Intellectual Property Protection in Deep Learning: A Survey on Watermarking, Model Security, and Adversarial Defense

Abstract

Intellectual property protection in deep learning encompasses a range of strategies, including watermarking, model security, and adversarial defenses, to safeguard AI models against unauthorized use and theft. This survey paper examines the current landscape of intellectual property protection, highlighting key challenges and innovative strategies. Watermarking techniques, such as structural watermarking and digital passports, are essential for ownership verification and model security. However, existing methods face challenges from adversarial attacks and removal attempts, necessitating more robust and adaptive solutions. The integration of encryption techniques with watermarking enhances model security, while frameworks like AuthNet and NeRFProtector exemplify the use of embedded credentials to prevent unauthorized access. Real-world applications across industries, including healthcare, automotive, and entertainment, demonstrate the effectiveness of these strategies in protecting sensitive data and proprietary algorithms. The paper also explores the expansion of watermarking techniques to large language models and federated learning, addressing the unique challenges posed by these domains. As the field evolves, continuous innovation in evaluation practices and methodologies is crucial to ensure the resilience and effectiveness of intellectual property protection strategies. By advancing these techniques, the community can foster sustainable growth and innovation in artificial intelligence, maintaining the integrity and security of AI models in an increasingly adversarial environment.

1 Introduction

1.1 Importance of Intellectual Property Protection

Intellectual property protection is essential in deep learning due to the rising risk of model infringement, especially in image processing tasks [1]. The rapid evolution of watermarking techniques necessitates robust mechanisms to maintain the integrity and authenticity of AI models [2]. As deep neural networks (DNNs) become integral across applications, ensuring the security of model parameters and detecting unauthorized modifications is critical [3].

The surge of digital content has intensified challenges in verifying the provenance and authenticity of AI models, highlighting the need for effective protection strategies [4]. Watermarking is vital for provenance verification, defending against unauthorized redistribution and model theft [5]. Existing methods often fall short of security and robustness standards, urging the development of innovative solutions that do not alter model parameters [6].

In generative models, intellectual property protection addresses unauthorized reuse and ensures accountability in generated content [7]. The value of trained models as digital assets further underscores the necessity for strong safeguards against unauthorized use, reinforcing the significance of intellectual property protection in deep learning [8]. These measures not only secure proprietary technology but also foster innovation and sustainable development in the field.

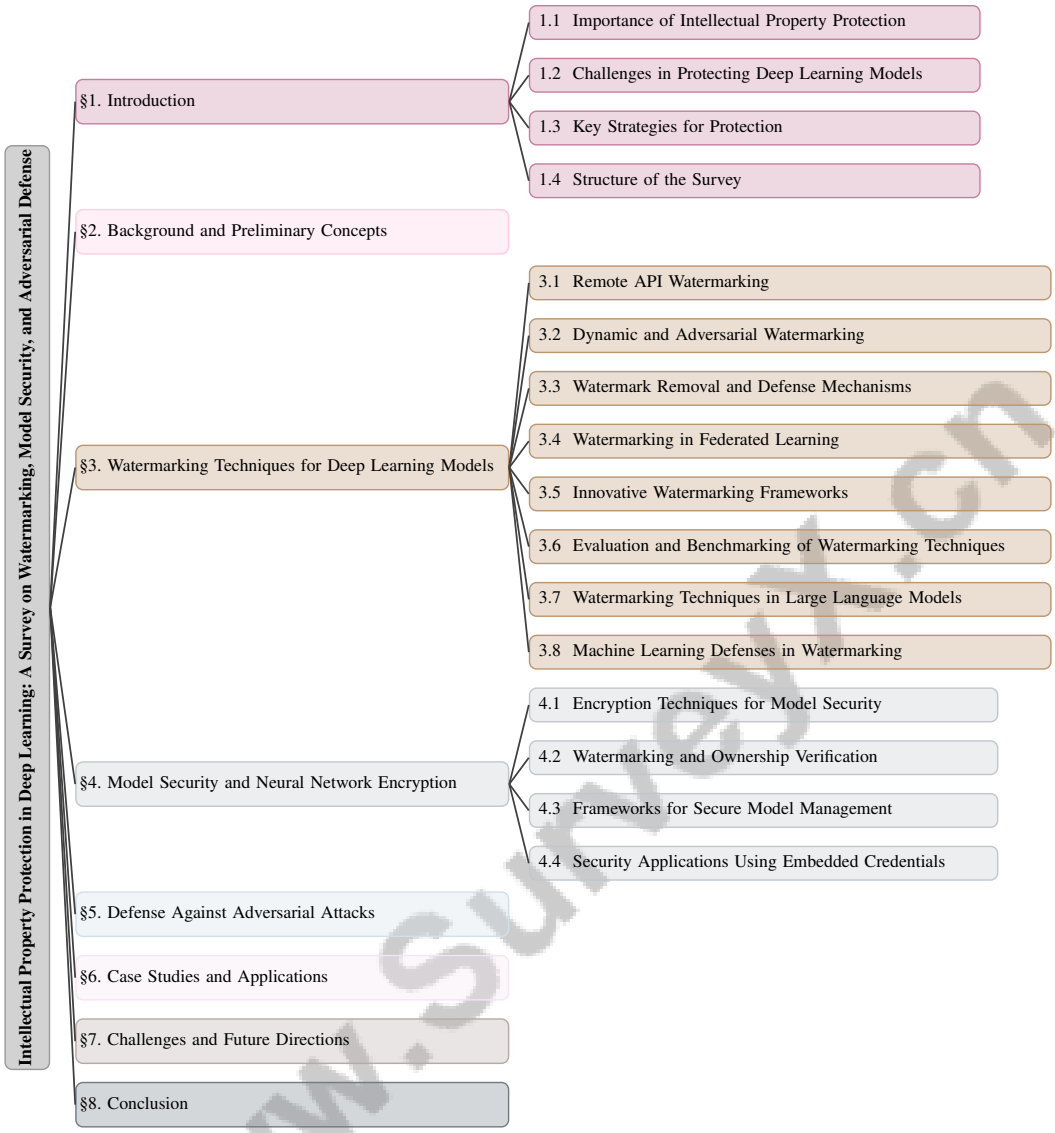


Figure 1: chapter structure

1.2 Challenges in Protecting Deep Learning Models

Protecting deep learning models from unauthorized access and theft presents significant challenges due to vulnerabilities in current watermarking and security techniques. A primary concern is the fragility of existing watermarking methods, which can be easily removed or fail to adequately secure against ownership claims by adversaries [6]. This vulnerability is exacerbated by functionality-stealing attacks, where adversaries replicate a model's functionality without direct access to the original model or its training data [9].

Attackers can also create surrogate models that closely mimic target models, even with limited access to their architecture and weights [1]. Current evaluation practices for trigger set-based watermarking methods are inadequate, undermining their reliability for ownership verification [10]. The lack of comprehensive, adaptive watermarking strategies to counter diverse and evolving threats to digital media remains a significant challenge [2].

Moreover, embedding watermarks can degrade neural network performance, complicating efforts to protect models without compromising functionality [8]. The security vulnerabilities of digital watermarking systems, particularly their susceptibility to copy and removal attacks, present substantial challenges to effective intellectual property protection [4]. Existing schemes often lack thorough

evaluations against a wide range of removal attacks, leading to uncertainty about their practical robustness [5].

In generative models, embedding watermarks without degrading output quality is crucial for effective authorship verification [7]. Additionally, the susceptibility of neural networks to backdoor attacks, where embedded watermarks can be exploited by adversaries, underscores the urgent need for more secure watermarking approaches [11].

These challenges highlight the critical need for innovative strategies to enhance the security and resilience of deep learning models. This is essential for safeguarding intellectual property rights in an increasingly adversarial landscape, where the risks of unauthorized reproduction, theft, and misuse of valuable deep neural networks are growing. Recent advancements in deep watermarking and fingerprinting techniques present potential solutions, yet the effectiveness and robustness of these methods against sophisticated evasion attacks remain key areas for further research and development [12, 13, 14, 15, 16].

1.3 Key Strategies for Protection

Intellectual property protection in deep learning is primarily achieved through three key strategies: watermarking, model security, and adversarial defenses. Watermarking techniques are vital for embedding hidden information within models to verify ownership and prevent unauthorized usage. Advanced methods, such as structural watermarking, utilize channel pruning to integrate watermarks into the architecture of deep neural networks (DNNs), enhancing robustness against typical attacks [17]. Additionally, incorporating the author's signature during training enables models to exhibit distinct behavior with signature-laden inputs, facilitating efficient ownership verification [18].

Innovative frameworks like ReDMark leverage Fully Convolutional Networks (FCNs) to develop new watermarking algorithms that operate in any desired transform space, improving the performance of existing techniques [19]. Furthermore, the Multi-view dATa (MAT) technique enhances watermarking effectiveness by using multi-view data to create robust trigger sets, making it more resilient against model extraction attacks [9]. A proposed deep watermarking framework embeds invisible watermarks into the outputs of image processing models, allowing for later extraction for ownership verification [1].

Model security is bolstered through encryption-based strategies and novel fingerprinting techniques. The introduction of pooled membership inference (PMI) allows for ownership determination while preserving the original DNN model, representing a new approach to intellectual property protection [20]. However, existing methods like authentication and encryption can be compromised by system breaches or malicious querying, emphasizing the need for more resilient security measures [21].

Adversarial defenses are crucial for protecting models against various attacks. The ROSE protocol, a lightweight and secure black-box DNN watermarking approach, employs cryptographic one-way functions and in-task key image-label pairs injected during training to enhance security [6]. Leveraging inherent model fingerprints for authorship verification offers an alternative to traditional watermarking techniques, which often compromise output quality [7].

These strategies highlight the importance of comprehensive protection mechanisms to secure deep learning models in an increasingly adversarial environment, addressing growing concerns over misuse and legal implications of AI-generated content. The integration of advanced watermarking techniques, including a novel deep watermarking framework that embeds invisible watermarks into outputs, alongside robust security frameworks, exemplifies innovative strategies employed to safeguard intellectual property in deep learning. This approach enhances model resilience against unauthorized use while adapting to various image characteristics, addressing the pressing concerns of intellectual property infringement in the field [13, 2, 22].

