Deep Neural Network Watermarking: A Survey

www.surveyx.cn

Abstract

Deep Neural Network (DNN) watermarking has emerged as a critical strategy for safeguarding intellectual property and ensuring model security, particularly in the context of Machine Learning as a Service (MLaaS) platforms and the proliferation of DNNs in mobile and embedded devices. This survey paper systematically explores the methodologies and applications of DNN watermarking, focusing on white-box, black-box, and no-box techniques. White-box techniques embed watermarks directly into model architectures, enhancing robustness against unauthorized modifications. Black-box methods, essential when model internals are inaccessible, focus on embedding and verifying watermarks through outputs or modified inputs. No-box approaches address scenarios with neither access to model internals nor outputs, requiring innovative strategies for watermark embedding and verification. The paper provides a comparative analysis of these techniques, evaluating their effectiveness, robustness, and trade-offs. Key findings highlight the strengths of adaptive methods in enhancing tampering detection and the challenges posed by adversarial attacks and scalability issues. Future directions emphasize the integration of watermarking with emerging technologies to bolster defenses against sophisticated threats. The survey underscores the necessity of ongoing research to advance watermarking techniques, ensuring robust protection of neural network models in diverse and adversarial environments.

1 Introduction

1.1 Concept and Importance of DNN Watermarking

Deep Neural Network (DNN) watermarking is crucial for protecting intellectual property and ensuring model security, particularly as DNNs are increasingly utilized across various domains [1]. The significant investments in model training and data preparation necessitate safeguarding these assets from unauthorized redistribution, especially within Machine Learning as a Service (MLaaS) platforms [2]. Watermarking techniques embed identifiable marks within models, facilitating ownership verification and preventing unauthorized use.

The widespread commercialization of DNNs in mobile and embedded devices further emphasizes the need for robust intellectual property protection mechanisms [3]. As DNNs become integral to numerous applications, effective protection strategies are essential [4]. The emergence of generative AI technologies, such as latent diffusion models (LDMs), complicates the distinction between genuine and AI-generated content, increasing the risk of misuse in contexts like deepfakes [5]. In this environment, watermarking is vital for ensuring the authenticity of models and their outputs, particularly in sensitive applications like facial recognition and medical imaging [6].

In Federated Learning (FL) settings, where data privacy and model ownership are critical, watermarking embeds identifiable markers within models, enabling ownership verification without compromising privacy [7]. The limitations of existing methods that rely solely on watermarking for dataset ownership verification (DOV) highlight the need for innovative verification approaches [8]. Moreover, the demand for secure multimedia information management in deep-learning contexts

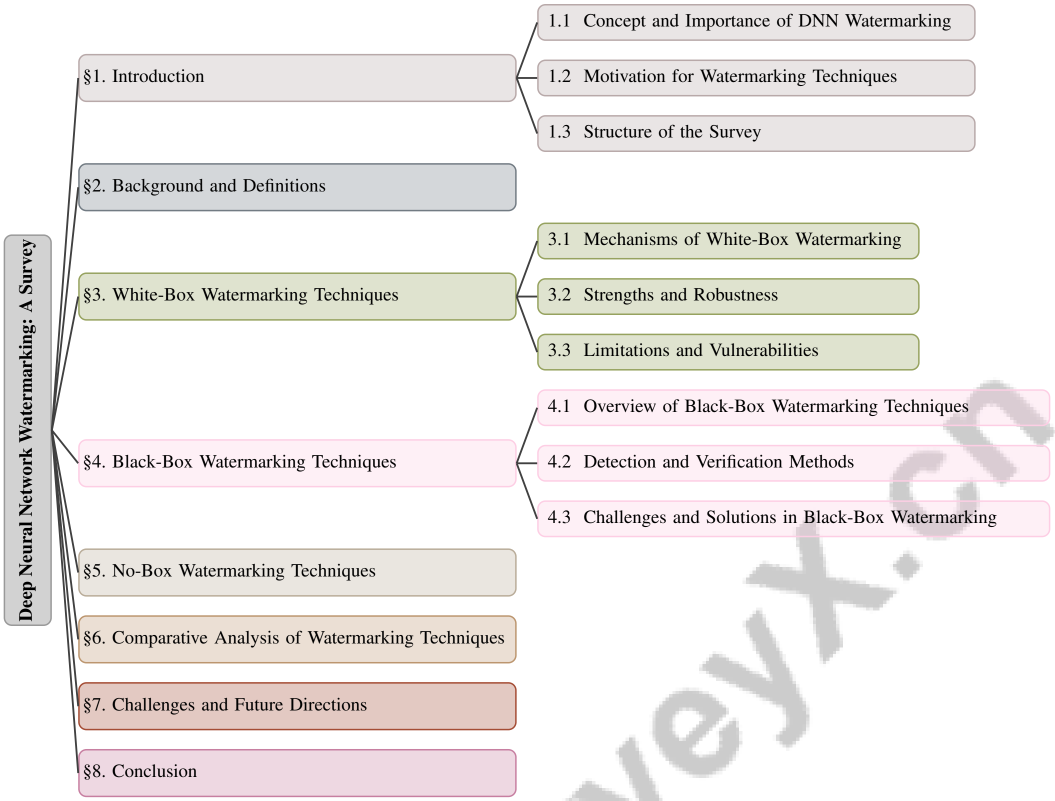

Deep Neural Network Watermarking: A Survey
§1. Introduction
1.1 Concept and Importance of DNN Watermarking
1.2 Motivation for Watermarking Techniques
1.3 Structure of the Survey
§2. Background and Definitions
§3. White-Box Watermarking Techniques
3.1 Mechanisms of White-Box Watermarking
3.2 Strengths and Robustness
3.3 Limitations and Vulnerabilities
§4. Black-Box Watermarking Techniques
4.1 Overview of Black-Box Watermarking Techniques
4.2 Detection and Verification Methods
4.3 Challenges and Solutions in Black-Box Watermarking
§5. No-Box Watermarking Techniques
§6. Comparative Analysis of Watermarking Techniques
§7. Challenges and Future Directions
§8. Conclusion

Figure 1: chapter structure

has driven the development of comprehensive watermarking techniques [9], essential for preventing unauthorized modifications during deployment [10].

Innovative frameworks that introduce new labels to crafted key samples during training enhance the protection of DNN models' intellectual property [11]. By embedding watermarks, these frameworks secure neural network models against unauthorized access and distribution, safeguarding the intellectual and financial investments involved in developing advanced machine learning models.

1.2 Motivation for Watermarking Techniques

The development of watermarking techniques for deep neural networks (DNNs) is driven by the need to protect intellectual property and secure models against unauthorized use. As AI technologies proliferate, the demand for robust copyright protection mechanisms has intensified, reflecting significant investments in sophisticated model development. A critical challenge is the ineffectiveness of existing watermarking methods in black-box scenarios within embedded systems, exposing them to unauthorized access [3].

Watermarking serves as a primary method for detecting and attributing AI-generated content, necessitating advancements in robust techniques to counter forgery and model inversion attacks [5]. Such attacks pose substantial risks, potentially compromising model integrity without degrading performance [10]. Traditional watermarking methods, often inspired by multimedia applications, frequently fail to preserve watermark integrity against manipulations like fine-tuning and pruning [9].

The limitations of existing techniques are particularly evident in federated learning environments, where model leakage by malicious clients is a pressing concern, underscoring the need for effective ownership verification mechanisms [7]. The FedTracker approach illustrates the importance of watermarking techniques compatible with encrypted model parameters, essential for maintaining data privacy while ensuring ownership verification [12]. Furthermore, protecting DNN models trained on textual data from copyright infringement highlights the constraints of current algorithms, primarily designed for image classification tasks [13].

The unauthorized replication and redistribution of DNN models present significant security and legal challenges for model owners and service providers, driving the advancement of sophisticated watermarking techniques [11]. Addressing these challenges is crucial for safeguarding intellectual and financial investments in neural network models. Effective watermarking solutions are vital for ensuring model traceability and ownership verification, particularly in scenarios where unauthorized access and model leakage pose substantial risks [14]. The benchmarks established by [1] further underscore the necessity of evaluating the robustness and reliability of backdoor-based watermarking schemes to prevent unauthorized model redistribution.

1.3 Structure of the Survey

This survey is structured to provide a comprehensive examination of deep neural network watermarking, detailing methodologies such as parameter regularization, margin-based techniques, and probabilistic embedding strategies, while exploring applications aimed at protecting intellectual property and verifying ownership in both black-box and white-box scenarios [15, 16, 2]. The paper begins with an introduction to the concept and importance of DNN watermarking, emphasizing its role in model security and intellectual property protection. Following this, the motivation for developing watermarking techniques is discussed, highlighting the necessity for robust security mechanisms in the evolving AI landscape.

The subsequent section provides a comprehensive background and definitions, elucidating key concepts and terminologies in neural network watermarking. Following this, the relevance of watermarking to intellectual property security and model protection is explored.

The core of the survey is divided into three sections, each dedicated to a specific watermarking technique: white-box, black-box, and no-box. Each section analyzes the mechanisms, strengths, limitations, and challenges associated with these techniques, offering a detailed evaluation of their effectiveness in ensuring model protection and intellectual property security.

