4 Proposed model

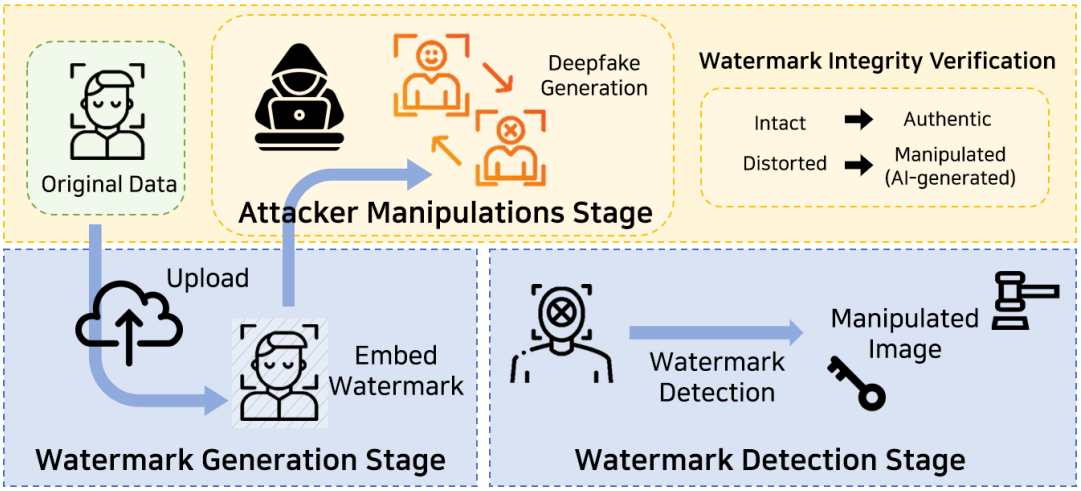

Figure 1: Watermarked Image Deepfake Framework

To address the persistent issue of detection gaps stemming from delays in technological advancements, a proactive approach is essential to move away from the current dependence on post-detection methods. This study introduces a proactive framework that embeds digital watermarking at the initial distribution stage of media content. The framework embeds invisible watermarks into facial images and videos, creating a mechanism to rapidly detect tampering or unlawful distribution by extracting and verifying the watermarks during the distribution process. The ultimate goal is to safeguard privacy, copyright and provenance.

Figure 1 provides a schematic overview of the process for generating and detecting watermark-based deepfake content, which consists of three main stages. Watermark Generation Stage, Attacker Manipulations Stage, and Watermark Detection Stage. First, in the Watermark Generation Stage, an invisible digital watermark is embedded into the original facial image or video at the time of upload, before its public release. The watermarked data are then uploaded to a shared or online platforms, resulting in a watermarked version of the content. These watermarked images are subsequently aggregated to form a public dataset. When deepfakes are generated using this dataset, the embedded watermark remains present but may become partially distorted due to manipulation. This characteristic enables the tracing of content origin and detection of tampering by verifying watermark integrity, thereby identifying unauthorized distributions and manipulated data.

Next, during the Attacker Manipulations Stage, malicious users may employ the watermarked dataset to generate manipulated or AI-synthesized content through deepfake generation techniques such

as face swapping, reenactment, or attribute editing. These manipulations inevitably alter or distort the embedded watermark, leaving detectable traces that can subsequently be used for identification and verification.

Finally, in the Watermark Detection Stage, the integrity of the embedded watermark is examined in the suspected media. If the watermark remains intact, the content is verified as authentic; if the watermark is distorted or missing, the content is identified as manipulated or AI-generated. This verification process enables the reliable detection of unauthorized modifications and ensures the traceability of media provenance.

5 Discussion and Limitations

The rapid advancement of deepfake technology raises serious ethical and social concerns, as it can be exploited for spreading misinformation, identity fraud, and creating content without consent. This paper suggests that we need to move beyond merely improving technical methods for dataset generation and detection. Instead, we must raise awareness about potential misuse and promote research into effective mitigation strategies. As one such approach, watermark-based image manipulation detection technology is proposed as a step toward building a safer and more trustworthy AI-based media environment.

Limitations. Although this study conducted a comprehensive analysis of existing deepfake datasets, the methodology for labeling dataset privacy is insufficiently specified. Our current evaluation relied on a binary classification system, categorizing datasets solely by the presence or absence of privacy considerations. This approach risks oversimplifying the complex range of privacy protection levels. To ensure transparency and reproducibility of the assessment results, future work will provide a clearly defined evaluation rubric. Specifically, we plan to establish a quantitative scoring framework grounded in four key criteria: (i) explicit consent or actor recruitment, (ii) data origin, (iii) license type, explicit usage purposes, and stated restrictions, and (iv) the availability of opt-out or deletion mechanisms. This rubric will enable privacy sensitivity to be expressed on a graded scale, allowing differentiation between weak and strong levels of privacy intrusion.