1.4 Structure of the Survey

This survey is meticulously structured to navigate the complex landscape of intellectual property protection in deep learning. We begin with an that underscores the critical importance of safeguarding AI models from unauthorized access and intellectual property theft, particularly given their status as valuable digital assets with commercial potential. This section lays the groundwork for subsequent discussions on innovative watermarking techniques, such as knowledge injection and inherent model

fingerprints, which aim to secure these models against misuse while maintaining output quality and ensuring the integrity of generated content [12, 7]. Following this, a detailed examination of the **Challenges in Protecting Deep Learning Models** delves into the vulnerabilities and limitations of current protection mechanisms.

In the **Key Strategies for Protection** section, we explore the primary methodologies employed to secure AI intellectual property, including watermarking, model security, and adversarial defenses. Each strategy is dissected to reveal its unique contributions and limitations within the context of deep learning.

The **Background and Preliminary Concepts** section provides foundational knowledge, defining essential terms and elucidating the roles of watermarking, model security, and adversarial attacks in the protection framework. This groundwork is crucial for understanding the advanced techniques discussed later.

We then transition into an in-depth analysis of **Watermarking Techniques for Deep Learning Models**, scrutinizing various methods for their effectiveness and limitations. This section also addresses innovative approaches such as remote API watermarking, dynamic and adversarial watermarking, and their applications in federated learning and large language models.

The focus shifts to **Model Security and Neural Network Encryption**, exploring strategies to enhance model security through encryption and the use of embedded credentials. The integration of watermarking techniques for ownership verification is explored in depth, highlighting innovative approaches that tackle the challenges of verifying intellectual property in generative models, particularly in the context of unauthorized reuse and the limitations of traditional methods. Various methodologies are presented, including the use of latent fingerprints in generative outputs and the novel Explanation as a Watermark (EaaW) paradigm, which embeds multi-bit watermarks within feature attribution explanations. Additionally, a new watermarking method for large language models (LLMs) utilizing knowledge injection is introduced, demonstrating high extraction success rates and robustness, thereby enhancing the reliability of ownership verification in AI-generated content [7, 23, 12].

In the survey titled , the authors comprehensively examine the multifaceted challenges presented by adversarial attacks on machine learning models, particularly in critical applications. They discuss various adaptive and robust defense strategies, including the integration of adversarial training and watermarking techniques, to enhance model resilience against evasion attacks and ensure the integrity and ownership of the models. This includes innovative approaches to mitigate vulnerabilities and improve the effectiveness of defenses against both model stealing and watermark removal attacks, thereby emphasizing the necessity for more resilient solutions in the face of evolving adversarial threats [24, 15, 25, 26].

Real-world applications and industry-specific implementations are showcased in the **Case Studies and Applications** section, providing practical insights into the deployment of protection strategies across various domains.

Finally, the survey concludes with a discussion on **Challenges and Future Directions**, identifying current obstacles in the field and proposing potential research avenues to advance the state of intellectual property protection in deep learning. This comprehensive roadmap navigates readers through the complex landscape of model protection, including innovative watermarking techniques and their applications in safeguarding intellectual property, while highlighting the necessity for continuous adaptation and advancement in response to evolving threats and opportunities within the field of machine learning [27, 28, 29, 30, 31].The following sections are organized as shown in Figure 1.

2 Background and Preliminary Concepts

2.1 Key Terms and Definitions

Watermarking in deep learning embeds identifiable information within neural network parameters to establish ownership and protect intellectual property, crucial for both model and content verification, especially in large language models [8]. The challenge lies in ensuring these identifiers do not

degrade model performance [6], as current techniques often lack robustness against manipulations, necessitating more effective methods to prevent watermark removal [11].

Model security involves measures to preserve the integrity and confidentiality of machine learning models, preventing unauthorized access and manipulation to safeguard data and intellectual property [6]. Neural network encryption, a subset of model security, encodes model parameters to prevent exploitation, especially in adversarial contexts.

Adversarial attacks exploit neural networks' vulnerabilities, leading to unauthorized and harmful outcomes, such as deepfake news and academic fraud, raising concerns about authenticity and ethics. Techniques like watermarking aim to embed identifiable markers in outputs for content verification and origin tracing, yet face challenges in robustness against adversarial attacks, emphasizing the need for resilient solutions [14, 15, 32]. Such attacks include model extraction, where adversaries replicate functionality by querying a model, risking intellectual property theft. Addressing these attacks is crucial for developing robust protection strategies, highlighting the vulnerabilities that must be addressed to secure AI models effectively.

2.2 Role of Watermarking and Model Security

Watermarking and model security are essential for protecting AI models from unauthorized use and intellectual property infringement. Watermarking techniques embed identifiable information within neural networks, crucial for ownership verification and traceability without degrading performance [33]. This is vital in scenarios where traditional methods may distort models, necessitating approaches that maintain integrity while securing ownership.

The effectiveness of watermarking is challenged by the need for both black-box and white-box detection, as existing methods may be vulnerable to attacks that limit authorship verification [18]. Embedding robust watermarks that withstand surrogate model attacks underscores watermarking's importance in AI model protection [1]. Developing benchmarks to evaluate DNN watermarking schemes' robustness against removal attacks provides a basis for assessing effectiveness [5].

Model security complements watermarking by ensuring AI models' integrity and confidentiality, particularly through evaluating ML-based digital watermarking techniques against adversarial attacks [4]. Comprehensive evaluations of watermarking methods in both white-box and black-box scenarios are crucial for understanding performance and resilience [34]. Innovative approaches, such as generative model fingerprinting, aim to address current watermarking techniques' limitations in effective authorship verification [7].

The significance of watermarking and model security is underscored by the necessity to verify DNN models' ownership without impairing performance, as traditional methods can distort models [20]. Benchmarks addressing ownership verification against adversarial attacks, which obscure or fraudulently claim ownership, are essential for advancing the field [10].

2.3 Adversarial Attacks and Defense Mechanisms

Adversarial attacks challenge the security and integrity of deep learning models by crafting inputs, known as adversarial examples, designed to deceive neural networks into incorrect predictions, exposing vulnerabilities [35]. These examples highlight the susceptibility of deep neural networks (DNNs) to subtle perturbations causing substantial misclassifications.

Adversarial attacks employ sophisticated strategies to exploit model weaknesses. For instance, the Easy Sample Matching Attack (ESMA) enhances adversarial transferability by selecting easy samples from the target class to guide perturbations, showcasing a strategic approach to maximizing adversarial attacks' impact [36]. Such methods underscore the adaptive nature of adversarial threats, necessitating advancements in defense mechanisms to safeguard AI models.

Defense mechanisms against adversarial attacks have evolved by integrating adversarial machine learning and digital watermarking techniques to fortify model security and enhance robustness [37]. This interdisciplinary approach is crucial for anticipating and mitigating future adversarial challenges, contributing to a deeper understanding of potential threats.

The DLOV E benchmark advances the evaluation of DNN-based watermarking techniques' robustness against adversarial attacks. By introducing the DLOV E attack, this benchmark provides a framework

for comparing models and techniques' effectiveness, facilitating more resilient defense strategies' development [38]. Systematic assessment and comparison of defense mechanisms are essential for identifying current approaches' strengths and weaknesses, guiding future research directions.

3 Watermarking Techniques for Deep Learning Models

Exploring watermarking techniques in deep learning models necessitates an understanding of various methodologies aimed at ensuring the security and integrity of these systems. This section particularly emphasizes remote API watermarking, a significant approach that employs application programming interfaces to embed watermarks in deep learning models. This technique facilitates ownership verification and safeguards against unauthorized usage, laying the groundwork for integrating watermarking within deep learning applications. Table 4 provides a comprehensive summary of various remote API watermarking methods, elucidating their respective security measures, integration techniques, and application domains in the context of deep learning models.

Figure 2 illustrates the hierarchical categorization of watermarking techniques in deep learning models, encompassing remote API watermarking, dynamic and adversarial methods, watermark removal and defense mechanisms, federated learning applications, innovative frameworks, evaluation and benchmarking practices, techniques for large language models, and machine learning defenses. Each category is further divided into specific methods, innovations, challenges, and solutions, providing a comprehensive overview of the current landscape and advancements in watermarking strategies for deep learning. This visual representation enhances our understanding of the multifaceted approaches to watermarking, highlighting the interconnectedness of various methodologies and their respective contributions to the field.

3.1 Remote API Watermarking

Method Name	Integration Techniques	Security Measures	Application Domains
LBW-DEM[39]	Black-box Watermarking	Sensitive Samples	Ensemble Models
DP[40]	Digital Passports	Digital Passports	Text, Images, Audio
ROSE[6]	Cryptographic Hashes	Digital Passports	Text, Images
RD[19]	Deep Learning Models	Differentiable Attack Layer	Grayscale Images

Table 1: Summary of remote API watermarking methods detailing their integration techniques, security measures, and application domains in the context of deep learning models. The table highlights diverse approaches, including black-box watermarking, digital passports, and cryptographic hashes, demonstrating their application across various media types such as text, images, and audio.

Method Name	Integration Techniques	Security Measures	Application Domains
LBW-DEM[39]	Black-box Watermarking	Sensitive Samples	Ensemble Models
DP[40]	Digital Passports	Digital Passports	Text, Images, Audio
ROSE[6]	Cryptographic Hashes	Digital Passports	Text, Images
RD[19]	Deep Learning Models	Differentiable Attack Layer	Grayscale Images

Table 2: Summary of remote API watermarking methods detailing their integration techniques, security measures, and application domains in the context of deep learning models. The table highlights diverse approaches, including black-box watermarking, digital passports, and cryptographic hashes, demonstrating their application across various media types such as text, images, and audio.

Table 2 provides a comprehensive summary of various remote API watermarking methods, elucidating their respective security measures, integration techniques, and application domains in the context of deep learning models. Remote API watermarking is pivotal for embedding watermarks into deep learning models via their APIs, enabling ownership verification and protection against unauthorized use. This approach modifies decision boundaries, allowing specific queries to detect watermark presence without altering the original model. The Lossless Black-box Watermarking for DEMs (LBW-DEM) method exemplifies this by verifying the integrity of deep ensemble models without modifying the original model [39].

The integration of digital passports into deep neural network (DNN) models further showcases advanced watermarking techniques in remote API environments. This method embeds digital passports, ensuring normal operation only with valid credentials and disrupting unauthorized networks [40].