A comparative analysis section follows, evaluating the effectiveness, robustness, and applicability of different watermarking techniques. This section discusses trade-offs and comparative strengths and weaknesses, providing a nuanced understanding of their practical implications.

The survey concludes with a discussion on the challenges and future directions in deep neural network watermarking, identifying current limitations and proposing potential research avenues to enhance the security and effectiveness of watermarking techniques. This aligns with the systematic approach of creating a taxonomy of watermarking methods and a unified threat model, as suggested by [15].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Key Concepts in Neural Network Watermarking

Neural network watermarking involves embedding unique identifiers within models to assert ownership and protect intellectual property from unauthorized access [11]. This process is vital for ensuring the traceability and security of deep neural networks (DNNs), which is crucial for maintaining model integrity. Techniques such as controlling channel pruning rates are employed to embed watermarks into the model architecture, facilitating reliable watermark recovery [17]. These mechanisms enhance security by making unauthorized modifications more detectable.

The Free Fine-tuning Watermarking Scheme (PTYNet) exemplifies an innovative method of embedding watermarks by injecting a proprietary model, enabling ownership verification without extensive fine-tuning [18]. This underscores the efficient integration of watermarking techniques. In Federated Learning, the FedCrypt technique demonstrates a dynamic white-box watermarking approach for homomorphically encrypted DNNs, ensuring model security while preserving data privacy [7]. This is essential in scenarios where model parameters are shared across devices, necessitating robust protection mechanisms.

The ZeroMark method offers a novel solution for dataset ownership verification without revealing dataset-specific watermarks, leveraging intrinsic DNN properties [8]. This highlights the adaptability of watermarking techniques to various data modalities. In black-box settings, where model internals are inaccessible, verifying ownership post-theft is challenging. Techniques like ID watermarking of

AI-generated, for reference only.

neural networks (IDwNet) showcase advanced digital watermarking methods that embed unique user IDs into poisoned images [4], demonstrating innovative strategies to ensure watermark robustness across different model types.

The challenge of forgery attacks on semantic watermarks emphasizes the need for robust strategies to withstand adversarial manipulations [5]. The evolving taxonomy of watermarking methods, including white-box versus black-box and static versus dynamic approaches, provides a framework for advancing this critical research area. Neural network watermarking integrates diverse strategies to protect intellectual property, ensure model integrity, and prevent unauthorized usage. The benchmark established by [6] offers a standardized framework for evaluating the robustness of various DNN watermarking schemes against removal attacks, advancing the field with empirical insights into the effectiveness of these techniques.

2.2 Relevance to Intellectual Property Security and Model Protection

Watermarking is crucial for securing intellectual property and protecting neural network models from unauthorized use. By embedding unique identifiers within models, watermarking facilitates ownership verification while maintaining model performance and robustness against attacks [11]. This capability is particularly important in environments where independence from the training dataset enhances the applicability of watermarking methods, especially in black-box scenarios.

The robustness of watermarking frameworks is highlighted by their ability to withstand adversarial attempts, including parameter pruning and functionality stealing attacks. Margin-based Watermarking fortifies models against these threats, safeguarding intellectual property and ensuring model integrity [2]. However, many existing schemes lack systematic evaluation against a comprehensive set of removal attacks, creating uncertainty regarding their practical deployment [6]. This underscores the necessity for advanced watermarking techniques capable of withstanding diverse attack vectors while ensuring AI model reliability.

In federated learning environments, watermarking protects intellectual property by facilitating ownership verification and identifying clients who may unlawfully distribute models. Advanced frameworks like FedTracker employ a dual-layer protection strategy, incorporating a global watermark for model ownership authentication and a local fingerprint mechanism to trace the model's origin back to the responsible client. These measures are crucial in mitigating unauthorized model distribution risks, ensuring the integrity and rights of all contributors in the collaborative training process [19, 20, 12, 21]. Techniques like FedCrypt address the challenge of embedding watermarks into DNNs trained with Homomorphic Encryption, preserving client data privacy and model intellectual property, thus maintaining control over distributed AI models without directly sharing data or model parameters.

Despite advancements in watermarking techniques for digital images and DNNs, significant challenges remain in achieving an optimal balance between imperceptibility, capacity, and robustness. Many existing methods claim resilience against watermark removal attacks; however, empirical evaluations often reveal vulnerabilities to a comprehensive range of adaptive attacks and novel removal strategies, raising doubts about their practical applicability. The lack of systematic testing against diverse threats highlights intrinsic flaws in current robustness assessments, emphasizing the need for rigorous evaluation frameworks to ensure the reliability and effectiveness of watermarking solutions in safeguarding intellectual property and verifying provenance [22, 13, 23, 6]. Additionally, the limitations of existing black-box model watermarking methods, particularly in constructing poisoned images that are susceptible to forgery and fail to uniquely identify users in multi-user settings, further underscore the necessity for innovative solutions.

Watermarking is a critical component in securing neural network models, providing a framework for ownership verification and model protection in an increasingly competitive and adversarial landscape. The development of robust watermarking techniques capable of withstanding various attack vectors, such as forgery and tampering, is essential for ensuring the integrity and reliability of AI models [1].

In recent years, the development of watermarking techniques has gained significant attention, particularly in the realm of digital content protection. Among these, white-box watermarking has emerged as a prominent method due to its unique capabilities and applications. As illustrated in Figure 2, the hierarchical structure of white-box watermarking techniques is categorized into three main sections: mechanisms, strengths, and limitations. The mechanisms section delves into the various integration and training methods employed in these techniques, while the strengths emphasize their resilience

AI-generated, for reference only.

against attacks and their role in safeguarding intellectual property. Conversely, the limitations section addresses the inherent challenges faced by these methods and underscores the necessity for ongoing adaptation and optimization. This comprehensive overview not only highlights the multifaceted nature of white-box watermarking strategies but also sets the stage for a deeper exploration of their implications in the digital landscape.

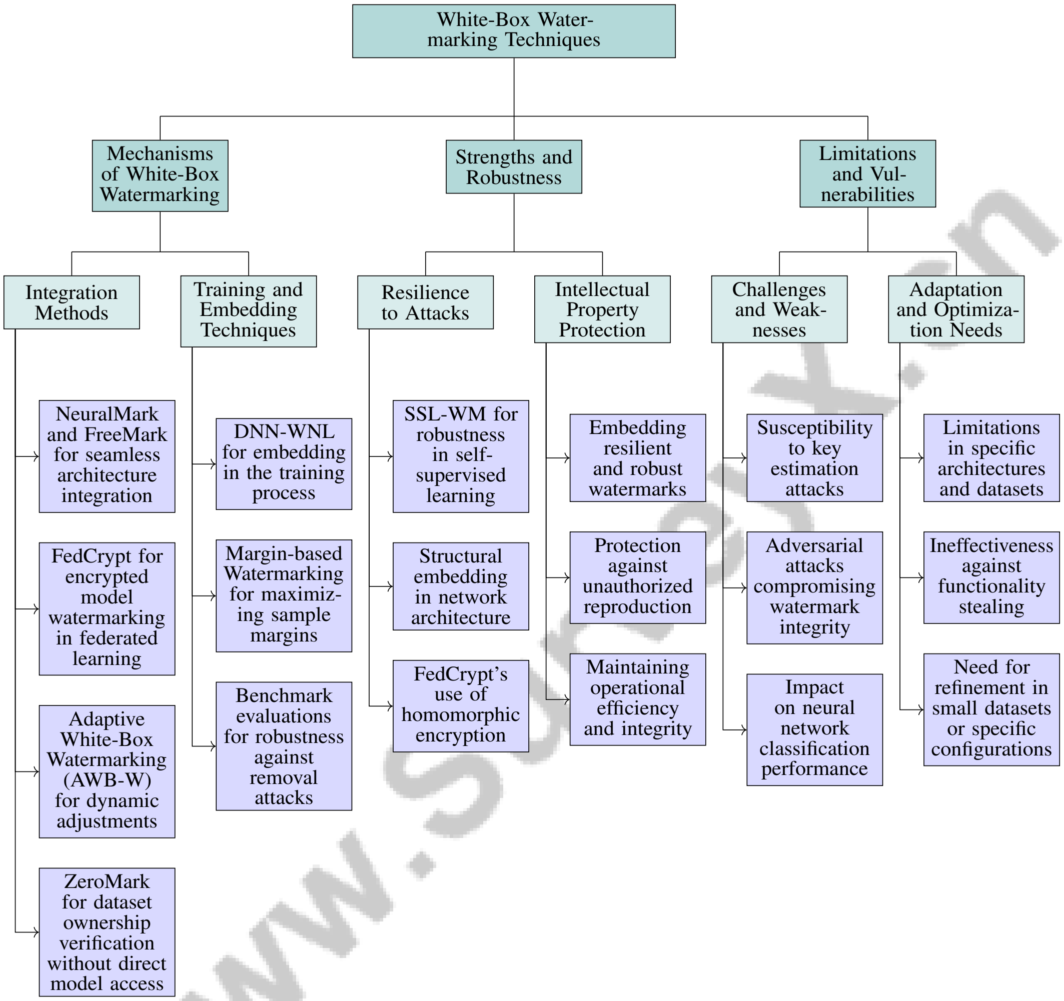

White-Box Watermarking Techniques
Mechanisms of White-Box Watermarking
Strengths and Robustness
Limitations and Vulnerabilities
Integration Methods
Training and Embedding Techniques
Resilience to Attacks
Intellectual Property Protection
Challenges and Weaknesses
Adaptation and Optimization Needs
NeuralMark and FreeMark for seamless architecture integration
FedCrypt for encrypted model watermarking in federated learning
Adaptive White-Box Watermarking (AWB-W) for dynamic adjustments
ZeroMark for dataset ownership verification without direct model access
DNN-WNL for embedding in the training process
Margin-based Watermarking for maximizing sample margins
Benchmark evaluations for robustness against removal attacks
SSL-WM for robustness in self-supervised learning
Structural embedding in network architecture
FedCrypt's use of homomorphic encryption
Embedding resilient and robust watermarks
Protection against unauthorized reproduction
Maintaining operational efficiency and integrity
Susceptibility to key estimation attacks
Adversarial attacks compromising watermark integrity
Impact on neural network classification performance
Limitations in specific architectures and datasets
Ineffectiveness against functionality stealing
Need for refinement in small datasets or specific configurations

Figure 2: This figure illustrates the hierarchical structure of white-box watermarking techniques, categorized into mechanisms, strengths, and limitations. The mechanisms section details integration and training methods, while the strengths highlight resilience and intellectual property protection. Limitations focus on challenges and the need for adaptation and optimization, providing a comprehensive overview of white-box watermarking strategies.