Discussion. Although this study primarily focused on the technical and ethical dimensions of deepfake dataset privacy, a more comprehensive evaluation should also consider whether current dataset collection and sharing practices comply with existing legal and regulatory frameworks. As the creation and distribution of face data often involve identifiable individuals, potential conflicts may arise under both domestic and international privacy laws. Therefore, a systematic legal–regulatory analysis is essential to determine whether the collection, sharing, and utilization of such datasets are consistent with recognized privacy principles and rights of data subjects. To strengthen the reliability and applicability of the proposed framework, future work should establish a clear legal foundation by mapping the privacy evaluation criteria to authoritative governance instruments. At the international level, comparative reference to frameworks such as the EU General Data Protection Regulation (GDPR)—particularly Articles 6 and 17 concerning lawful processing and erasure—the California Consumer Privacy Act (CCPA, Section 1798.105), the OECD Privacy Guidelines, the ISO/IEC 29100 Privacy Framework, and the NIST Privacy Framework 1.0 can provide structured standards for assessing consent, lawful use, and data subject rights.

At the domestic level, incorporating provisions from the Personal Information Protection Act (PIPA) of South Korea and related enforcement decrees would enable context-specific evaluation of consent mechanisms, purpose limitation, and opt-out procedures in accordance with national legal obligations.

By integrating these comparative legal perspectives, future work can assess not only whether datasets are ethically and technically sound, but also whether they may violate or fall short of compliance with existing national or international privacy obligations. Such analysis is expected to

demonstrate the international importance of privacy governance in datasets, potentially fostering legal accountability.

6 Conclusion

This paper demonstrates, through an analysis of the recent deepfake datasets, that while recent deepfake-related research has primarily focused on improving technical performance, there is a significant lack of consideration for privacy issues during the dataset construction and utilization process. Among 22 deepfake generation datasets, only two explicitly considered privacy. Among 41 detection datasets, only 8 reflected privacy considerations. Most datasets rely on post-hoc opt-out mechanisms after web crawling and platform-based data collection, revealing persistent structural risks: (i) downstream secondary and tertiary infringements can accumulate during derivation and reprocessing, and (ii) once data are disclosed, effective recall is difficult, limiting the impact of ex-post measures. These conditions call for a more fundamental remedy.

To address these challenges, we propose a preemptive protection framework applied at the initial stage of data distribution. The core of this framework lies in its ability to detect unauthorized manipulations throughout the data distribution process based on watermark traces, thereby verifying potential infringements of copyright and personal information. As technology continues to advance, the obligation to explicitly label AI-generated content is likely to be strengthened, while malicious actors will simultaneously develop techniques to evade such detection. In this context, the advancement of AI detection technologies becomes an increasingly critical element.

The proposed approach aims to proactively mitigate potential risks associated with dataset utilization and enhance detection reliability, while complementing the limitations of existing post-hoc detection methods. Future research will focus on presenting experimental results to validate the feasibility of the proposed framework and conducting an in-depth analysis from legal and regulatory perspectives. In particular, experimental verification of watermark durability, robustness in real conversion environments, false positive and false negative rates, and computational overhead aims to demonstrate that these watermarking techniques can effectively detect AI-generated manipulations and mitigate associated privacy risks. In addition, by analyzing the degree of infringement across different stages, this study seeks to derive more reliable and interpretable results. Ultimately, this research holds significant academic value as it proposes a proactive protection framework that addresses the vulnerabilities of existing technologies and provides a new direction for deepfake-related studies.

References

[1] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401–4410).

[2] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8110–8119).

[3] Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. (2023). Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*. Retrieved from arXiv:2311.15127

[4] Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., & Dai, B. (2023). AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*. Retrieved from arXiv:2307.04725

[5] Liu, R., Ma, B., Zhang, W., Hu, Z., Fan, C., Lv, T., Ding, Y., & Cheng, X. (2024). Towards a simultaneous and granular identity-expression control in personalized face generation. *arXiv preprint arXiv:2401.01207*. Retrieved from arXiv:2401.01207

[6] Kang, C., Jeong, S., Lee, J., Choi, D., Woo, S. S., & Han, J. (2025). HiDF: A Human-Indistinguishable Deepfake Dataset. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2* (pp. 5527-5538).

[7] Li, C., Wang, L., Ji, S., Zhang, X., Xi, Z., Guo, S., & Wang, T. (2022). Seeing is living? rethinking the security of facial liveness verification in the deepfake era. In *31st USENIX Security Symposium (USENIX Security 22)*.