Figure 2: This figure illustrates the hierarchical categorization of watermarking techniques in deep learning models, encompassing remote API watermarking, dynamic and adversarial methods, watermark removal and defense mechanisms, federated learning applications, innovative frameworks, evaluation and benchmarking practices, techniques for large language models, and machine learning defenses. Each category is further divided into specific methods, innovations, challenges, and solutions, providing a comprehensive overview of the current landscape and advancements in watermarking strategies for deep learning.

Such innovations highlight the adaptability of remote API watermarking across diverse applications, enhancing model security and intellectual property protection.

Recent surveys affirm the effectiveness of various watermarking techniques across modalities such as text, images, and audio, emphasizing their robustness [41]. The ROSE protocol, a black-box DNN watermarking method, employs secret key trigger-label pairs during training, facilitating ownership verification during testing and further illustrating the utility of remote API watermarking in safeguarding AI models [6].

Moreover, the ReDMark framework utilizes Fully Convolutional Neural Networks with residual structures for blind secure watermarking, demonstrating its potential in remote API scenarios across various transform domains [19]. This adaptability is crucial for maintaining watermark integrity against unauthorized modifications.

Remote API watermarking is rapidly evolving, incorporating sophisticated techniques to enhance model security against adversarial threats. Innovations include self-watermarking methods leveraging latent fingerprints for ownership verification, adaptive watermarking strategies preserving content quality while defending against model extraction attacks, and knowledge injection techniques embedding watermarks directly within model architectures. These developments address critical issues such as misinformation and academic misconduct, ensuring the integrity of generated content and the authenticity of its sources [7, 27, 42, 12]. By leveraging these approaches, remote API watermarking

AI-generated, for reference only.

provides a robust layer of intellectual property protection, securing deep learning models against unauthorized access.

3.2 Dynamic and Adversarial Watermarking

Dynamic and adversarial watermarking techniques are essential for safeguarding deep learning models, employing innovative methodologies to enhance watermark robustness and stealth. For instance, the DynaMarks approach dynamically alters model output responses to embed watermarks in surrogate models without modifying the original training process, effectively protecting intellectual property against model extraction attacks. Similarly, the Adversarial Watermarking Transformer (AWT) employs a jointly trained encoder-decoder framework to unobtrusively encode messages within generated text while minimizing alterations to semantics, thus maintaining utility and resisting adversarial attacks [43, 2, 32]. These advancements not only enhance watermark security but also expand future research potential in deep learning-based watermarking.

The DAWN framework exemplifies dynamic adversarial watermarking by altering predictions for a fraction of queries made to a model's prediction API, effectively embedding watermarks and adapting to evolving adversarial attacks [44]. This technique highlights the potential of dynamic watermarking in addressing adversarial challenges.

Integrating improved discrete wavelet transform (DWT) and discrete cosine transform (DCT) watermarking algorithms marks a significant innovation in generating adversarial examples, providing a sophisticated approach to watermarking in adversarial contexts [35]. By leveraging these transformations, watermarking methods achieve greater resilience against adversarial manipulations.

Furthermore, an adversarial training framework combining watermarking and overwriting networks showcases a competitive training process, enhancing watermarking robustness through continuous adaptation [45]. The U-Net based architecture trained through adversarial learning effectively generates fake watermarked images, demonstrating the application of adversarial techniques in watermarking [46].

3.3 Watermark Removal and Defense Mechanisms

Watermark removal and defense mechanisms are crucial for ensuring the security and integrity of deep learning models against unauthorized exploitation. Traditional watermark removal techniques often require access to original training samples or specific knowledge of the watermarking scheme, limiting their practicality [47]. This necessitates robust defense mechanisms capable of withstanding removal attacks.

The laundering algorithm proposed by Aiken et al. systematically removes black-box watermarks from neural networks, allowing adversaries to regain control over the model without significant accuracy loss [11]. This highlights vulnerabilities in current watermarking techniques and the need for resilient defenses. The Dehydra method exemplifies a watermark-agnostic removal attack, exploiting DNN internals to recover and unlearn watermark messages [48].

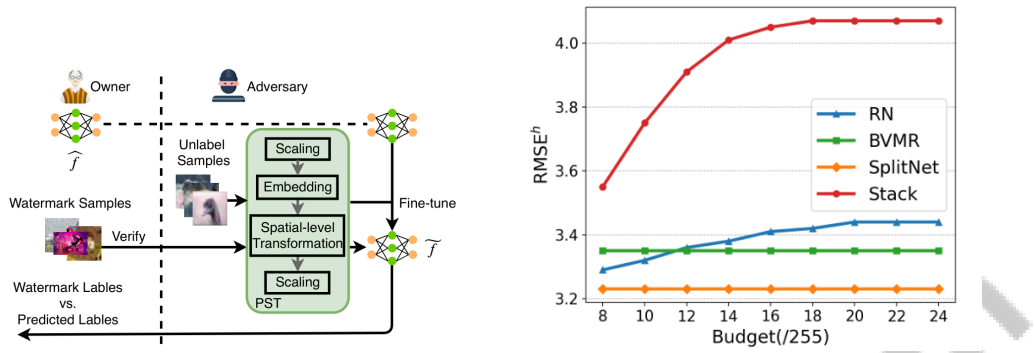
Innovative defense strategies have emerged in response to these vulnerabilities. The Watermark Vaccine method generates adversarial perturbations applied to host images prior to watermarking, enhancing watermark resilience against removal attempts [49]. Additionally, the UBW method creates poisoned datasets leading to dispersible predictions, complicating watermark removal [50].

Xiang et al. propose a method allowing ownership verification without compromising model performance, successfully evading detection by existing algorithms, thus demonstrating a sophisticated approach to watermark resilience [51]. The EaaW method embeds a multi-bit watermark within feature attribution of specific trigger samples, enabling ownership verification without impacting model performance [23].

Cryptographic techniques in watermarking, as highlighted by Li et al., secure watermarks against knowledgeable adversaries while maintaining neural network performance during watermarking [52]. This approach is crucial for protecting watermarks in adversarial environments.

The effectiveness of radioactive data, demonstrated by Tekgul et al., lies in its ability to maintain the integrity of true labels while embedding watermarks robustly against adversarial modifications

and visual detection [53]. Such methods are essential for ensuring watermarking robustness against removal attempts.



(a) Watermark Detection with Spatial-Level Transformation[54]

(b) Comparison of RMSEh values across different budget sizes for different models[49]

Figure 3: Examples of Watermark Removal and Defense Mechanisms

As shown in Figure 3, watermarking techniques in deep learning serve as crucial tools for protecting intellectual property and verifying model authenticity. Figure 1 illustrates two significant aspects of this domain. Subfigure (a) showcases a watermark detection system employing spatial-level transformations, highlighting interactions between the owner's network and the adversary's network. Subfigure (b) presents a line graph comparing Root Mean Squared Error (RMSE) values across models and budget sizes, providing insights into the effectiveness of various watermarking strategies. Analyzing RMSE values aids researchers in gauging model robustness against adversarial attacks, contributing to the development of more resilient watermarking techniques.

3.4 Watermarking in Federated Learning

Watermarking in federated learning introduces unique challenges and opportunities due to decentralized model training across multiple devices, enhancing privacy while increasing the risk of model theft. Traditional watermarking techniques, designed for centralized training, face limitations, necessitating innovative approaches like WAFFLE. This method incorporates a retraining step at the server after aggregating local models, embedding a resilient watermark without requiring access to training data, while incurring minimal degradation in model accuracy [55, 56]. As federated learning evolves, addressing these constraints is crucial for protecting intellectual property in sensitive applications.

WAFFLE exemplifies a robust approach to embedding watermarks in DNN models trained via federated learning. This method involves server-side retraining to embed a resilient watermark after aggregating local models, enhancing protection against unauthorized usage and ensuring traceability [56]. By leveraging server-side capabilities, WAFFLE effectively incorporates watermarks without compromising federated learning's distributed nature.

The survey by Lansari et al. provides an extensive overview of watermarking techniques applicable to federated learning, highlighting methods, limitations, and security considerations [55]. This analysis underscores the need for innovative watermarking strategies addressing challenges posed by federated environments, such as local data heterogeneity and potential adversarial attacks during model aggregation.

FedIPR introduces a unified framework facilitating the embedding of feature-based and backdoor-based watermarks within federated learning systems [57]. This dual approach enables robust ownership verification while maintaining federated model integrity and performance. By allowing flexible watermarking strategies, FedIPR enhances adaptability to diverse federated learning scenarios.

Integrating watermarking techniques within federated learning safeguards intellectual property rights by enabling model ownership verification while enhancing the overall security and trustworthiness of distributed AI systems. This approach addresses challenges posed by federated learning, such as model theft and unauthorized redistribution, by embedding private watermarks independently verifiable by participants without exposing sensitive training data. Recent advancements, including

frameworks like Merkle-Sign and WAFFLE, demonstrate the effectiveness of these techniques in ensuring ownership verification and protecting DNNs against malicious attacks, fostering a more secure collaborative environment for AI development [55, 56, 57, 58]. As federated learning gains traction across domains, developing effective watermarking methods remains critical for secure and privacy-preserving AI solutions.

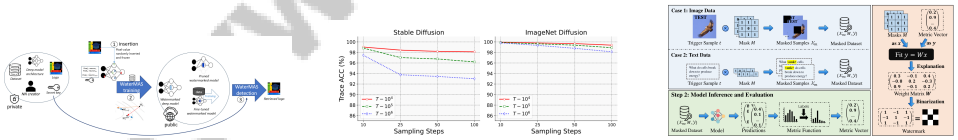
3.5 Innovative Watermarking Frameworks

Innovative watermarking frameworks have emerged as vital solutions for enhancing robustness and effectiveness in intellectual property protection for deep learning models. These frameworks address critical challenges in AI security, including detectability, resilience against removal attacks, and maintaining AI model functionality. By integrating advanced watermarking techniques and adversarial training, they strengthen AI models against unauthorized exploitation and adversarial threats, ultimately safeguarding intellectual property and ensuring content authenticity in applications such as large language models (LLMs) [12, 27, 15, 25].

A notable framework employs parameter regularizers to embed watermarks directly into the training process, preserving the model's original task performance while seamlessly integrating the watermark [8]. This approach enhances watermark stealth and persistence, making it a robust solution for intellectual property protection.