3 White-Box Watermarking Techniques

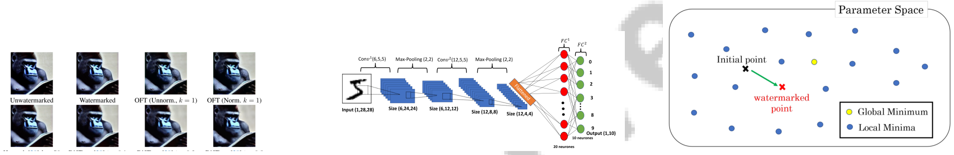
3.1 Mechanisms of White-Box Watermarking

White-box watermarking techniques embed unique markers within neural network parameters, enhancing intellectual property protection and ownership verification. Methods such as NeuralMark and FreeMark integrate seamlessly into various architectures, resisting forgery, fine-tuning, and pruning attacks, thereby strengthening ownership claims [24, 25, 26, 27, 28]. These techniques exploit model structures to incorporate resilient watermarks that maintain performance under adversarial conditions.

FedCrypt exemplifies watermark embedding into encrypted models by training a projection function on model activations using a trigger set, ensuring secure watermarking in federated learning environments [7]. This underscores the need for compatible mechanisms with encrypted models, especially in distributed settings. Adaptive White-Box Watermarking (AWB-W) integrates self-mutual check parameters with adaptive bit adjustments to detect and restore tampered parameters, enhancing robustness through dynamic adjustments [10]. ZeroMark offers a unique solution by generating closest boundary samples and analyzing gradients for dataset ownership verification [8], demonstrating adaptability across data modalities without direct model access.

The DNN-WNL method embeds watermarks directly into the training process by generating key samples with new labels [11], ensuring robust integration against removal attempts. Margin-based Watermarking maximizes watermarked sample margins during training, ensuring correct classification under adversarial conditions [2]. A benchmark by [6] evaluates state-of-the-art schemes against diverse removal attacks, underscoring the need for robust techniques that withstand various attack vectors.

These mechanisms illustrate diverse strategies in white-box watermarking, enhancing neural network security against unauthorized use. By integrating digital watermarks into architectures and employing advanced cryptographic techniques, these strategies bolster model security and verifiability, particularly in adversarial environments. Watermarking protects intellectual property and enables ownership verification in black-box scenarios, safeguarding against functionality stealing attacks. Experiments demonstrate effective watermark embedding during model training without performance degradation, maintaining integrity through fine-tuning and parameter pruning [16, 2].



(a) Comparison of Watermarking Techniques on a Gorilla Image[29]

(b) Convolutional Neural Network (CNN) Architecture[30]

(c) Parameter Space[25]

Figure 3: Examples of Mechanisms of White-Box Watermarking

As shown in Figure 3, the exploration of white-box watermarking techniques illustrates three distinct mechanisms: a comparison of watermarking techniques on a gorilla image, a CNN architecture, and a parameter space representation. The first subfigure presents a comparative analysis of three techniques applied to a gorilla image, showcasing transformations through various stages, including the Omnipresent Fast Transform (OFT) algorithm. This visual comparison is essential for evaluating the efficacy of different approaches. The second subfigure highlights the architecture of a CNN, detailing its layered structure, crucial for processing and embedding watermarks in digital content. Finally, the third subfigure visualizes a parameter space, depicting the proximity of a watermarked point to an initial point, emphasizing the precision required in watermarking processes. Together, these visual elements provide a comprehensive overview of the mechanisms involved in white-box watermarking, illustrating both theoretical and practical aspects of embedding watermarks in digital media [29, 30, 25].

3.2 Strengths and Robustness

White-box watermarking techniques are notable for embedding identifiable markers within neural networks, offering robust protection against unauthorized access and tampering while maintaining performance. These methods are resilient to various attacks, ensuring watermark integrity. Techniques like SSL-WM demonstrate robustness against typical attacks, providing strong ownership verification in self-supervised learning models [31]. This robustness is further exemplified by methods leveraging DNNs' inherent capacity to accommodate additional constraints, allowing effective signature embedding without compromising performance [3].

Structural embedding of watermarks directly into network architecture enhances security by making unauthorized modifications more detectable. This approach is effective against attacks targeting

AI-generated, for reference only.

Method Name	Security Mechanisms	Performance Preservation	Attack Resilience
SSL-WM[31]	Black-box Watermarking	Consistent Output Behavior	Typical Attacks
DNN-WM[3]	Black-box Detection	Maintaining Performance	Robustness Against Attacks
ZM[8]	Boundary Gradients	Maintaining Watermark Confidentiality	Withstand Various Attacks
FC[7]	Homomorphic Encryption	Performance Loss	Removal Attacks
NNL[32]	Watermark Recovery	Preserving Model Performance	Reset Identified Neurons

Table 1: Comparison of various white-box watermarking techniques highlighting their security mechanisms, performance preservation capabilities, and resilience against attacks. The table provides insights into how each method maintains model integrity and security while withstanding diverse attack vectors.

parameter-based methods, as it embeds watermarks into the network structure itself. ZeroMark's effectiveness is highlighted by its ability to utilize the correlation between boundary gradients of benign samples and watermark patterns for secure verification without direct model access [8].

White-box techniques excel in maintaining model fidelity and performance. For instance, FedCrypt uses homomorphic encryption to perform computations on encrypted data, enabling watermark embedding without decrypting model parameters [7]. This ensures the watermarking process does not degrade performance, preserving both security and efficiency in federated learning environments.

Innovative approaches, such as those described by [32], demonstrate effective watermark removal while preserving model performance, even with limited retraining datasets. This underscores the importance of embedding mechanisms that maintain functionality while providing covert protection against detection attacks.

The strengths of white-box watermarking techniques lie in their ability to embed resilient and robust watermarks that withstand a wide range of attacks and manipulations. As illustrated in Figure 4, these methods, including SSL-WM, FedCrypt, and ZeroMark, highlight key security enhancements such as structural embedding and homomorphic encryption. Additionally, the figure emphasizes ownership verification through deep watermarking and fingerprinting, which play a crucial role in safeguarding intellectual property rights associated with DNN models, valuable assets due to high development costs and performance capabilities. By implementing techniques such as deep watermarking and fingerprinting, these strategies protect against unauthorized reproduction and misuse while striving to maintain operational efficiency and integrity. This balance is essential for the effectiveness of modern AI security frameworks, particularly as the risk of intellectual property infringement grows in the rapidly evolving landscape of AI [33, 34, 35]. Table 1 presents a comparative analysis of different white-box watermarking methods, focusing on their security mechanisms, performance preservation, and attack resilience, as discussed in the following section.

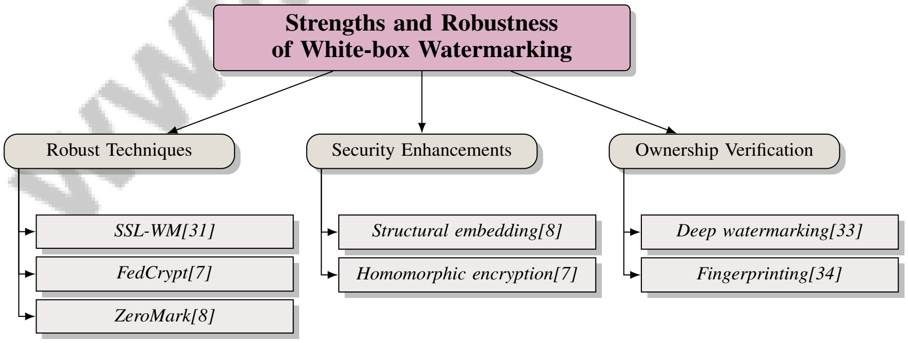


Figure 4: This figure illustrates the strengths and robustness of white-box watermarking techniques, highlighting key methods such as SSL-WM, FedCrypt, and ZeroMark, along with security enhancements like structural embedding and homomorphic encryption. The figure also emphasizes ownership verification through deep watermarking and fingerprinting.