[8] Tariq, S., Abuadbba, A., & Moore, K. (2023). Deepfake in the metaverse: Security implications for virtual gaming, meetings, and offices. In *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes, WDC '23* (pp. 16–19). Association for Computing Machinery.

[9] Faceswap. (2018). Retrieved from GitHub: https://github.com/MarekKowalski/FaceSwap/

[10] Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *CVPR*.

[11] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Canton Ferrer, C. (2020). The deepfake detection challenge (DFDC) dataset. *arXiv preprint arXiv:2006.07397*.

[12] Khalid, H., Tariq, S., Kim, M., & Woo, S. S. (2021). FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*.

[13] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1–11).

[14] Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition*.

[15] Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *ICCV*.

[16] Han, L., Ren, J., Lee, H. Y., Barbieri, F., Olszewski, K., Minaee, S., Metaxas, D., & Tulyakov, S. (2022). Show me what and tell me how: Video synthesis via multimodal conditioning. In *CVPR*.

[17] Yu, J., Zhu, H., Jiang, L., Loy, C. C., Cai, W., & Wu, W. (2023). CelebV-Text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14805-14814).

[18] Barrington, S., Bohacek, M., & Farid, H. (2024). DeepSpeak Dataset v1. 0. *arXiv e-prints, arXiv-2408*.

[19] Moore, S., & Bowden, R. (2010). Multi-view pose and facial expression recognition. In *BMVC*.

[20] Parkhi, O., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *BMVC*.

[21] Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). VGGface2: A dataset for recognising faces across pose and age. In *FG*.

[22] Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of gans for improved quality, stability, and variation. In *ICLR*.

[23] Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2018). Deep audio-visual speech recognition. *TPAMI*.

[24] Afouras, T., Chung, J. S., & Zisserman, A. (2018). LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv*.

[25] Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: a large-scale speaker identification dataset. In *Interspeech*.

[26] Chung, J. S., Nagrani, A., & Zisserman, A. (2018). VoxCeleb2: Deep speaker recognition. In *Interspeech*.

[27] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *CVPR*.

[28] Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., & Loy, C. C. (2020). MEAD: A large-scale audiovisual dataset for emotional talking-face generation. In *ECCV*.

[29] Lee, C. H., Liu, Z., Wu, L., & Luo, P. (2020). MaskGAN: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5549-5558).

[30] Sun, J., Li, Q., Wang, W., Zhao, J., & Sun, Z. (2021). Multicaption text-to-face synthesis: Dataset and algorithm. In *ACM MM*.

[31] Xia, W., Yang, Y., Xue, J. H., & Wu, B. (2021). TediGAN: Text-guided diverse face image generation and manipulation. In *CVPR*.

[32] Zhang, Z., Li, L., Ding, Y., & Fan, C. (2021). Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3661-3670).

[33] Wang, T. C., Mallya, A., & Liu, M. Y. (2021). One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*.

[34] Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., & Loy, C. C. (2022). CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*.

[35] Chen, X., Ni, B., Liu, Y., Liu, N., Zeng, Z., & Wang, H. (2023). SimSwap++: Towards faster and high-quality identity swapping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(1), 576-592.

[36] Paraperas Papantoniou, F., Lattas, A., Moschoglou, S., Deng, J., Kainz, B., & Zafeiriou, S. (2024). Arc2Face: A Foundation Model for ID-Consistent Human Faces. In *ECCV*.

[37] Yang, X., Li, Y., & Lyu, S. (2019). Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8261–8265). IEEE.

[38] Korshunov, P., & Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection. *arXiv*.

[39] Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2017). Two-stream neural networks for tampered face detection. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 1831-1839). IEEE.

[40] Google Research. (2019). Contributing data to deepfake detection research. Retrieved from Google Research Blog: https://blog.research.google/2019/09/contributing-data-to-deepfake-detection.html

[41] Wang, R., Juefei-Xu, F., Ma, L., Xie, X., Huang, Y., Wang, J., & Liu, Y. (2019). FakeSpotter: A simple yet robust baseline for spotting AI-synthesized fake faces. *arXiv preprint arXiv:1909.06122*.

[42] Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020). On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[43] Gandhi, A., & Jain, S. (2020). Adversarial perturbations fool deepfake detectors. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8).

[44] Jiang, L., Li, R., Wu, W., Qian, C., & Loy, C. C. (2020). DeeperForensics1.0: A large-scale dataset for real-world face forgery detection. In *CVPR* (pp. 2889–2898).

[45] Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2019). The deepfake detection challenge (DFDC) preview dataset. *arXiv*.

[46] Zi, B., Chang, M., Chen, J., Ma, X., & Jiang, Y. G. (2020). WildDeepfake: A challenging real-world dataset for deepfake detection. In *ACM MM*.