The three-step laundering process, including watermark recovery, neuron resetting, and retraining, represents a significant advancement in addressing watermark removal complexities [11]. This method effectively counters challenges associated with removing embedded watermarks, reinforcing model security.

These innovative frameworks collectively represent significant advancements in watermarking technologies. By addressing critical challenges and improving robustness and effectiveness in intellectual property protection, these frameworks underscore the need for continued innovation in watermarking techniques for diverse applications. The focus on cross-disciplinary approaches emphasizes the critical need to integrate insights from various fields, such as watermarking techniques in AI-generated content and deep learning methodologies for text detection, to enhance the robustness, security, and operational effectiveness of AI models. This integration addresses vulnerabilities related to intellectual property theft and content misattribution, enhancing the reliability of distinguishing between human-authored and AI-generated texts, ultimately leading to more secure and functional AI applications [12, 59, 29].



(a) Watermarking and Detection of Watermarked Images Using Deep Neural Networks[60]

(b) Comparison of Trace Accuracy for Stable Diffusion and ImageNet Diffusion[61]

(c) Masked Data Analysis for Text and Image Data[23]

Figure 4: Examples of Innovative Watermarking Frameworks

As shown in Figure 4, watermarking techniques have become essential in securing intellectual property and ensuring model integrity. The examples presented highlight three distinct frameworks illustrating the sophistication of modern watermarking strategies. The first example illustrates a comprehensive process involving the insertion of pixel values, training a watermarked model, and subsequent detection using fine-tuned models, ensuring robust watermark embedding. The second example compares trace accuracy between diffusion models, providing insights into watermark traceability effectiveness across architectures. The third example demonstrates masked data handling, crucial for maintaining data privacy while enabling model training. These frameworks not only enhance deep learning model security but also pave the way for advancements in model protection techniques.

AI-generated, for reference only.

Benchmark	Size	Domain	Task Format	Metric
ML-Watermark-Benchmark[4]	800	Digital Watermarking	Watermark Detection	Probability of Miss, Bit Error Rate
WMG[15]	296	Text Generation	Watermarking Evaluation	Quality Score, Watermark Rate
EGG[62]	12,000	Image Processing	Image Deraining	PSNR, MS-SSIM
WILD[47]	60,000	Computer Vision	Watermark Removal	Test Accuracy, Watermark Retention
BW-Benchmark[63]	60,000	Image Classification	Watermark Detection	Accuracy, Watermark Retention
LLM-WM[64]	500	Watermarking	Watermark Detection	Z-score, PPL
LLM-WM[30]	40,000	Natural Language Processing	Watermark Removal	Precision, GRR
VPB[65]	1,000	Image Watermarking	Watermark Detection	CMMD, Detectability Rate

Table 3: This table provides a comprehensive comparison of various benchmarks used in the evaluation of watermarking techniques across different domains. It details the size, domain, task format, and performance metrics for each benchmark, offering insights into their applicability and effectiveness in assessing watermarking strategies. Such a comparative overview is essential for understanding the strengths and limitations of each benchmark in the context of watermark detection and removal tasks.

3.6 Evaluation and Benchmarking of Watermarking Techniques

Table 3 offers a detailed comparison of key benchmarks utilized in the evaluation of watermarking techniques, highlighting their respective domains, task formats, and performance metrics.

Evaluating and benchmarking watermarking techniques are pivotal in assessing their effectiveness, robustness, and applicability across diverse deep learning scenarios. These assessments confirm that advanced watermarking strategies can effectively protect AI models' intellectual property by embedding latent fingerprints in generated outputs, ensuring model performance and resilience against various adversarial attacks, including watermark removal and forging. This approach enhances ownership verification and mitigates risks associated with unauthorized data reuse and misinformation, reinforcing the integrity of AI-generated content [12, 59, 7, 42, 41].

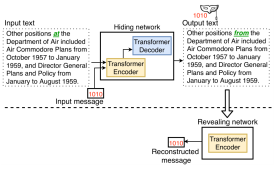
A comprehensive evaluation involves several key metrics and methodologies. The survey by Zhong et al. categorizes methodologies within different frameworks, emphasizing their effectiveness and unique approaches to embedding and extracting watermarks [2]. This comparative analysis is crucial for understanding each technique's strengths and limitations, providing a roadmap for future improvements.

Performance metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) are commonly used to evaluate image quality in watermarking applications. The RedMark framework utilizes these metrics alongside Bit Error Rate (BER) to assess watermark extraction accuracy under various attack conditions [19]. Such metrics are essential for quantifying the impact of watermarking on model outputs and ensuring fidelity preservation.

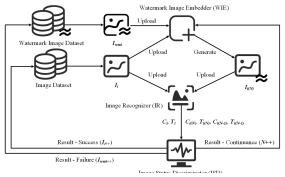
The introduction of novel benchmarks, focusing on the latent space of foundation models, systematically evaluates watermarking techniques against emerging attack strategies [4]. These benchmarks highlight model vulnerabilities and the need for robust defense mechanisms, facilitating the development of more resilient watermarking methods.

The ROSE method exemplifies the importance of evaluating watermark recovery rates and test accuracy under different attack scenarios, including pruning, weight quantization, and JPEG compression [6]. Such evaluations are critical for understanding watermarking techniques' robustness and their ability to withstand adversarial conditions.

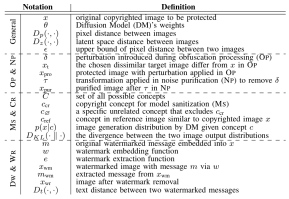
As shown in Figure 5, the exploration of watermarking techniques for deep learning models provides a comprehensive overview through three distinct visual representations. The first image illustrates a flowchart depicting the text encryption and decryption process using a Transformer Encoder and Decoder, foundational for integrating watermarking into text processing systems. The second image introduces a watermark image embedding system, composed of a watermark image dataset, an image recognizer, and an image status discriminator, training models to recognize and embed watermarks within images. The third image presents a table titled "Notation Definition," categorizing various notations and definitions, providing a structured framework for understanding watermarking termi-



(a) The image shows a flowchart of a text encryption and decryption process using a Transformer Encoder and Decoder.[32]



(b) Watermark Image Embedding System[35]



	Notation	Definition
General	x	original copyrighted image to be protected
	θ	Diffusion Model (DM)'s weights
	$D_S(\cdot,\cdot)$	pixel distance between images
	$D_z(\cdot,\cdot)$	latent space distance between images
	ϵ	upper bound of pixel distance between two images
Obf. & Pur.	δ	perturbation introduced during obfuscation processing (Obf)
	x_t	the chosen dissimilar target image differ from x in Obf
	x_{pro}	protected image with perturbation applied in Obf
	τ	transformation applied in noise purification (NP) to remove δ
	x_{pur}	purified image after τ in NP
Ms. & Cn.	$\mathcal{C}$	set of all possible concepts
	c_{cr}	copyright concept for model sanitization (Ms)
	c_{cr}	a specific unrelated concept that excludes c_{cr}
	c_{ref}	concept in reference image similar to copyrighted image x
	$p(x\|c)$	image generation distribution by DM given concept c
	$D_{KL}(\cdot\|\|\cdot)$	the divergence between the two image output distributions
Dw. & Wr.	m	original watermarked message embedded into x^w
	w	watermark embedding function
	e	watermark extraction function
	x_{wm}	watermarked image with message m via w
	m_{wm}	extracted message from x_{wm}
	x_{rm}	image after watermark removal
	$D_t(\cdot,\cdot)$	text distance between two watermarked messages

(c) The image is a table with a header row and multiple rows of data.[66]

Figure 5: Examples of Evaluation and Benchmarking of Watermarking Techniques

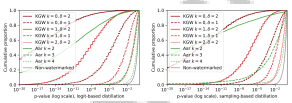
nologies. Together, these visual elements facilitate the evaluation and benchmarking of watermarking techniques within deep learning models.

3.7 Watermarking Techniques in Large Language Models

Watermarking large language models (LLMs) presents unique challenges due to their complex architectures and vast data processing capabilities. These models are particularly susceptible to misuse, making effective watermarking strategies crucial for ensuring intellectual property protection and ownership verification [67]. The inherent complexity of LLMs necessitates robust watermarking techniques that can withstand systematic removal attacks while maintaining model performance.

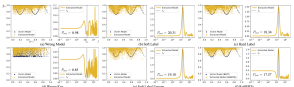
Recent benchmarks reveal vulnerabilities in existing watermarking techniques for LLMs, underscoring the need for improved strategies enhancing watermark resilience [30]. The effectiveness of proposed watermark removers highlights significant vulnerabilities in victim models, demonstrating the ease with which adversaries can compromise watermark integrity and ownership verification [62].

To address these challenges, innovative approaches like WaterPool have been developed to mitigate trade-offs between watermark robustness and model performance. By strategically embedding watermarks in ways that do not degrade functionality, WaterPool offers a promising solution to LLM watermarking challenges [67]. This approach exemplifies the need for continuous innovation in watermarking techniques to protect the intellectual property of large language models effectively.



(a) Comparison of Logit-Based and Sampling-Based Distillation Methods for Watermarking[68]

(b) A diagram illustrating a neural network architecture for data hiding and revealing[32]



(c) Model Selection and Labeling in Deep Learning: A Comparative Study[69]

Figure 6: Examples of Watermarking Techniques in Large Language Models

As shown in Figure 6, watermarking techniques have emerged as crucial tools for securing intellectual property and ensuring model integrity, particularly in large language models. The figure provides an overview of three distinct approaches to watermarking in this domain. The first subfigure illustrates a comparative analysis of logit-based and sampling-based distillation methods, highlighting their efficacy in watermarking. The second subfigure delves into a neural network architecture designed for data hiding and revealing, showcasing components like an encoder-decoder transformer. The third subfigure presents a study on model selection and labeling, contrasting the performance of different models under various conditions. Together, these examples illustrate diverse methodologies and applications of watermarking in safeguarding advancements in large language models.

3.8 Machine Learning Defenses in Watermarking

Machine learning techniques provide a robust framework for defending against watermark removal, enhancing the resilience and integrity of watermarking strategies in deep learning models. These techniques harness machine learning's adaptive capabilities to respond dynamically to evolving threats, safeguarding intellectual property in a complex environment characterized by sophisticated attacks and effective defenses, as highlighted by recent research categorizing threats and proposing protection mechanisms such as watermarking and fingerprinting for machine learning models [15, 28].