3.3 Limitations and Vulnerabilities

White-box watermarking techniques, despite their effectiveness in embedding identifiable markers within neural networks, face several limitations and vulnerabilities. Table 2 provides a comprehen-

AI-generated, for reference only.

Method Name	Security Challenges	Generalizability Issues	Performance Impact
BTSK[36]	Key Estimation Attacks	Different Cnn Architectures	Classification Accuracy
PDE[14]	Model Leakage	Dataset Applicability	Prediction Accuracy
DLW[37]	Various Attacks	Different Neural Networks	Perceptual Quality
PAPL[38]	Adversarial Attacks	Broader Range Applicability	Accuracy Drops
BW-DNN-IDF[13]	Parameter Pruning	Limit Its Applicability	Preserving Model Performance
IDwNet[4]	Forgery Risks	Multi-user Distribution	Classification Performance Impact
AWB-W[10]	Extensive Tampering	Different Architectures	Performance Degradation
MBW[2]	Functionality Stealing Attacks	-	Accuracy Trade-off

Table 2: Comparison of White-Box Watermarking Methods: A detailed examination of various white-box watermarking techniques highlighting their security challenges, generalizability issues, and performance impact. This table provides insights into the limitations and vulnerabilities associated with each method, drawing on recent advancements and research findings in the field.

sive overview of the limitations and vulnerabilities of current white-box watermarking techniques, emphasizing their security challenges, generalizability issues, and performance impact. A significant concern is their susceptibility to key estimation attacks, necessitating careful key management to ensure security [36]. Traditional methods often suffer from passive protection and inefficiencies in computation and key storage, limiting their effectiveness in dynamic environments [14]. Further exploration into specific deep learning architectures and training datasets is crucial to optimize performance for watermarking, as current methods may not be universally applicable [37].

Adversarial attacks pose a formidable challenge, with the potential to forge evidence and compromise watermark integrity and reliability [38]. The security of methods relying on specific data characteristics, such as selecting words with low TF-IDF scores, may be limited in different contexts, affecting generalizability [13]. Additionally, the impact of watermarking on neural network classification performance, as observed in methods like IDwNet, necessitates further research to minimize adverse effects and maintain efficacy [4].

The opaque nature of neural networks and the limitations of black-box approaches hinder effective location and recovery of tampered parameters, challenging watermarking robustness [10]. Moreover, the ineffectiveness of existing methods against functionality stealing highlights the need for advanced strategies to safeguard intellectual property [2].

While white-box watermarking techniques provide valuable protection for neural network models, addressing these limitations and vulnerabilities is essential for enhancing their applicability and effectiveness in securing intellectual property in increasingly adversarial environments. The potential for diminished effectiveness in scenarios involving small datasets or specific configurations underscores the need for ongoing refinement and adaptation of watermarking techniques. Research indicates that methods like radioactive data demonstrate varying levels of efficacy depending on dataset size and complexity; while effective in white-box settings for large datasets, these techniques may falter in low-class or low-sample scenarios, necessitating further exploration to enhance robustness and applicability in diverse contexts. Continuous innovation in watermarking strategies is vital to ensure reliable ownership demonstration and protection against unauthorized use in an increasingly data-driven landscape [35, 39, 40].

4 Black-Box Watermarking Techniques

Black-box watermarking techniques are pivotal in protecting neural network models, especially when internal architectures are inaccessible. This section delves into various methodologies, their applications, and innovative strategies for embedding watermarks, emphasizing their significance in intellectual property protection and ownership verification across numerous domains.

4.1 Overview of Black-Box Watermarking Techniques

Black-box watermarking techniques embed identifiable markers in neural network models without requiring access to internal parameters, ensuring intellectual property protection and ownership verification while maintaining model confidentiality. The SSL-WM approach exemplifies this by embedding watermarks into self-supervised learning encoders, facilitating ownership verification through downstream classifier outputs [31]. The MOVE framework further demonstrates versatility, operating effectively in both white-box and black-box contexts [41].

In textual data, watermark embedding can be achieved by modifying documents based on TF-IDF scores, enabling integration without model internals [13]. Additionally, binary encoding for words allows selective substitution to inject watermarks, showcasing innovative linguistic manipulations in black-box settings [42].

For audio processing, black-box watermarking techniques have been adapted for Automatic Speech Recognition models, embedding watermarks without requiring model parameter access to protect integrity [43]. The WMD method similarly demonstrates effectiveness in black-box settings by utilizing clean datasets to enhance detection capabilities [22].

The KMP method underscores the importance of unique key management for model protection, drawing inspiration from adversarial defense techniques [44]. Evolutionary trigger set generation techniques optimize trigger patterns across popular DNN models, enhancing black-box watermarking effectiveness [45].

Embedding the author's signature into a DNN's training dataset ensures robust ownership verification linked to the data without direct model access [3].

These techniques illustrate the adaptability of black-box methods across multiple domains, such as text generation, audio, and image processing. They enhance AI-generated content detection, improve intellectual property protection, and bolster accountability in digital media, addressing challenges posed by generative models and unauthorized content use [46, 47, 22, 42]. By embedding watermarks without requiring access to model internals, these methods provide robust solutions for ownership verification and intellectual property protection in diverse environments.

As shown in Figure 5, this figure illustrates the classification of black-box watermarking techniques across various domains, highlighting key methodologies and their applications in text, audio, and image processing. The first image, "Evasion Rate and Bit-wise Accuracy vs Number of Models (k)," explores the trade-off between evasion rate and bit-wise accuracy across models, highlighting scalability and robustness of techniques like HiDDEN, MBRS, StegaStamp, and RivaGAN. The second image, "SimCLR: Simultaneous Representation Learning for Large-Scale Image Classification," illustrates representation learning applications across diverse datasets, emphasizing SimCLR's classification capabilities on datasets like CIFAR10 and STL-10. Lastly, the "Deep Learning Model Training and Authentication System" image outlines a dual-component system for watermarking, where the training component generates samples through a neural network and compares them to a trigger set for loss calculation, while the authentication component ensures model integrity and ownership verification. Together, these examples underscore the diverse strategies and considerations in implementing black-box watermarking techniques in machine learning [29, 48, 49].

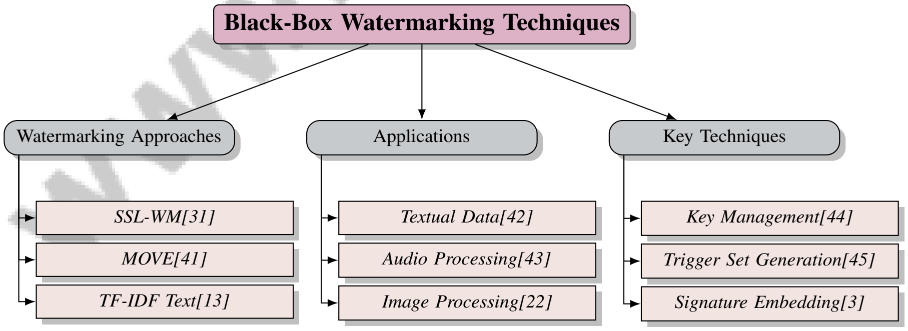


Figure 5: This figure illustrates the classification of black-box watermarking techniques across various domains, highlighting key methodologies and their applications in text, audio, and image processing.

4.2 Detection and Verification Methods

Detecting and verifying black-box watermarks in neural networks is crucial for safeguarding model integrity and authenticity by identifying unauthorized modifications, such as malicious fine-tuning and backdoor attacks, without internal parameter access. This is increasingly important as deep neural networks (DNNs) are deployed in commercial applications, making them susceptible to various threats that can compromise performance and reliability. Advanced watermarking techniques using trigger image sets and innovative training methods can effectively detect integrity breaches while enabling copyright authentication, ensuring models maintain their original classification capabilities [32, 49]. Various methods have been developed to enhance these processes, leveraging statistical analyses and performance metrics.

One approach employs statistical tests tailored for specific watermarking schemes, such as Red-Green, Fixed-Sampling, and Cache-Augmented methods, to detect watermarks in models [47]. These tests identify output deviations indicating an embedded watermark, facilitating robust verification.

The Neural Dehydration (Dehydra) attack benchmark emphasizes challenges in removing black-box watermarks, particularly under limited data conditions, underscoring the need for robust detection methods that withstand adversarial attacks [50]. This highlights the necessity of developing resilient verification strategies against watermark removal attempts.

In self-supervised learning (SSL), the method proposed by [48] trains encoders to produce consistent representation vectors for watermarked inputs, ensuring reliable detection and verification of watermarks by downstream classifiers.

For speech recognition models, watermarking schemes are assessed through fidelity, robustness, and integrity tests, measuring metrics such as word error rate (WER) and character error rate (CER) against various attacks [51]. These metrics provide a comprehensive evaluation of watermark resilience to adversarial manipulations.

The DVBW method focuses on verifying backdoor patterns through model predictions, allowing detection of watermarks embedded within decision-making processes [52]. This ensures that watermarks can be verified based on model output behavior, even in black-box settings.

Advanced evaluation techniques, as described by [53], utilize metrics like ROC-AUC and partial AUC (pAUC) to compare watermarked text against non-watermarked text generated by the same model, offering a quantitative assessment of watermark detection performance.