[47] Huang, J., Wang, X., Du, B., Du, P., & Xu, C. (2021). DeepFake MNIST+: A DeepFake facial animation dataset. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)* (pp. 1973–1982).

[48] He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., Sheng, L., Shao, J., & Liu, Z. (2021). ForgeryNet: A versatile benchmark for comprehensive forgery analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4360–4369). CVPR.

[49] Pu, J., Mangaokar, N., Kelly, L., Bhattacharya, P., Sundaram, K., Javed, M., Wang, B., & Viswanath, B. (2021). Deepfake videos in the wild: Analysis and detection. *arXiv:2103.04263*.

[50] Kwon, P., You, J., Nam, G., Park, S., & Chae, G. (2021). KODF: A large-scale Korean deepfake detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10744–10753).

[51] Zhou, T., Wang, W., Liang, Z., & Shen, J. (2021). Face forensics in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5778–5788).

[52] Peng, B., Fan, H., Wang, W., Dong, J., Li, Y., Lyu, S., Li, Q., Sun, Z., Chen, H., Chen, B., et al. (2021). DFGC 2021: A deepfake game competition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)* (pp. 1–8). IEEE.

[53] Le, T. N., Nguyen, H. H., Yamagishi, J., & Echizen, I. (2021). OpenForensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10117-10127).

[54] Cai, Z., Stefanov, K., Dhall, A., & Hayat, M. (2022). Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In *DICTA*.

[55] Ju, Y., Jia, S., Cai, J., Guan, H., & Lyu, S. (2023). GLFF: Global and local feature fusion for AI-synthesized image detection. *IEEE Transactions on Multimedia*.

[56] Narayan, K., Agarwal, H., Thakral, K., Mittal, S., Vatsa, M., & Singh, R. (2022). DeepHy: On deepfake phylogeny. In *IJCB*.

[57] Song, H., Huang, S., Dong, Y., & Tu, W. W. (2023). Robustness and generalizability of deepfake detection: A study with diffusion models. *arXiv preprint arXiv:2309.02218*.

[58] Narayan, K., Agarwal, H., Thakral, K., Mittal, S., Vatsa, M., & Singh, R. (2023). DF-Platter: multi-face heterogeneous deepfake dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9739–9748). CVPR.

[59] Shao, R., Wu, T., & Liu, Z. (2023). Detecting and grounding multimodal media manipulation. In *CVPR*.

[60] Cai, Z., Ghosh, S., Adatia, A. P., Hayat, M., Dhall, A., & Stefanov, K. (2024). AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset. In *ACM MM*.

[61] Cheng, H., Guo, Y., Wang, T., Nie, L., & Kankanhalli, M. (2024). Diffusion facial forgery detection. In *ACM MM*.

[62] Bhattacharyya, C., Wang, H., Zhang, F., Kim, S., & Zhu, X. (2024). Diffusion deepfake. *arXiv preprint arXiv:2404.01579*.

[63] Yan, Z., et al. (2024). DF40: Toward next-generation deepfake detection. *Advances in Neural Information Processing Systems*, *37*, 29387-29434.

[64] Hou, Y., Fu, H., Chen, C., Li, Z., Zhang, H., & Zhao, J. (2024). PolyglotFake: A novel multilingual and multimodal deepfake dataset. *arXiv preprint arXiv:2405.08838*.

[65] Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., Farhat, K., Caffee, B., Paik, S., Lee, C., et al. (2025). DeepFake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv preprint arXiv:2503.02857*.

[66] Xiong, X., Patel, P., Fan, Q., Wadhwa, A., Selvam, S., Guo, X., & Sengupta, R. (2025). TalkingHeadBench: A Multi-Modal Benchmark & Analysis of Talking-Head DeepFake Detection. *arXiv preprint arXiv:2505.24866*.

[67] He, X., Zhou, Y., Fan, B., Li, B., Zhu, G., & Ding, F. (2025). VLForgery Face Triad: Detection, Localization and Attribution via Multimodal Large Language Models. *arXiv preprint arXiv:2503.06142*.

[68] Li, Y., Zhu, D., Cui, X., & Lyu, S. (2025). Celeb-DF++: A large-scale challenging video deepfake benchmark for generalizable forensics. *arXiv preprint arXiv:2507.18015*.

[69] Pei, G., Zhang, J., Hu, M., Zhang, Z., Wang, C., Wu, Y., & Tao, D. (2024). Deepfake generation and detection: A benchmark and survey. *arXiv preprint arXiv:2403.17881*.

[70] Liu, P., Tao, Q., & Zhou, J. T. (2024). Evolving from Single-modal to Multi-modal Facial Deepfake Detection: Progress and Challenges. *arXiv preprint arXiv:2406.06965*.