One approach involves adversarial training methods, enhancing watermark robustness by simulating potential removal attacks during training [45]. Incorporating adversarial examples into training allows models to recognize and resist watermark removal attempts, maintaining ownership verification and model integrity.

The integration of machine learning with digital watermarking also facilitates adaptive defense mechanisms. Techniques like the Watermark Vaccine method utilize adversarial perturbations to fortify watermarks against removal attempts, demonstrating machine learning's potential to enhance watermarking security [49]. This method exemplifies the innovative application of machine learning in creating robust defenses against adversarial threats.

Moreover, employing neural network architectures trained through adversarial learning offers a promising avenue for watermark defense. Models generating adversarially robust watermarks ensure watermark integrity against sophisticated removal attacks [46]. This approach highlights the strategic use of machine learning to reinforce watermark resilience and protect intellectual property.

Investigating machine learning defenses in watermarking underscores the critical need for ongoing innovation to develop more resilient protection mechanisms. This is particularly important as emerging techniques, such as trigger set watermarking and knowledge injection methods, reveal vulnerabilities to adversarial attacks, including watermark removal and forging. Recent advancements demonstrate that leveraging diffusion models and deep learning architectures can enhance watermarking effectiveness, ensuring better intellectual property protection for machine-generated content and deep learning models [12, 22, 59, 42, 15]. By leveraging machine learning's adaptive and predictive capabilities, researchers can create more resilient watermarking techniques that effectively counteract removal attempts, safeguarding the integrity and ownership of deep learning models in a rapidly evolving adversarial environment.

Feature	Remote API Watermarking	Dynamic and Adversarial Watermarking	Watermark Removal and Defense Mechanisms
Integration Technique	Api Embedding	Dynamic Alteration	Laundering Algorithm
Security Measure	Ownership Verification	Adversarial Resistance	Resilience Enhancement
Application Domain	Deep Learning Models	Surrogate Models	Model Security

Table 4: This table presents a comparative analysis of three distinct watermarking methodologies used in deep learning models: Remote API Watermarking, Dynamic and Adversarial Watermarking, and Watermark Removal and Defense Mechanisms. Each method is evaluated based on its integration technique, security measure, and application domain, providing a comprehensive overview of their respective strengths and applications in enhancing model security and intellectual property protection.

4 Model Security and Neural Network Encryption

4.1 Encryption Techniques for Model Security

Encryption techniques are vital for enhancing the security of deep learning models by protecting the confidentiality of model parameters and reinforcing intellectual property rights. These methods often integrate with watermarking strategies to create a robust security framework. A notable approach is the AuthNet mechanism, which embeds authentication logic into deep learning models by repurposing redundant neurons to generate authentication bits, thereby providing native authentication [21].

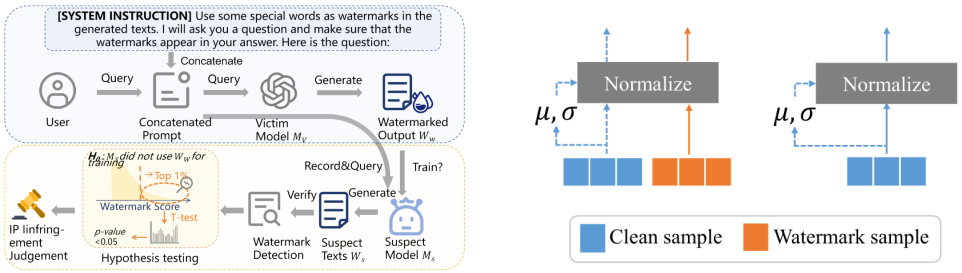
The Deep Serial Number (DSN) framework serves as a watermarking technique that embeds unique serial numbers into deep neural networks (DNNs), ensuring that models operate only with the correct serial number input [70]. This highlights the synergy between encryption and watermarking, which protects model outputs and prevents unauthorized use. Similarly, the Digital Watermarking for

Deep Neural Networks (DWN) method embeds a watermark into model parameters during training, ensuring its resilience even after fine-tuning or parameter pruning [8].

Advanced methods include the digital passporting technique, which alters DNN behavior based on the validity of a digital passport, providing immediate protection against unauthorized usage [40]. The EncryIP framework exemplifies encryption's potential by generating multiple protected versions of a model from a single training iteration, facilitating secure access to model outputs [71].

Moreover, the NeRFProtector integrates watermarking and encryption by embedding binary messages into Neural Radiance Fields (NeRFs) during optimization [72]. The DeepSigns framework employs a masking strategy during training to distribute watermark information across the model, enhancing its robustness against removal [73].

Collectively, encryption techniques and watermarking strategies significantly bolster the confidentiality, integrity, and ownership verification of deep learning models, addressing threats such as unauthorized model extraction and copyright infringement. Recent advancements in adaptive watermarking frameworks maintain model performance while enhancing resilience against adversarial attacks, ensuring the integrity of deep learning models in an increasingly hostile environment [27, 74, 13, 42, 14].



(a) A System for Detecting IP Infringement Judgments Using Watermarking[27]

(b) Watermark Detection in Normalized Images[75]

Figure 7: Examples of Encryption Techniques for Model Security

As shown in Figure 7, innovative techniques for neural network encryption and model security continue to evolve, safeguarding intellectual property and ensuring the integrity of machine learning models. The first example, "A System for Detecting IP Infringement Judgments Using Watermarking," illustrates a framework designed to identify unauthorized use of intellectual property by processing user queries and verifying outputs against suspect models trained on potentially infringing texts. The second example, "Watermark Detection in Normalized Images," demonstrates watermarking in image processing, where normalization differentiates between 'Clean sample' and 'Watermark sample.' These examples collectively underscore the potential of encryption techniques to fortify neural networks against unauthorized exploitation and preserve intellectual property [27, 75].

4.2 Watermarking and Ownership Verification

Watermarking is a crucial technique for verifying model ownership, providing a robust mechanism to protect against unauthorized use and model theft while preserving performance on original tasks [76]. By embedding identifiable information within deep learning models, watermarking facilitates ownership verification, ensuring the watermark remains intact despite modifications [33]. This capability is essential for safeguarding intellectual property in AI models, particularly in adversarial contexts.

The integration of watermarking with authentication mechanisms, such as the AuthNet framework, enhances model security by embedding authentication logic directly into the model, blocking unauthorized access without additional layers, thus preserving performance [21]. Techniques like the MAT watermarking method further strengthen ownership verification for deep neural networks (DNNs), ensuring ownership can be asserted even against adaptive attacks [9].

A comprehensive evaluation of multiple watermarking techniques under various conditions, as documented by Chen et al., allows for a thorough comparison of their effectiveness in ownership

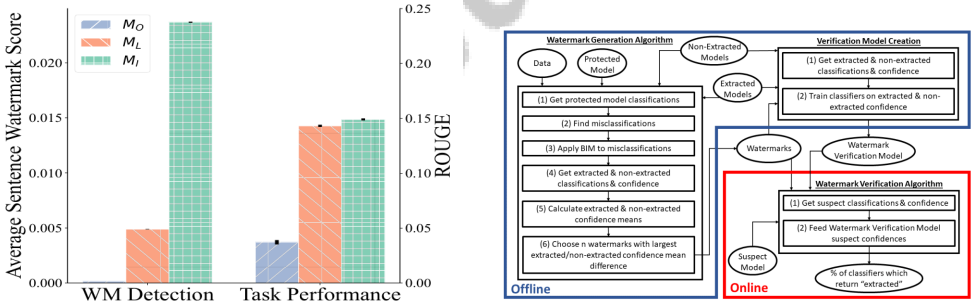
verification [34]. This analysis is crucial for identifying the strengths and limitations of different watermarking approaches, guiding future advancements in the field.

4.3 Frameworks for Secure Model Management

Frameworks for secure model management are vital for ensuring the integrity and security of deep learning models, especially in environments where computational resources and data processing loads must be optimized. The Diffusion-based Robust Adversarial Attack (DRAA) framework, for example, focuses on monitoring data processing loads and redistributing computational resources to enhance performance [77]. This approach optimizes model operations while reinforcing security by dynamically adjusting to varying demands.

Integrating model management frameworks with robust watermarking techniques further strengthens the security posture of AI models. These frameworks manage the lifecycle of models, including deployment, monitoring, and updating, while ensuring embedded watermarks remain intact and effective against unauthorized access or modifications. By employing adaptive strategies that respond to evolving adversarial threats, these frameworks significantly enhance the resilience of machine learning models against security breaches. For instance, self-watermarking mechanisms allow models to autonomously embed identifiers into their outputs, safeguarding intellectual property while maintaining content quality. Evaluations across diverse datasets demonstrate that these frameworks consistently outperform existing solutions, effectively countering model extraction and watermark removal attacks [27, 25, 59, 78, 15].

The emphasis on resource optimization in frameworks like DRAA underscores the importance of balancing performance with security. By leveraging advanced monitoring and redistribution techniques, these frameworks ensure efficient model operation without compromising integrity or exposing vulnerabilities. This balance is critical for maintaining the reliability of AI systems, particularly in increasingly intricate and adversarial contexts where intellectual property theft and content authenticity are significant concerns. Effective watermarking techniques that embed identifiable markers within AI-generated outputs play a crucial role in this process by enabling verification and tracing of content origins, thereby reinforcing user trust and ethical standards in AI applications [12, 15].



(a) Comparison of Sentence Watermark Detection and Task Performance Metrics[27]

(b) Watermarking for Model Classification Protection[78]

Figure 8: Examples of Frameworks for Secure Model Management

As shown in Figure 8, ensuring the security and integrity of machine learning models is paramount. The example titled "Model Security and Neural Network Encryption; Frameworks for Secure Model Management" delves into innovative approaches for safeguarding these models. The first figure, "Comparison of Sentence Watermark Detection and Task Performance Metrics," presents a comparative analysis of performance metrics—M$_O$, M_L, $and M_I$ $^-across watermark detection and general task per formance. This visual comparison highlight 78].

4.4 Security Applications Using Embedded Credentials

The incorporation of embedded credentials into deep learning models represents a significant advancement in model security, providing sophisticated methods to protect intellectual property against

AI-generated, for reference only.

unauthorized access and exfiltration. Given the vulnerabilities of deep neural networks to intellectual property theft, where attackers can replicate models through fine-tuning or surrogate training, innovative techniques like deep watermarking and integrated authentication mechanisms enhance resilience, ensuring unauthorized users cannot effectively utilize compromised models without the appropriate credentials [21, 13, 79]. Embedded credentials function as digital signatures or watermarks, enabling models to authenticate access and verify ownership, thus preventing unauthorized use and tampering.