The B4 method's performance is evaluated using fidelity (P-SP) and efficacy (ROC-AUC) metrics, illustrating its capability in detecting and verifying black-box watermarks [54]. This method emphasizes balancing detection accuracy with minimal impact on model performance.

Finally, the evaluation method proposed by [42] employs Receiver Operating Characteristic (ROC) curves and Z-scores to assess watermark strength, providing a robust framework for evaluating detection and verification effectiveness while ensuring covert yet detectable watermarks.

Collectively, these methods highlight diverse strategies for detecting and verifying black-box watermarks, each tailored to enhance neural network security against unauthorized use and tampering. By employing advanced statistical analyses and performance metrics, these watermarking techniques bolster the security and verifiability of neural networks in adversarial contexts, effectively safeguarding against functionality stealing attacks and unauthorized redistribution while maintaining predictive accuracy [55, 56, 2].

4.3 Challenges and Solutions in Black-Box Watermarking

Black-box watermarking techniques encounter several challenges impacting their effectiveness and robustness, necessitating innovative solutions for enhanced applicability and security. A significant challenge is the reliance on specific watermark characteristics, which may limit the generalizability of methods like Dehydra across all watermarking schemes, especially those with unique designs [50]. This underscores the need for adaptable watermarking approaches accommodating diverse scheme designs.

The complexity of generating effective trigger audios for speech recognition models presents another challenge, as these triggers must not degrade model accuracy while ensuring robust watermark embedding [51]. This complexity highlights the importance of developing sophisticated audio watermarking techniques that balance effectiveness with model performance.

While SSL-WM is notable for its robustness against common watermark removal techniques, extreme model modifications such as excessive pruning can significantly reduce watermark effectiveness [31]. Addressing this limitation requires resilient watermarking strategies that maintain integrity under substantial model alterations.

Potential vulnerability to replay attacks, where attackers reuse previously obtained copyright evidence, necessitates enhancements to protect against such exploits [57]. Solutions may involve dynamic watermarking techniques that periodically update or alter watermark patterns to prevent reuse.

The susceptibility of methods like RIGA to advanced model stealing and distillation attacks in black-box settings further emphasizes the need for robust defenses [24]. Developing secure watermarking frameworks that withstand these sophisticated attacks is crucial for ensuring intellectual property protection.

The extractor-gradient-guided (EGG) remover exemplifies a significant vulnerability in existing black-box watermarking techniques by successfully removing watermarks, indicating the necessity for advanced countermeasures to prevent such removals [58]. Enhancing watermark stealth and resilience against removal techniques is essential for maintaining security.

Finally, the requirement for fine-tuning parameters, such as thresholds for synonym selection in text watermarking, presents a challenge in optimizing watermark strength and fidelity [42]. Developing automated or adaptive parameter tuning methods could improve watermark embedding efficiency and effectiveness.

To effectively tackle evolving challenges in deep neural network (DNN) model integrity, advancing watermarking methodologies that enhance robustness, adaptability, and resistance to various attacks is essential. This includes developing innovative black-box watermarking techniques capable of detecting malicious fine-tuning and ensuring secure copyright authentication. By employing strategies such as trigger image sets and specialized two-stage training approaches, these methodologies can maintain original model performance while safeguarding against threats like poisoning and backdoor attacks. Ultimately, these advancements will facilitate secure and reliable protection of neural network models in black-box environments, ensuring their integrity in commercial applications [59, 49].

5 No-Box Watermarking Techniques

The development of no-box watermarking techniques is essential for enhancing neural network security, particularly when traditional methods fail due to restricted access to model parameters and outputs. This section outlines the principles of no-box watermarking, highlighting its importance in protecting intellectual property. No-box watermarking is introduced as an innovative strategy for ensuring integrity and ownership verification of neural networks under challenging conditions. The discussion focuses on methodologies such as the Wide Flat Minimum Watermarking (WFM) technique, showcasing advancements in this field.

5.1 Introduction to No-Box Watermarking

No-box watermarking represents a novel approach to neural network security, operating without access to a model's internal parameters or outputs. This paradigm effectively tackles the challenges of intellectual property protection and ownership verification in environments where traditional methods, such as white-box and black-box techniques, are insufficient [12]. The FedTracker technique exemplifies this by enabling ownership verification and traceability without accessing model internals or outputs.

Challenges in no-box watermarking stem from the need to embed and verify watermarks without direct model interaction. This requires independent strategies, as seen in structural watermarking methods that maintain watermark reliability and robustness while preserving the fidelity of the DNN's original task [17]. The method proposed by [60] highlights robust protection against fine-tuning attacks, where adversaries adapt stolen models using small datasets, emphasizing the need for effective solutions under adversarial conditions.

The emergence of no-box watermarking necessitates a reevaluation of traditional strategies, leading to robust solutions designed to protect model security and intellectual property in restricted-access environments. This shift is crucial as existing box-free methods remain vulnerable to various removal attacks, requiring enhanced protective measures effective under black-box conditions [61, 29, 58, 27, 28]. This evolving field offers significant opportunities for research and development to address the unique challenges of securing neural network models in complex and adversarial environments.

5.2 Wide Flat Minimum Watermarking (WFM)

The Wide Flat Minimum Watermarking (WFM) method marks a significant advancement in no-box watermarking, offering a robust solution for embedding watermarks during the training of Generative Adversarial Networks (GANs). This technique ensures resilience against a wide range of attacks, including both black-box and white-box scenarios [62]. WFM embeds watermarks that remain robust and detectable despite adversarial attempts at alteration or removal.

The efficacy of the WFM approach is evaluated using a diverse dataset of 1,000 images, rigorously testing watermarking techniques under various scenarios [29]. This evaluation confirms the WFM method's theoretical soundness and practical effectiveness across real-world applications.

A key feature of the WFM technique is its resilience to novel attack methodologies, as evidenced by benchmarks incorporating both black-box and white-box strategies [63]. These benchmarks challenge traditional approaches by addressing previously overlooked vulnerabilities, providing a rigorous assessment of watermark robustness.

Furthermore, WFM addresses the challenge of sophisticated forgery attacks that use unrelated models to forge or remove watermarks [5]. By embedding watermarks that exploit the wide flat minima of the loss landscape, WFM enhances security and reliability, making it more resistant to advanced forgery techniques.

The WFM method exemplifies innovative strategies in no-box watermarking, offering a comprehensive framework that effectively safeguards the security and integrity of neural network models in scenarios with limited access to model parameters or outputs. This method ensures ownership protection against unauthorized redistribution while maintaining model performance and demonstrating resilience against adversarial attacks like parameter pruning and brute-force attempts. By integrating watermarking techniques during training and using advanced metrics like Term Frequency and Inverse Document Frequency, the WFM method enhances verification while preserving operational efficacy [15, 28, 27, 13, 59].

5.3 Security and Vulnerability Considerations

No-box watermarking techniques introduce unique security challenges and vulnerabilities due to their operation in environments lacking access to model internals or outputs. Ensuring watermark robustness under these constraints requires innovative security strategies capable of withstanding various adversarial attacks. A primary concern is the potential for adversaries to exploit limited access to model-specific information, attempting to remove or forge watermarks without direct interaction with the model's architecture [32].

The robustness of no-box watermarking is tested against sophisticated forgery attacks, where adversaries use unrelated models to bypass watermark detection mechanisms. Such attacks pose substantial risks by undermining the integrity of the watermarking process, crucial for verifying model ownership. Unauthorized use or distribution of machine learning models can occur, allowing malicious users to evade detection and exploit the intellectual property embedded within these models. This concern is heightened as current watermarking techniques, particularly backdoor-based methods, remain vulnerable to evasion strategies that obscure ownership verification, facilitating model theft [19, 64, 56]. The WFM approach addresses these vulnerabilities by embedding watermarks that exploit the wide flat minima of the loss landscape, enhancing resistance to advanced forgery techniques.

Despite advancements, no-box watermarking techniques may not entirely eliminate all types of watermarks, particularly those robust against adversarial attacks. This limitation underscores the necessity for ongoing research and development to identify and mitigate potential vulnerabilities, ensuring watermarking methods remain effective in increasingly adversarial environments [32].

The adaptability of no-box watermarking strategies is crucial for ensuring robust security across various applications and model types, especially given their susceptibility to removal attacks in black-box settings. Recent studies reveal the challenges of effectively safeguarding intellectual property in diverse deep learning contexts [61, 29, 58, 42]. As the field evolves, developing standardized benchmarks and evaluation metrics will be vital for assessing the security and reliability of these techniques, providing a framework for continuous improvements and innovations in neural network protection.

AI-generated, for reference only.

6 Comparative Analysis of Watermarking Techniques

The advancement of deep neural networks (DNNs) has highlighted the necessity of watermarking techniques as vital tools for safeguarding intellectual property and maintaining model integrity. This section evaluates the effectiveness and robustness of various watermarking methods, focusing on their performance metrics and resilience against adversarial attacks to discern key factors influencing their practical success or limitations.