A notable application of embedded credentials is the Digital Passport method, which embeds a digital passport within the neural network, allowing the model to function normally only when the correct credentials are presented [40]. This effectively paralyzes unauthorized networks, ensuring that only entities with valid credentials can access the model's functionalities.

The AuthNet framework exemplifies the use of embedded credentials by integrating authentication logic directly into the model, utilizing redundant neurons to create authentication bits [21]. This method enhances model security by providing native authentication capabilities, eliminating the need for additional security layers while maintaining performance.

Furthermore, the NeRFProtector framework embeds binary messages into Neural Radiance Fields (NeRFs) during optimization, ensuring model outputs are protected against unauthorized use [72]. This approach underscores the potential of combining watermarking with embedded credentials to secure model outputs in complex environments.

These applications demonstrate the effectiveness of embedded credentials in enhancing model security, providing a robust layer of protection that ensures the integrity and confidentiality of AI models. By employing advanced techniques such as deep watermarking and deep fingerprinting, researchers can enhance the security and resilience of deep learning systems, effectively safeguarding the intellectual property of trained models against theft, unauthorized reproduction, and misuse in an increasingly adversarial environment. This approach addresses significant concerns regarding the protection of valuable assets in the form of deep neural networks, which require substantial resources and expertise to develop [14, 16].

5 Defense Against Adversarial Attacks

5.1 Adaptive and Robust Defense Strategies

Adaptive and robust defense strategies are pivotal for bolstering deep learning models against adversarial threats, ensuring their integrity and functionality in challenging environments. These strategies employ dynamic and proactive measures to counteract evolving attacks while maintaining optimal performance. Structural watermarking enhances robustness by embedding watermarks within the network's architecture rather than its parameters, thereby providing resilience against parameter-based attacks [17]. The MAT technique complicates the transfer of predictions to surrogate models, safeguarding intellectual property and emphasizing the importance of obstructing attack vectors to protect model functionality and ownership [9].

Deep watermarking frameworks exhibit resilience against various attacks, achieving high success rates in watermark extraction, crucial for maintaining model integrity [1]. The ROSE method maintains high watermark recovery rates with minimal performance impact, highlighting the necessity of balancing security measures with model efficiency [6]. Robust watermarking schemes that withstand adaptive attacks are essential for ensuring model security and intellectual property protection [5].

Incorporating black-box detection capabilities, as demonstrated by Guo et al., enhances security frameworks with minimal overhead for authorship proof and resistance to various attacks [18]. Understanding the robustness of detection methods is crucial for developing effective countermeasures against the misuse of generative AI technologies, as emphasized by Saberi et al. [80].

6 Case Studies and Applications

The exploration of watermarking and model security through case studies highlights their crucial role in safeguarding intellectual property and enhancing data security across diverse domains. This section delves into specific instances that demonstrate the effectiveness and adaptability of watermarking methods in overcoming contemporary data protection challenges.

6.1 Case Studies and Real-world Applications

Watermarking and model security techniques are vital in real-world applications for protecting intellectual property and ensuring data security. The anti-neuron watermarking method exemplifies this by preventing unauthorized use of user data in neural network training, thus safeguarding data privacy [81]. In the IoT domain, automated robust image watermarking techniques enhance security by embedding user credentials into QR codes, limiting access to authorized users and mitigating significant risks of unauthorized access [82]. These cases underscore the versatility of watermarking strategies in securing intellectual property and ensuring data integrity. Advanced watermarking methods enhance security frameworks against unauthorized access, particularly in large language models (LLMs) and AI-generated content, by embedding latent fingerprints for data ownership verification and mitigating intellectual property theft. They address traditional vulnerabilities, such as removal and forgery, and are crucial as digital landscapes evolve, ensuring digital asset integrity and regulatory compliance [12, 7, 59, 83].

6.2 Industry-Specific Implementations

Intellectual property protection strategies in deep learning are tailored across industries to meet specific security and operational needs. In healthcare, watermarking techniques protect sensitive patient data processed by AI models, fostering trust in AI-driven solutions [81]. The automotive industry employs robust watermarking frameworks to secure autonomous vehicle models from unauthorized access and tampering, ensuring the reliability and safety of navigation and safety systems [82]. Financial services use watermarking and encryption to protect AI models in fraud detection and risk assessment, verifying ownership and preventing algorithm misappropriation, which is critical for maintaining customer trust and regulatory compliance. These methods safeguard intellectual property and ensure training data integrity, mitigating unauthorized usage risks and enhancing compliance with copyright regulations [12, 84, 29]. In entertainment, watermarking protects AI-generated content, such as music and videos, by embedding hidden signals to distinguish these works from human creations. As generative AI evolves, effective watermarking becomes essential to counter misinformation and intellectual property theft, enhancing trustworthiness and safety in AI outputs and contributing to regulatory discussions on generative AI's impact in creative fields [41, 85]. By embedding digital signatures, content creators can assert ownership and track unauthorized distribution, protecting against piracy and unauthorized reproduction. These industry-specific implementations highlight the adaptability and necessity of intellectual property protection strategies in deep learning. Customizing watermarking techniques to address sector-specific challenges significantly enhances AI model security, protecting intellectual property and preserving operational integrity in a digital and interconnected landscape. Recent advancements in watermarking, such as knowledge injection, have proven highly effective in protecting AI-generated content, ensuring regulatory compliance and fostering trust in AI technologies [12, 59, 28, 16, 41].

7 Challenges and Future Directions

The dynamic evolution of deep learning necessitates a reevaluation of watermarking techniques to address vulnerabilities, particularly regarding forgery. This section focuses on the challenges of watermark forgery, emphasizing adversarial attacks and the limitations of current methodologies. Addressing these challenges is vital for developing innovative solutions that enhance the security and efficacy of watermarking in intellectual property protection.

7.1 Challenges in Watermark Forgery

The sophistication of adversarial techniques presents significant challenges for watermarking in deep learning models. Current methods are vulnerable to removal attacks that erase embedded watermarks without affecting model performance [11]. Even robust techniques like ROSE are susceptible to extreme pruning attacks, which can eliminate watermarks while degrading model accuracy [6]. Watermarking methods for machine-generated texts also face vulnerabilities, necessitating comprehensive evaluations and enhancements to withstand adaptive removal strategies [12, 5, 15, 4, 86]. The ineffectiveness of current methods when surrogate models deviate from target architectures further complicates protection efforts.

Watermarked images are at risk from image processing algorithms and intentional attacks that can corrupt hidden data, compromising integrity and security. Verifying authorship amid significantly altered content is challenging, particularly with generative models facilitating unauthorized reuse, posing substantial risks to intellectual property rights. Traditional techniques struggle against sophisticated removal attacks, complicating authorship attribution. Innovative approaches leveraging latent fingerprints through regeneration are essential for enhancing ownership verification integrity and addressing concerns related to misinformation and copyright infringement [7, 87, 15]. The vulnerability of watermarks to overwriting attacks, where a third party embeds a different watermark, remains a critical concern.

To effectively tackle these challenges, a multifaceted approach is essential. This involves strengthening techniques against attacks—such as watermark removal and forgery—while expanding applicability across various domains and datasets. Innovations like Multi-bit Watermark via Position Allocation enhance robustness and allow for embedding longer messages without degrading text quality, alongside a comprehensive understanding of deep learning-based image watermarking that categorizes methodologies and highlights emerging research avenues [2, 88, 59]. Developing strategies capable of counteracting diverse attack vectors while maintaining watermark integrity will advance secure and reliable intellectual property protection in deep learning models.

7.2 Enhancing Watermark Robustness and Adaptability

Enhancing the robustness and adaptability of watermarking techniques is crucial for securing deep learning models in adversarial environments. The Ingrain method demonstrates improved resilience against distillation attacks, underscoring the need for robust strategies against sophisticated threats [89]. Future research should focus on fortifying watermarking diffusion models (WDM) against fine-tuning attacks, which significantly challenge embedded watermark integrity. Exploring methods for embedding watermarks in other generative models can enhance adaptability, ensuring applicability across diverse model architectures and use cases [90].

Developing robust and adaptable methods requires integrating principles from adversarial machine learning, cryptography, and digital watermarking. This approach should prioritize exploring pretext and post-text strategies, resilience against adversarial attacks, and content quality preservation during the watermarking process [59, 42, 15, 32, 26]. Leveraging interdisciplinary perspectives will foster innovative solutions enhancing security and resilience, ensuring effective intellectual property protection in an increasingly adversarial landscape.

7.3 Expanding Applicability Across Diverse Domains

Expanding intellectual property protection strategies across diverse domains is imperative to address unique challenges and requirements. In healthcare, watermarking is vital for safeguarding sensitive patient data processed by AI models, ensuring privacy and regulatory compliance [81]. This underscores the necessity for robust methods adaptable to stringent security demands.

In the automotive industry, protecting autonomous vehicle models from unauthorized access and tampering is paramount. Watermarking frameworks tailored to this sector ensure proprietary algorithms in navigation and safety systems remain secure, maintaining reliability and safety [82]. Embedding watermarks into vehicle software is essential for detecting unauthorized modifications and preserving operational integrity.

The financial services industry benefits from watermarking and encryption techniques to protect AI models used in fraud detection and risk assessment. Embedding watermarks within these models allows financial institutions to verify ownership and prevent misappropriation by competitors or malicious actors. Effective watermarking techniques for LLMs are crucial for intellectual property protection and enhancing text source detection reliability, mitigating risks associated with unauthorized data use and potential misuse of generated content. Innovations like DeepTextMark and TextMarker safeguard copyrighted material and verify AI-generated texts' authenticity [12, 84, 29].

In the entertainment industry, watermarking is pivotal for protecting intellectual property associated with AI-generated content, such as music and video. Embedding digital signatures within creative works allows content creators to assert ownership and monitor unauthorized distribution, leveraging innovative techniques that integrate proactive and passive strategies. This approach safeguards

assets from piracy and unauthorized reproduction, enhancing traceability in the face of evolving generative technologies, ensuring robust defense against intellectual property violations, and fostering accountability in the digital landscape [7, 87, 91, 92].

The varied applications of watermarking in AI-generated content regulation, large language model protection, and content verification highlight the necessity for adaptable and robust protection strategies tailored to each domain's unique challenges, particularly in addressing vulnerabilities to attacks like removal and forging [27, 67, 15, 59]. By developing tailored solutions for different industries, researchers can enhance AI models' security and integrity, ensuring effective intellectual property protection in an increasingly interconnected digital world.

7.4 Improving Evaluation Practices and Methodologies

Evaluating watermarking techniques in deep learning models is crucial for assessing effectiveness, robustness, and practical applicability. Current practices often focus on metrics like Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Bit Error Rate (BER) to quantify watermarking's impact on model outputs, ensuring fidelity preservation [19]. However, these metrics may not fully capture resilience against diverse adversarial threats, necessitating more comprehensive evaluation frameworks.

Systematic assessment against a wide range of attack scenarios, including pruning, weight quantization, and JPEG compression, is essential for ensuring robustness in adversarial environments [6]. Incorporating benchmarks evaluating the latent space of foundation models can provide insights into vulnerabilities and guide the development of more resilient methods [4].

Introducing novel benchmarks, such as the DLOV E benchmark, represents a significant advancement in evaluating DNN-based watermarking techniques' robustness against adversarial attacks. These benchmarks facilitate developing more resilient defense strategies by providing a framework for comparing models and techniques [38].

Exploring cross-disciplinary approaches that integrate insights from adversarial machine learning, cryptography, and digital watermarking can enhance evaluation practices by providing a holistic view of watermarking techniques' security and effectiveness. By integrating insights from various disciplines, researchers can create advanced evaluation methodologies that effectively tackle intellectual property protection challenges for deep learning models, given the significant resources required for training and the rising risks of unauthorized reproduction, redistribution, and abuse. This includes exploring strategies like deep watermarking and fingerprinting and developing comprehensive frameworks like DEEPJUDGE to assess model integrity against potential attacks, ensuring robust protection of these valuable intellectual assets [13, 93, 94, 16, 95].

7.5 Innovative Approaches and Emerging Trends

Innovative approaches and emerging trends in intellectual property protection for AI models are crucial for enhancing watermarking techniques' robustness and adaptability. Future research directions emphasize developing more resilient schemes capable of withstanding combined attack strategies and a wider array of threats, enhancing evaluation frameworks [5]. This involves improving resilience against preprocessing attacks and extending application beyond image processing models to tasks like object detection and speech recognition [79].

Exploring generative model fingerprinting methods across applications, such as Natural Language Generation and Image Generation, represents a significant trend [7]. These methods provide robust ownership verification and enhance AI-generated content security. Integrating content-aware watermark embedding techniques and innovative security protocols can further bolster model robustness against untrained noise and adversarial threats [2].

Future work should focus on exploring a broader range of foundation models and comparing vulnerabilities to classical techniques. This innovative approach enables a deeper understanding of different models' strengths and weaknesses [4]. Enhancing watermarking techniques' robustness and investigating additional strategies for detection and removal in more complex neural network architectures are essential areas for exploration [11].

Developing adaptive control mechanisms for embedding strength in processes could optimize resource allocation and expand applicability across various domains beyond machine learning [96]. Addressing these innovative approaches and emerging trends will develop more secure and robust protection strategies, ensuring AI models' integrity and ownership in an increasingly adversarial landscape. These advancements will play a crucial role in fortifying model security and adapting to evolving intellectual property protection challenges in AI.

8 Conclusion

The intellectual property protection landscape in deep learning is marked by both innovation and ongoing challenges, necessitating comprehensive strategies to safeguard AI models. Recent advancements, such as high-accuracy watermarking frameworks for Graph Neural Networks (GNNs), highlight the efficacy of these methods in verifying model ownership while maintaining task performance. For example, frameworks like Merkle-Sign successfully address ownership verification challenges in federated learning, ensuring the protection of deep neural network (DNN) models and facilitating independent verification.

Emerging methods, including passport-based DNN ownership verification schemes, effectively counter ambiguity attacks, thereby enhancing secure ownership verification. Additionally, key-based protection strategies for image captioning models achieve ownership verification without compromising performance, emphasizing the importance of intellectual property safeguards. Experimental validations further demonstrate the practicality and effectiveness of protecting deep ensemble models (DEMs), offering viable solutions for securing AI assets.

Moreover, the proposed pooled membership inference (PMI) method verifies DNN model ownership without degrading performance, showcasing significant potential for practical applications in intellectual property protection. These innovative techniques highlight the critical need for continued research and development in watermarking for industrial applications, validating the robustness of these methods against attacks and underscoring their wider applicability.

References

[1] Jie Zhang, Dongdong Chen, Jing Liao, Han Fang, Weiming Zhang, Wenbo Zhou, Hao Cui, and Nenghai Yu. Model watermarking for image processing networks, 2020.

[2] Xin Zhong, Arjon Das, Fahad Alrasheedi, and Abdullah Tanvir. A brief yet in-depth survey of deep learning-based image watermarking, 2023.

[3] 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.

[4] Vitaliy Kinakh, Brian Pulfer, Yury Belousov, Pierre Fernandez, Teddy Furon, and Slava Voloshynovskiy. Evaluation of security of ml-based watermarking: Copy and removal attacks, 2024.

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

[6] Kassem Kallas and Teddy Furon. Rose: A robust and secure dnn watermarking, 2022.

[7] Aditya Desu, Xuanli He, Qiongkai Xu, and Wei Lu. Generative models are self-watermarked: Declaring model authentication through re-generation, 2024.

[8] Yuki Nagai, Yusuke Uchida, Shigeyuki Sakazawa, and Shin'ichi Satoh. Digital watermarking for deep neural networks, 2018.

[9] Yuxuan Li, Sarthak Kumar Maharana, and Yunhui Guo. Not just change the labels, learn the features: Watermarking deep neural networks with multi-view data, 2024.

[10] Suyoung Lee, Wonho Song, Suman Jana, Meeyoung Cha, and Sooel Son. Evaluating the robustness of trigger set-based watermarks embedded in deep neural networks, 2023.

[11] William Aiken, Hyoungshick Kim, and Simon Woo. Neural network laundering: Removing black-box backdoor watermarks from deep neural networks, 2020.

[12] Shuai Li, Kejiang Chen, Kunsheng Tang, Jie Zhang, Weiming Zhang, Nenghai Yu, and Kai Zeng. Turning your strength into watermark: Watermarking large language model via knowledge injection, 2024.

[13] Jie Zhang, Dongdong Chen, Jing Liao, Weiming Zhang, Huamin Feng, Gang Hua, and Nenghai Yu. Deep model intellectual property protection via deep watermarking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4005–4020, 2021.

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

[15] Zesen Liu, Tianshuo Cong, Xinlei He, and Qi Li. On evaluating the performance of watermarked machine-generated texts under adversarial attacks, 2024.

[16] 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, 2023.

[17] Xiangyu Zhao, Yinzhe Yao, Hanzhou Wu, and Xinpeng Zhang. Structural watermarking to deep neural networks via network channel pruning, 2021.

[18] 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.

[19] Mahdi Ahmadi, Alireza Norouzi, S. M. Reza Soroushmehr, Nader Karimi, Kayvan Najarian, Shadrokh Samavi, and Ali Emami. Redmark: Framework for residual diffusion watermarking on deep networks, 2018.

[20] Hanzhou Wu. Robust and lossless fingerprinting of deep neural networks via pooled membership inference, 2022.

[21] Yuling Cai, Fan Xiang, Guozhu Meng, Yinzhi Cao, and Kai Chen. Authnet: Neural network with integrated authentication logic, 2024.

[22] Bishwa Karki, Chun-Hua Tsai, Pei-Chi Huang, and Xin Zhong. Deep learning-based text-in-image watermarking, 2024.

[23] 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, 2024.

[24] Chaoning Zhang, Chenguo Lin, Philipp Benz, Kejiang Chen, Weiming Zhang, and In So Kweon. A brief survey on deep learning based data hiding, 2022.

[25] Janvi Thakkar, Giulio Zizzo, and Sergio Maffeis. Elevating defenses: Bridging adversarial training and watermarking for model resilience, 2024.

[26] Salah Ghamizi, Maxime Cordy, Mike Papadakis, and Yves Le Traon. Adversarial embedding: A robust and elusive steganography and watermarking technique, 2019.

[27] Kaiyi Pang, Tao Qi, Chuhan Wu, Minhao Bai, Minghu Jiang, and Yongfeng Huang. Modelshield: Adaptive and robust watermark against model extraction attack, 2025.

[28] Isabell Lederer, Rudolf Mayer, and Andreas Rauber. Identifying appropriate intellectual property protection mechanisms for machine learning models: A systematization of watermarking, fingerprinting, model access, and attacks, 2023.

[29] Travis Munyer, Abdullah Tanvir, Arjon Das, and Xin Zhong. Deeptextmark: A deep learning-driven text watermarking approach for identifying large language model generated text, 2024.

[30] Zhaoxi Zhang, Xiaomei Zhang, Yanjun Zhang, Leo Yu Zhang, Chao Chen, Shengshan Hu, Asif Gill, and Shirui Pan. Large language model watermark stealing with mixed integer programming, 2024.

[31] Ruisi Zhang and Farinaz Koushanfar. Watermarking large language models and the generated content: Opportunities and challenges, 2024.

[32] Sahar Abdelnabi and Mario Fritz. Adversarial watermarking transformer: Towards tracing text provenance with data hiding, 2021.

[33] Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. Embedding watermarks into deep neural networks, 2017.

[34] Huili Chen, Bita Darvish Rouhani, Xinwei Fan, Osman Cihan Kilinc, and Farinaz Koushanfar. Performance comparison of contemporary dnn watermarking techniques, 2018.

[35] Yuexin Xiang, Tiantian Li, Wei Ren, Tianqing Zhu, and Kim-Kwang Raymond Choo. Generating image adversarial examples by embedding digital watermarks, 2022.

[36] Biqing Qi, Junqi Gao, Yiang Luo, Jianxing Liu, Ligang Wu, and Bowen Zhou. Investigating deep watermark security: An adversarial transferability perspective, 2024.

[37] Erwin Quiring, Daniel Arp, and Konrad Rieck. Fraternal twins: Unifying attacks on machine learning and digital watermarking, 2017.