6.1 Effectiveness and Robustness of Watermarking Techniques

Benchmark	Size	Domain	Task Format	Metric
WRT[6]	1,260,000	Image Classification	Watermarking Evaluation	Watermark Accuracy, Stealing Loss
Dehydra[50]	1,000,000	Computer Vision	Watermark Removal	Rescaled Watermark Accuracy
OFT[29]	1,000	Image Watermarking	Evasion Attack	Evasion Rate, Bit-wise Accuracy
AudioMarkBench[65]	20,000	Audio Watermarking	Robustness Evaluation	FNR, FPR
WRT[23]	1,260,000	Image Classification	Watermarking Evaluation	Watermark Accuracy, Stealing Loss
MLP-Bench[66]	46,753	Mobile Security	Model Protection Assessment	Model Extraction Rate, Protection Effectiveness
WMD[64]	1,000	Image Classification	Watermark Verification	Accuracy, Verification Rate
INB[67]	9	Deep Neural Networks	Watermark Removal	Bit Error Rate, Model Utility

Table 3: This table presents a comprehensive overview of various benchmarks utilized in the evaluation of watermarking techniques across different domains. It details the size, task formats, and metrics employed to assess the robustness and effectiveness of watermarking methods, highlighting key studies in the field.

Watermarking techniques play a crucial role in securing the intellectual property and integrity of DNNs. Table 3 provides a detailed overview of representative benchmarks used to evaluate the effectiveness and robustness of watermarking techniques in securing deep neural networks. Adaptive watermarking enhances tampering detection and recovery, increasing watermark capacity under lossless conditions [10]. Margin-based watermarking effectively counters functionality stealing attacks, achieving perfect accuracy on trigger sets [2], thereby proving its robustness under challenging conditions. However, no single scheme is universally robust against all attacks, as highlighted by [6], necessitating the development of more resilient techniques. The method by [11] achieves zero false-positive rates and enhanced attack resilience, underscoring the importance of precision in watermark embedding and detection.

Black-box methods like SSL-WM and WMD effectively verify ownership and detect invisible watermarks, addressing misuse concerns while preserving model integrity [57, 49, 42, 47]. Continuous refinement is crucial for enhancing watermarking techniques' effectiveness and robustness, as seen in studies by [9] and [8], which emphasize the need for robust defenses against unauthorized model exploitation.

6.2 Trade-offs in Watermarking Approaches

Selecting watermarking approaches for DNNs involves trade-offs impacting security, performance, and applicability. A key trade-off exists between robustness and model performance; for example, FedReverse embeds reversible watermarks in federated learning but may introduce distortion with more clients, affecting performance [68]. Complexity versus effectiveness is another trade-off, as techniques countering adversarial attacks often require complex processes and significant resources. Advanced methods secure DNNs with unique identifiers during training, protecting intellectual property while ensuring accuracy, yet they may be resource-intensive [56, 2, 13, 59].

Adaptability to various architectures and data types also presents trade-offs. Some methods are task-optimized but lack broader application flexibility, complicating effectiveness when models adapt or retrain. Even advanced methods can be circumvented by adversaries, highlighting the need for ongoing innovation [56, 40]. The transparency versus security trade-off is significant; covert watermarking may reduce robustness against removal attempts, as adversaries can exploit this to remove watermarks without understanding the scheme [46].

Ultimately, watermarking approach choices must align with specific security requirements and constraints. Balancing robustness, payload capacity, and unobtrusiveness is vital for effective DNN watermarking solutions that protect intellectual property while preserving performance. Techniques like margin-based watermarking and distribution-optimized weight settings enhance security without compromising functionality, making them crucial in the MLaaS landscape [69, 11, 56, 2].

6.3 Comparative Strengths and Weaknesses

Evaluating the strengths and weaknesses of watermarking techniques in DNNs is essential for assessing their effectiveness in intellectual property protection. The approach by [70] effectively watermarks neural networks without direct model weight access, facilitating remote ownership verification. DEEP JUDGE, as outlined by [71], offers non-invasive, efficient watermarking with robustness against adaptive attacks, maintaining model performance while providing strong security measures.

Despite strengths, limitations exist. The method by [72] improves traditional techniques but may still face sophisticated attacks, underscoring the need for innovation. Similarly, [73] may struggle when adversaries adapt to watermarking strategies, revealing a potential adaptability weakness. [1] confirms vulnerabilities in backdoor-based schemes to removal attacks, indicating a critical reliance weakness on these methods alone.

Various watermarking techniques offer distinct security and efficiency advantages but also exhibit weaknesses requiring ongoing research and development. Balancing these is crucial for advancing the field and ensuring protection against unauthorized access and exploitation, especially amid emerging threats like functionality stealing and model modification. As DNNs increase in value as intellectual property, innovative approaches like margin-based and exponential weighting methods must be developed to ensure ownership verification while maintaining performance in real-world applications [55, 59, 2].

7 Challenges and Future Directions

Enhancing the security and integrity of neural network models requires addressing the complex challenges inherent in watermarking techniques. This section delves into the obstacles that must be overcome to bolster these techniques' resilience against adversarial attacks, a vital component of effective intellectual property protection. By examining current watermarking methodologies, we identify key areas for improvement and future research directions essential for the evolution of secure watermarking practices.

7.1 Robustness Against Attacks

Ensuring watermarking techniques' robustness against diverse adversarial attacks is crucial for safeguarding neural network models. Existing methodologies often struggle to balance model performance with defense against sophisticated threats [9]. The complexity of adversarial environments necessitates continuous advancements in watermarking strategies to address vulnerabilities like key estimation and forgery [3].

The SSL-WM approach, while effective against common removal techniques, requires enhancements to withstand more sophisticated attacks [31]. Similarly, ZeroMark offers significant security by concealing watermark patterns during verification but needs further exploration to ensure robustness against resourceful attackers [8]. Adaptive techniques, such as those by [10], offer parameter-level localization and recovery, enhancing robustness against various attacks. However, the computational complexity of methods using Homomorphic Encryption (HE) may hinder practical application by slowing the watermark embedding process [7].

Margin-based watermarking demonstrates exceptional robustness against functionality stealing attacks, ensuring ownership verification even against sophisticated adversarial techniques [2]. Nonetheless, methods like those proposed by [11] may not prevent all attack forms, especially those exploiting the model's learning characteristics, highlighting the challenge of ensuring watermark robustness.

AI-generated, for reference only.

Future research should focus on developing more resilient watermarking techniques and exploring additional attack methodologies to assess model security [1]. Expanding benchmarks to include various attack vectors and evaluation metrics is crucial for enhancing watermarking schemes' robustness [6]. Advancing watermarking techniques involves a multifaceted approach, including device-bound and key-storageless methods, flexible embedding and extraction processes, and additional security measures against sophisticated attacks. Addressing these challenges will propel the field toward more secure and reliable watermarking methods, ensuring effectiveness in increasingly adversarial environments.

7.2 Scalability and Performance

The scalability and performance of watermarking techniques are critical for their practical application in real-world scenarios. As neural networks grow in complexity, the ability of watermarking methods to scale without compromising performance becomes increasingly important. A primary challenge is optimizing key prompt selection and exploring fine-grained locking strategies to enhance scalability [74]. This involves developing methods that efficiently handle large-scale models while maintaining the integrity and robustness of embedded watermarks.

The computational complexity of embedding and verifying watermarks in deep neural networks significantly impacts performance, especially in resource-constrained environments. This complexity arises from the need to integrate watermarking techniques without degrading model accuracy or efficiency. Techniques requiring intricate embedding processes or substantial computational resources may hinder scalability, limiting deployment in large-scale applications. Future research should focus on optimizing these processes to reduce computational overhead and improve efficiency.

Evaluating watermarking techniques in terms of their ability to maintain model accuracy and functionality is essential. Ensuring these methods do not compromise model performance is crucial, as degradation could lead to harmful misclassification behaviors and ownership ambiguity, undermining ownership verification and intellectual property protection in machine learning models [19, 56]. A careful balance between embedding robustness and model fidelity is necessary to ensure that watermarking processes do not introduce significant performance penalties.

Addressing challenges related to scalability and performance involves developing efficient watermarking techniques that seamlessly integrate into large-scale neural networks without compromising effectiveness. By optimizing key prompt selection and investigating advanced locking strategies, researchers can significantly enhance scalability and performance, ensuring these techniques remain effective in diverse environments [19, 75, 22, 40].

7.3 Integration with Emerging Technologies

Integrating watermarking techniques with emerging technologies represents a strategic advancement in enhancing the security and robustness of deep neural networks (DNNs). As artificial intelligence evolves, there is a growing need for sophisticated intellectual property protection methods adaptable to new technological paradigms. One promising avenue for future research is exploring adversarial purification mechanisms to counter evidence forgery in watermarking techniques [38], bolstering defenses against increasingly sophisticated forgery attacks.

Developing resilient watermarking techniques that do not rely on inversion methods is crucial for effectively counteracting forgery attacks [5]. By focusing on robust watermarking schemes, researchers can address challenges posed by model security and explore integrating encryption with watermarking to enhance protection [9]. This integration could provide a dual-layered security framework, enhancing the overall resilience of neural network models.

Optimizing verification processes for faster execution and exploring the applicability of techniques like ZeroMark across different watermarking methods are essential for advancing the field [8]. Additionally, enhancing adaptive watermarking techniques could improve their applicability across various neural network architectures, increasing versatility and effectiveness [10].

Emerging technologies such as query rejection mechanisms can further strengthen watermarking methods' resistance against novel attacks, ensuring applicability across diverse neural network types [11]. Furthermore, optimizing the balance between watermarking effectiveness and model performance is critical for future research, ensuring that security measures do not compromise AI models' functionality [2].

Integrating watermarking techniques with emerging technologies offers significant potential for advancing neural network security. Investigating and implementing innovative integration strategies for DNN watermarking can enhance resilience against emerging threats, such as functionality stealing attacks and piracy. This will ensure watermarking remains a crucial mechanism for protecting the integrity and intellectual property of DNN models, particularly as they are increasingly deployed in complex and adversarial environments, such as Machine Learning as a Service (MLaaS) platforms. These advancements will bolster ownership verification in black-box scenarios and safeguard substantial investments in developing these models, maintaining their value in critical applications, including healthcare, autonomous driving, and natural language processing [76, 15, 77, 56, 2].

8 Conclusion

The survey emphasizes the critical importance of deep neural network watermarking for intellectual property protection and model security across various applications. Techniques such as the IDwNet method achieve remarkable invisibility and validation success rates, with over 99% accuracy for poisoned images while minimally affecting classification accuracy [4]. This demonstrates the effectiveness of embedding identifiers within models for robust ownership verification.

The proposed mechanism addresses security challenges in AI model transmission and deployment by offering a device-bound and key-storageless solution, enhancing model protection against unauthorized access and distribution, particularly in federated learning environments [14].

Experimental findings indicate significant improvements in watermark reliability and model performance, showcasing resilience against adversarial attacks [38]. These results highlight the necessity for robust watermarking techniques capable of enduring sophisticated adversarial strategies while preserving model integrity.

Future research will aim to strengthen watermarking methods against input preprocessing attacks and investigate their applicability beyond image classification domains [18]. This expansion is vital for adapting watermarking techniques to emerging AI technologies and applications, ensuring comprehensive protection across diverse contexts.

The survey underscores the need for continued research and development in neural network watermarking. As AI technologies progress, innovative solutions are crucial to confront emerging challenges and enhance protection mechanisms, safeguarding the integrity and security of neural networks in an increasingly adversarial landscape.

AI-generated, for reference only.

References

[1] Masoumeh Shafieinejad, Nils Lukas, Jiaqi Wang, Xinda Li, and Florian Kerschbaum. On the robustness of backdoor-based watermarking in deep neural networks. In *Proceedings of the 2021 ACM workshop on information hiding and multimedia security*, pages 177–188, 2021.

[2] Byungjoo Kim, Suyoung Lee, Seanie Lee, Sooel Son, and Sung Ju Hwang. Margin-based neural network watermarking. In *International Conference on Machine Learning*, pages 16696–16711. PMLR, 2023.

[3] Jia Guo and Miodrag Potkonjak. Watermarking deep neural networks for embedded systems. In *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 1–8. IEEE, 2018.

[4] Mouke Mo, Chuntao Wang, and Shan Bian. A unique identification-oriented black-box watermarking scheme for deep classification neural networks. *Symmetry*, 16(3):299, 2024.

[5] Andreas Müller, Denis Lukovnikov, Jonas Thietke, Asja Fischer, and Erwin Quiring. Black-box forgery attacks on semantic watermarks for diffusion models. *arXiv preprint arXiv:2412.03283*, 2024.

[6] Nils Lukas, Edward Jiang, Xinda Li, and Florian Kerschbaum. Sok: How robust is image classification deep neural network watermarking? (extended version), 2021.

[7] Mohammed Lansari, Reda Bellafqira, Katarzyna Kapusta, Kassem Kallas, Vincent Thouvenot, Olivier Bettan, and Gouenou Coatrieux. Fedcrypt: A dynamic white-box watermarking scheme for homomorphic federated learning. 2024.

[8] Junfeng Guo, Yiming Li, Ruibo Chen, Yihan Wu, Heng Huang, et al. Zeromark: Towards dataset ownership verification without disclosing watermark. *Advances in Neural Information Processing Systems*, 37:120468–120500, 2024.

[9] Himanshu Kumar Singh and Amit Kumar Singh. Comprehensive review of watermarking techniques in deep-learning environments. *Journal of Electronic Imaging*, 32(3):031804–031804, 2023.

[10] Zhenzhe Gao, Zhaoxia Yin, Hongjian Zhan, Heng Yin, and Yue Lu. Adaptive white-box watermarking with self-mutual check parameters in deep neural networks, 2023.

[11] Qi Zhong, Leo Yu Zhang, Jun Zhang, Longxiang Gao, and Yong Xiang. Protecting ip of deep neural networks with watermarking: A new label helps. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II 24*, pages 462–474. Springer, 2020.

[12] Shuo Shao, Wenyuan Yang, Hanlin Gu, Zhan Qin, Lixin Fan, Qiang Yang, and Kui Ren. Fedtracker: Furnishing ownership verification and traceability for federated learning model, 2024.

[13] Mohammad Mehdi Yadollahi, Farzaneh Shoeleh, Sajjad Dadkhah, and Ali A Ghorbani. Robust black-box watermarking for deep neural network using inverse document frequency. In *2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 574–581. IEEE, 2021.

[14] Qianqian Pan, Mianxiong Dong, Kaoru Ota, and Jun Wu. Device-bind key-storageless hardware ai model ip protection: A puf and permute-diffusion encryption-enabled approach, 2022.

[15] Franziska Boenisch. A systematic review on model watermarking for neural networks. *Frontiers in big Data*, 4:729663, 2021.

[16] Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, pages 269–277, 2017.

[17] Xiangyu Zhao, Yinzhe Yao, Hanzhou Wu, and Xinpeng Zhang. Structural watermarking to deep neural networks via network channel pruning. In *2021 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2021.

[18] Run Wang, Jixing Ren, Boheng Li, Tianyi She, Chenhao Lin, Liming Fang, Jing Chen, Chao Shen, and Lina Wang. Free fine-tuning: A plug-and-play watermarking scheme for deep neural networks, 2022.

[19] Shuo Shao, Yiming Li, Hongwei Yao, Yiling He, Zhan Qin, and Kui Ren. Explanation as a watermark: Towards harmless and multi-bit model ownership verification via watermarking feature attribution. *arXiv preprint arXiv:2405.04825*, 2024.

[20] Fang-Qi Li, Shi-Lin Wang, and Alan Wee-Chung Liew. Towards practical watermark for deep neural networks in federated learning, 2021.

[21] Shuo Shao, Wenyuan Yang, Hanlin Gu, Zhan Qin, Lixin Fan, and Qiang Yang. Fedtracker: Furnishing ownership verification and traceability for federated learning model. *IEEE Transactions on Dependable and Secure Computing*, 2024.

[22] Minzhou Pan, Zhenting Wang, Xin Dong, Vikash Sehwag, Lingjuan Lyu, and Xue Lin. Finding needles in a haystack: A black-box approach to invisible watermark detection. In *European Conference on Computer Vision*, pages 253–270. Springer, 2024.

[23] Nils Lukas, Edward Jiang, Xinda Li, and Florian Kerschbaum. Sok: How robust is image classification deep neural network watermarking? In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 787–804. IEEE, 2022.

[24] Tianhao Wang and Florian Kerschbaum. Riga: Covert and robust white-box watermarking of deep neural networks, 2021.

[25] Minoru Kuribayashi, Tatsuya Yasui, and Asad Malik. White box watermarking for convolution layers in fine-tuning model using the constant weight code. *Journal of Imaging*, 9(6):117, 2023.

[26] Fang-Qi Li, Shi-Lin Wang, and Yun Zhu. Fostering the robustness of white-box deep neural network watermarks by neuron alignment, 2021.

[27] Yuan Yao, Jin Song, Jian Jin, and Zhenzhen Jiao. Neuralmark: Advancing white-box neural network watermarking.

[28] Yuzhang Chen, Jiangnan Zhu, Yujie Gu, Minoru Kuribayashi, and Kouichi Sakurai. Freemark: A non-invasive white-box watermarking for deep neural networks. *arXiv preprint arXiv:2409.09996*, 2024.

[29] Qilong Wu and Varun Chandrasekaran. The efficacy of transfer-based no-box attacks on image watermarking: A pragmatic analysis. *arXiv preprint arXiv:2412.02576*, 2024.

[30] Reda Bellafqira and Gouenou Coatrieux. Diction: Dynamic robust white box watermarking scheme, 2022.

[31] Peizhuo Lv, Pan Li, Shenchen Zhu, Shengzhi Zhang, Kai Chen, Ruigang Liang, Chang Yue, Fan Xiang, Yuling Cai, Hualong Ma, Yingjun Zhang, and Guozhu Meng. Ssl-wm: A black-box watermarking approach for encoders pre-trained by self-supervised learning, 2024.

[32] William Aiken, Hyoungshick Kim, Simon Woo, and Jungwoo Ryoo. Neural network laundering: Removing black-box backdoor watermarks from deep neural networks. *Computers & Security*, 106:102277, 2021.

[33] Yuchen Sun, Tianpeng Liu, Panhe Hu, Qing Liao, Shaojing Fu, Nenghai Yu, Deke Guo, Yongxiang Liu, and Li Liu. Deep intellectual property protection: A survey. *arXiv preprint arXiv:2304.14613*, 2023.