[38] Sudev Kumar Padhi and Sk. Subidh Ali. Dlove: A new security evaluation tool for deep learning based watermarking techniques, 2024.

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

[40] Lixin Fan, KamWoh Ng, and Chee Seng Chan. Digital passport: A novel technological strategy for intellectual property protection of convolutional neural networks, 2019.

[41] Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramer, Somesh Jha, Lei Li, Yu-Xiang Wang, and Dawn Song. Sok: Watermarking for ai-generated content, 2024.

[42] Hongyu Zhu, Sichu Liang, Wentao Hu, Fangqi Li, Ju Jia, and Shilin Wang. Reliable model watermarking: Defending against theft without compromising on evasion, 2024.

[43] Abhishek Chakraborty, Daniel Xing, Yuntao Liu, and Ankur Srivastava. Dynamarks: Defending against deep learning model extraction using dynamic watermarking, 2022.

[44] Sebastian Szyller, Buse Gul Atli, Samuel Marchal, and N. Asokan. Dawn: Dynamic adversarial watermarking of neural networks, 2021.

[45] Huajie Chen, Tianqing Zhu, Chi Liu, Shui Yu, and Wanlei Zhou. High-frequency matters: An overwriting attack and defense for image-processing neural network watermarking, 2023.

[46] Ruowei Wang, Chenguo Lin, Qijun Zhao, and Feiyu Zhu. Watermark faker: Towards forgery of digital image watermarking, 2021.

[47] Xuankai Liu, Fengting Li, Bihan Wen, and Qi Li. Removing backdoor-based watermarks in neural networks with limited data, 2020.

[48] Yifan Lu, Wenxuan Li, Mi Zhang, Xudong Pan, and Min Yang. Neural dehydration: Effective erasure of black-box watermarks from dnns with limited data, 2024.

[49] Xinwei Liu, Jian Liu, Yang Bai, Jindong Gu, Tao Chen, Xiaojun Jia, and Xiaochun Cao. Watermark vaccine: Adversarial attacks to prevent watermark removal, 2022.

[50] Yiming Li, Yang Bai, Yong Jiang, Yong Yang, Shu-Tao Xia, and Bo Li. Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection, 2023.

[51] Tao Xiang, Chunlong Xie, Shangwei Guo, Jiwei Li, and Tianwei Zhang. Protecting your nlg models with semantic and robust watermarks, 2021.

[52] Fangqi Li and Shilin Wang. Knowledge-free black-box watermark and ownership proof for image classification neural networks, 2022.

[53] Buse Gul Atli Tekgul and N. Asokan. On the effectiveness of dataset watermarking in adversarial settings, 2022.

[54] Shangwei Guo, Tianwei Zhang, Han Qiu, Yi Zeng, Tao Xiang, and Yang Liu. Fine-tuning is not enough: A simple yet effective watermark removal attack for dnn models, 2021.

[55] Mohammed Lansari, Reda Bellafqira, Katarzyna Kapusta, Vincent Thouvenot, Olivier Bettan, and Gouenou Coatrieux. When federated learning meets watermarking: A comprehensive overview of techniques for intellectual property protection, 2023.

[56] Buse Gul Atli, Yuxi Xia, Samuel Marchal, and N. Asokan. Waffle: Watermarking in federated learning, 2021.

[57] Bowen Li, Lixin Fan, Hanlin Gu, Jie Li, and Qiang Yang. Fedipr: Ownership verification for federated deep neural network models, 2022.

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

[59] Guanlin Li, Yifei Chen, Jie Zhang, Shangwei Guo, Han Qiu, Guoyin Wang, Jiwei Li, and Tianwei Zhang. Warfare:breaking the watermark protection of ai-generated content, 2025.

[60] Carl De Sousa Trias, Mihai Mitrea, Attilio Fiandrotti, Marco Cagnazzo, Sumanta Chaudhuri, and Enzo Tartaglione. Watermas: Sharpness-aware maximization for neural network watermarking, 2024.

[61] Rui Min, Sen Li, Hongyang Chen, and Minhao Cheng. A watermark-conditioned diffusion model for ip protection, 2024.

[62] Haonan An, Guang Hua, Zhiping Lin, and Yuguang Fang. Box-free model watermarks are prone to black-box removal attacks, 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] Qi Pang, Shengyuan Hu, Wenting Zheng, and Virginia Smith. No free lunch in llm watermarking: Trade-offs in watermarking design choices, 2024.

[65] Niyar R Barman, Krish Sharma, Ashhar Aziz, Shashwat Bajpai, Shwetangshu Biswas, Vasu Sharma, Vinija Jain, Aman Chadha, Amit Sheth, and Amitava Das. The brittleness of ai-generated image watermarking techniques: Examining their robustness against visual paraphrasing attacks, 2024.

[66] Naen Xu, Changjiang Li, Tianyu Du, Minxi Li, Wenjie Luo, Jiacheng Liang, Yuyuan Li, Xuhong Zhang, Meng Han, Jianwei Yin, and Ting Wang. Copyrightmeter: Revisiting copyright protection in text-to-image models, 2024.

[67] Baizhou Huang and Xiaojun Wan. Waterpool: A watermark mitigating trade-offs among imperceptibility, efficacy and robustness, 2024.

[68] Chenchen Gu, Xiang Lisa Li, Percy Liang, and Tatsunori Hashimoto. On the learnability of watermarks for language models, 2024.

[69] Xuandong Zhao, Lei Li, and Yu-Xiang Wang. Distillation-resistant watermarking for model protection in nlp, 2022.

[70] Ruixiang Tang, Mengnan Du, and Xia Hu. Deep serial number: Computational watermarking for dnn intellectual property protection, 2023.

[71] Xin Mu, Yu Wang, Zhengan Huang, Junzuo Lai, Yehong Zhang, Hui Wang, and Yue Yu. Encryip: A practical encryption-based framework for model intellectual property protection, 2023.

[72] Qi Song, Ziyuan Luo, Ka Chun Cheung, Simon See, and Renjie Wan. Protecting nerfs' copyright via plug-and-play watermarking base model, 2024.

[73] Xiangyu Wen, Yu Li, Wei Jiang, and Qiang Xu. On function-coupled watermarks for deep neural networks, 2023.

[74] Franziska Boenisch. A systematic review on model watermarking for neural networks, 2021.

[75] Guanhao Gan, Yiming Li, Dongxian Wu, and Shu-Tao Xia. Towards robust model watermark via reducing parametric vulnerability, 2023.

[76] Yong Liu, Hanzhou Wu, and Xinpeng Zhang. Robust and imperceptible black-box dnn watermarking based on fourier perturbation analysis and frequency sensitivity clustering, 2023.

[77] Maedeh Jamali, Nader Karim, Pejman Khadivi, Shahram Shirani, and Shadrokh Samavi. Robust watermarking using diffusion of logo into autoencoder feature maps, 2021.

[78] Jacob Shams, Ben Nassi, Ikuya Morikawa, Toshiya Shimizu, Asaf Shabtai, and Yuval Elovici. Seeds don't lie: An adaptive watermarking framework for computer vision models, 2022.

[79] Jie Zhang, Dongdong Chen, Jing Liao, Weiming Zhang, Huamin Feng, Gang Hua, and Nenghai Yu. Deep model intellectual property protection via deep watermarking, 2021.

[80] Mehrdad Saberi, Vinu Sankar Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, Wenxiao Wang, and Soheil Feizi. Robustness of ai-image detectors: Fundamental limits and practical attacks, 2024.

[81] Zihang Zou, Boqing Gong, and Liqiang Wang. Anti-neuron watermarking: Protecting personal data against unauthorized neural networks, 2022.

[82] Xin Zhong, Pei-Chi Huang, Spyridon Mastorakis, and Frank Y. Shih. An automated and robust image watermarking scheme based on deep neural networks, 2020.

[83] Michael-Andrei Panaitescu-Liess, Zora Che, Bang An, Yuancheng Xu, Pankayaraj Pathmanathan, Souradip Chakraborty, Sicheng Zhu, Tom Goldstein, and Furong Huang. Can watermarking large language models prevent copyrighted text generation and hide training data?, 2024.

[84] Yixin Liu, Hongsheng Hu, Xun Chen, Xuyun Zhang, and Lichao Sun. Watermarking text data on large language models for dataset copyright, 2024.

[85] Kui Ren, Ziqi Yang, Li Lu, Jian Liu, Yiming Li, Jie Wan, Xiaodi Zhao, Xianheng Feng, and Shuo Shao. Sok: On the role and future of aigc watermarking in the era of gen-ai, 2024.

[86] Jiacheng Liang, Zian Wang, Lauren Hong, Shouling Ji, and Ting Wang. Waterpark: A robustness assessment of language model watermarking, 2024.

[87] Runyi Li, Xuanyu Zhang, Zhipei Xu, Yongbing Zhang, and Jian Zhang. Protect-your-ip: Scalable source-tracing and attribution against personalized generation, 2024.

[88] KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. Advancing beyond identification: Multi-bit watermark for large language models, 2024.

[89] Ziqi Yang, Hung Dang, and Ee-Chien Chang. Effectiveness of distillation attack and countermeasure on neural network watermarking, 2019.

[90] Sen Peng, Yufei Chen, Cong Wang, and Xiaohua Jia. Intellectual property protection of diffusion models via the watermark diffusion process, 2023.

[91] Xuefeng Fan, Dahao Fu, Hangyu Gui, Xinpeng Zhang, and Xiaoyi Zhou. Pcpt and acpt: Copyright protection and traceability scheme for dnn models, 2023.

[92] Jijia Yang, Sen Peng, and Xiaohua Jia. Embedding watermarks in diffusion process for model intellectual property protection, 2024.

[93] 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, 2021.

[94] Mingfu Xue, Yushu Zhang, Jian Wang, and Weiqiang Liu. Intellectual property protection for deep learning models: Taxonomy, methods, attacks, and evaluations, 2021.

[95] Mingfu Xue, Xin Wang, Yinghao Wu, Shifeng Ni, Yushu Zhang, and Weiqiang Liu. Infip: An explainable dnn intellectual property protection method based on intrinsic features, 2022.

[96] Mahnoosh Bagheri, Majid Mohrekesh, Nader Karimi, and Shadrokh Samavi. Adaptive control of embedding strength in image watermarking using neural networks, 2020.

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.