[34] Mingfu Xue, Yushu Zhang, Jian Wang, and Weiqiang Liu. Intellectual property protection for deep learning models: Taxonomy, methods, attacks, and evaluations. *IEEE Transactions on Artificial Intelligence*, 3(6):908–923, 2021.

[35] Jie Ren, Han Xu, Pengfei He, Yingqian Cui, Shenglai Zeng, Jiankun Zhang, Hongzhi Wen, Jiayuan Ding, Pei Huang, Lingjuan Lyu, et al. Copyright protection in generative ai: A technical perspective. *arXiv preprint arXiv:2402.02333*, 2024.

[36] MaungMaung AprilPyone and Hitoshi Kiya. A protection method of trained cnn model using feature maps transformed with secret key from unauthorized access, 2021.

[37] Vedran Vukotić, Vivien Chappelier, and Teddy Furon. Are deep neural networks good for blind image watermarking? In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018.

[38] Erjin Bao, Ching-Chun Chang, Hanrui Wang, and Isao Echizen. Agentic copyright watermarking against adversarial evidence forgery with purification-agnostic curriculum proxy learning, 2025.

[39] the effectiveness of dataset wat.

[40] Yuqing Liang, Jiancheng Xiao, Wensheng Gan, and Philip S Yu. Watermarking techniques for large language models: A survey. *arXiv preprint arXiv:2409.00089*, 2024.

[41] Yiming Li, Linghui Zhu, Xiaojun Jia, Yang Bai, Yong Jiang, Shu-Tao Xia, Xiaochun Cao, and Kui Ren. Move: Effective and harmless ownership verification via embedded external features, 2025.

[42] Xi Yang, Kejiang Chen, Weiming Zhang, Chang Liu, Yuang Qi, Jie Zhang, Han Fang, and Nenghai Yu. Watermarking text generated by black-box language models. *arXiv preprint arXiv:2305.08883*, 2023.

[43] Lina Lin and Hanzhou Wu. Verifying integrity of deep ensemble models by lossless black-box watermarking with sensitive samples, 2022.

[44] MaungMaung AprilPyone and Hitoshi Kiya. Training dnn model with secret key for model protection, 2020.

[45] Jia Guo and Miodrag Potkonjak. Evolutionary trigger set generation for dnn black-box watermarking, 2021.

[46] Erwin Quiring and Konrad Rieck. Adversarial machine learning against digital watermarking. In *2018 26th European signal processing conference (EUSIPCO)*, pages 519–523. IEEE, 2018.

[47] Thibaud Gloaguen, Nikola Jovanović, Robin Staab, and Martin Vechev. Black-box detection of language model watermarks. *arXiv preprint arXiv:2405.20777*, 2024.

[48] Peizhuo Lv, Pan Li, Shenchen Zhu, Shengzhi Zhang, Kai Chen, Ruigang Liang, Chang Yue, Fan Xiang, Yuling Cai, Hualong Ma, et al. Ssl-wm: A black-box watermarking approach for encoders pre-trained by self-supervised learning. *arXiv preprint arXiv:2209.03563*, 2022.

[49] Yunfei Song, Yujia Zhu, Liu Yang, and Daoxun Xia. Black box watermarking for dnn model integrity detection using label loss. *Journal of Computers*, 35(4):277–290, 2024.

[50] Yifan Lu, Wenxuan Li, Mi Zhang, Xudong Pan, and Min Yang. Neural dehydration: Effective erasure of black-box watermarks from dnns with limited data. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 675–689, 2024.

[51] Haozhe Chen, Weiming Zhang, Kunlin Liu, Kejiang Chen, Han Fang, and Nenghai Yu. Speech pattern based black-box model watermarking for automatic speech recognition, 2022.

[52] Yiming Li, Mingyan Zhu, Xue Yang, Yong Jiang, Tao Wei, and Shu-Tao Xia. Black-box dataset ownership verification via backdoor watermarking. *IEEE Transactions on Information Forensics and Security*, 18:2318–2332, 2023.

[53] Dara Bahri, John Wieting, Dana Alon, and Donald Metzler. A watermark for black-box language models. *arXiv preprint arXiv:2410.02099*, 2024.

[54] Baizhou Huang, Xiao Pu, and Xiaojun Wan. $B\Theta$4: A black-box scrubbing attack on llm watermarks. *arXiv preprint arXiv:2411.01222*, 2024.

[55] Ryota Namba and Jun Sakuma. Robust watermarking of neural network with exponential weighting. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, pages 228–240, 2019.

[56] Dorjan Hitaj and Luigi V. Mancini. Have you stolen my model? evasion attacks against deep neural network watermarking techniques, 2018.

[57] Na Zhao, Kejiang Chen, Weiming Zhang, and Nenghai Yu. Performance-lossless black-box model watermarking. *arXiv preprint arXiv:2312.06488*, 2023.

[58] Haonan An, Guang Hua, Zhiping Lin, and Yuguang Fang. Box-free model watermarks are prone to black-box removal attacks. *arXiv preprint arXiv:2405.09863*, 2024.

[59] Mohammad Mehdi Yadollahi, Farzaneh Shoeleh, Sajjad Dadkhah, and Ali A. Ghorbani. Robust black-box watermarking for deep neuralnetwork using inverse document frequency, 2021.

[60] AprilPyone MaungMaung and Hitoshi Kiya. A protection method of trained cnn model with secret key from unauthorized access, 2021.

[61] Haodong Zhao, Jinming Hu, Peixuan Li, Fangqi Li, Jinrui Sha, Peixuan Chen, Zhuosheng Zhang, and Gongshen Liu. Nsmark: Null space based black-box watermarking defense framework for pre-trained language models. *arXiv preprint arXiv:2410.13907*, 2024.

[62] Jianwei Fei, Zhihua Xia, Benedetta Tondi, and Mauro Barni. Wide flat minimum watermarking for robust ownership verification of gans. *IEEE Transactions on Information Forensics and Security*, 2024.

[63] Masoumeh Shafieinejad, Jiaqi Wang, Nils Lukas, Xinda Li, and Florian Kerschbaum. On the robustness of the backdoor-based watermarking in deep neural networks, 2019.

[64] Dorjan Hitaj and Luigi V Mancini. Have you stolen my model? evasion attacks against deep neural network watermarking techniques. *arXiv preprint arXiv:1809.00615*, 2018.

[65] Hongbin Liu, Moyang Guo, Zhengyuan Jiang, Lun Wang, and Neil Gong. Audiomarkbench: Benchmarking robustness of audio watermarking. *Advances in Neural Information Processing Systems*, 37:52241–52265, 2024.

[66] Zhichuang Sun, Ruimin Sun, Long Lu, and Alan Mislove. Mind your weight(s): A large-scale study on insufficient machine learning model protection in mobile apps, 2021.

[67] Xudong Pan, Mi Zhang, Yifan Yan, Yining Wang, and Min Yang. Cracking white-box dnn watermarks via invariant neuron transforms. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1783–1794, 2023.

[68] Junlong Mao, Huiyi Tang, Yi Zhang, Fengxia Liu, Zhiyong Zheng, and Shanxiang Lyu. Fedreverse: Multiparty reversible deep neural network watermarking, 2023.

[69] Benedetta Tondi, Andrea Costanzo, and Mauro Barni. Robust and large-payload dnn watermarking via fixed, distribution-optimized, weights. *IEEE Transactions on Dependable and Secure Computing*, 2024.

[70] Erwan Le Merrer, Patrick Perez, and Gilles Trédan. Adversarial frontier stitching for remote neural network watermarking, 2019.

[71] Jialuo Chen, Jingyi Wang, Tinglan Peng, Youcheng Sun, Peng Cheng, Shouling Ji, Xingjun Ma, Bo Li, and Dawn Song. Copy, right? a testing framework for copyright protection of deep learning models. In *2022 IEEE symposium on security and privacy (SP)*, pages 824–841. IEEE, 2022.

[72] Qi Cui, Ruohan Meng, Chaohui Xu, and Chip-Hong Chang. Steganographic passport: An owner and user verifiable credential for deep model ip protection without retraining, 2024.

[73] Yiming Li, Yang Bai, Yong Jiang, Yong Yang, Shu-Tao Xia, and Bo Li. Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection. *Advances in Neural Information Processing Systems*, 35:13238–13250, 2022.

AI-generated, for reference only.

[74] Yifeng Gao, Yuhua Sun, Xingjun Ma, Zuxuan Wu, and Yu-Gang Jiang. Modellock: Locking your model with a spell, 2024.

[75] Minzhou Pan, Yi Zeng, Xue Lin, Ning Yu, Cho-Jui Hsieh, Peter Henderson, and Ruoxi Jia. Jigmark: A black-box approach for enhancing image watermarks against diffusion model edits. *arXiv preprint arXiv:2406.03720*, 2024.

[76] Alaa Fkirin, Gamal Attiya, Ayman El-Sayed, and Marwa A Shouman. Copyright protection of deep neural network models using digital watermarking: a comparative study. *Multimedia Tools and Applications*, 81(11):15961–15975, 2022.

[77] Huiying Li, Emily Wenger, Shawn Shan, Ben Y Zhao, and Haitao Zheng. Piracy resistant watermarks for deep neural networks. *arXiv preprint arXiv:1910.01226*, 2019.

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.

AI-generated, for reference only.