applied sciences

MDPI

Article

A Brief, In-Depth Survey of Deep Learning-Based Image Watermarking

Xin Zhong *, Arjon Das, Fahad Alrasheedi and Abdullah Tanvir

Department of Computer Science, University of Nebraska Omaha, Omaha, NE 68182, USA; arjondas@unomaha.edu (A.D.); falrasheedi@unomaha.edu (F.A.); atanvir@unomaha.edu (A.T.)
* Correspondence: xzhong@unomaha.edu

Abstract: This paper presents a comprehensive survey of deep learning-based image watermarking; this technique entails the invisible embedding and extraction of watermarks within a cover image, aiming for a seamless combination of robustness and adaptability. We navigate the complex landscape of this interdisciplinary domain, linking historical foundations, current innovations, and prospective developments. Unlike existing literature, our study concentrates exclusively on image watermarking with deep learning, delivering an in-depth, yet brief analysis enriched by three fundamental contributions. First, we introduce a refined categorization, segmenting the field into embedder–extractor, deep networks for feature transformation, and hybrid methods. This taxonomy, inspired by the varied roles of deep learning across studies, is designed to infuse clarity, offering readers technical insights and directional guidance. Second, our exploration dives into representative methodologies, encapsulating the diverse research directions and inherent challenges within each category to provide a consolidated perspective. Lastly, we venture beyond established boundaries, outlining emerging frontiers and providing detailed insights into prospective research avenues.

Keywords: survey; deep learning; image watermarking

Citation: Zhong, X.; Das, A.; Alrasheedi, F.; Tanvir, A. A Brief, In-Depth Survey of Deep Learning-Based Image Watermarking. *Appl. Sci.* **2023**, *13*, 11852. https://doi.org/10.3390/app132111852

Academic Editor: Cosimo Nardi

Received: 1 October 2023
Revised: 25 October 2023
Accepted: 28 October 2023
Published: 30 October 2023

1. Introduction

Each year, the internet serves as a conduit for the upload, transfer, and sharing of billions of digital images [1]. The advent of sophisticated digital technologies has facilitated the effortless editing, dissemination, and reproduction of images, precipitating a surge in unauthorized usage and concomitant infringement of the intellectual property rights of original creators. In this context, digital image protection emerges as a critical mechanism to enforce the sanctity of the intellectual property of content creators. Digital images are not merely visual content; they are significant assets for individuals, corporate entities, and various organizations. The integrity of these assets is often threatened by unauthorized utilization and replication, a scenario that could culminate in substantial financial deficits and damage to reputation. Moreover, the illicit use of personal images can inflict emotional distress, exacerbated when images or videos are circulated without the owner's consent.

Digital image watermarking has emerged as a preeminent technique in the domain of image protection, garnering widespread application and acclaim [2]. Central to this technique is the covert embedding of informational elements, ranging from logos to copyright notices, directly into the visual content. This surreptitious integration ensures that only individuals endowed with the appropriate authorization can extract the watermark, maintaining the confidentiality and integrity of the embedded data [3]. Watermarks serve a multifaceted purpose. They act as an indelible signature, affirming ownership and bolstering copyright protection by dissuading unauthorized replication and distribution. They stand as testaments to authenticity, facilitating the licensing and tracking of image utilization. Moreover, they serve as conduits for covert communication, encapsulating hidden messages that are seamlessly woven into the visual content. Lastly, they function as

intrinsic detectors, revealing alterations and tampering, thereby upholding the integrity of the original content. The versatility of digital watermarking extends its applicability across a plethora of fields. It has become an instrumental tool in forensic analyses, enhancing the traceability and verification of digital content. In the burgeoning spheres of 5G communication and the Internet of Things (IoT) [4], watermarking is pivotal in fortifying security and enhancing data integrity. Smart cities, characterized by their intricate networks of interconnected digital systems, leverage watermarking to safeguard data and ensure the seamless, secure interchange of information [5].

Image watermarking and image steganography [6] are closely related fields, yet with distinct technical and application-specific differences. Both areas explore the intricate process of subtly embedding data within images, ensuring the modifications remain unnoticeable to the unaided eye. However, the differing goals they aim to accomplish lead to distinct technical focuses. Image steganography primarily aims to provide a covert channel for information transmission, avoiding detection by unauthorized parties. It hinges on the principles of unpredictability and high payload capacity [7,8]. The former emphasizes resistance against steganalysis techniques, while the latter denotes the ability to embed a significant amount of data without affecting the perceptual quality of the cover image. On the other hand, image watermarking seeks to protect the integrity of both the cover image and the embedded watermark. The fundamental aspect of this field is robustness, which refers to the enduring readability of the watermark amidst various potential attacks. This characteristic is crucial, although there are situations where fragile watermarking is preferable and necessary, especially in scenarios like medical imaging, where maintaining the original quality of the image is vital [9].

Traditional image watermarking approaches are predominantly characterized by handcrafted embedding and extraction mechanisms. These processes often entail the intricate utilization of prior knowledge and a substantial level of expertise in the domain of image processing. The inherent dependency on prior information culminates in designs that are tailored for specific cases and exhibit a marked lack of adaptability [6]. Such designs are characterized by a uniform application of watermarking patterns across a myriad of images, neglecting the unique content characteristics and quality attributes inherent to each individual image. In the realm of robustness, a significant limitation manifests. Each handcrafted method accentuates a specific attribute or set of attributes, resulting in a fragmented and isolated approach to enhancing watermark robustness. The absence of a comprehensive strategy that encapsulates a broad spectrum of potential attacks is conspicuously evident. For instance, the watermarking technique premised on quantization index modulation is primarily tailored to counter JPEG compression artifacts [10]. Concurrently, methods founded upon the principles of log-polar coordinates are intricately designed to mitigate the impacts of rotational manipulations [11]. This scenario illuminates an overarching challenge—the lack of a holistic, adaptive, and universally applicable watermarking strategy. The juxtaposition of these isolated techniques against the dynamic and multifaceted nature of digital media manipulation threats underscores a significant vulnerability. The evolution of manipulative attacks, characterized by their increasing sophistication, demands a parallel evolution in watermarking techniques that is anchored in adaptability, comprehensive threat mitigation, and contextual applicability.

Deep learning [12] is characterized by algorithms inspired by the structure and function of the brain, known as artificial neural networks. These networks are adept at learning from large volumes of data, enabling the extraction of complex patterns and representations. Deep learning has catalyzed significant advancements in various fields, including image and speech recognition, natural language processing, and autonomous systems. The depth of the networks, characterized by multiple layers of interconnected nodes, contributes to their capacity to perform intricate computations, offering superior performance and predictive accuracy in diverse applications. Each layer transforms its input data into increasingly abstract and complex representations, enabling nuanced decision-making and predictions.

In the quest for enhanced robustness and adaptability in image watermarking, deep learning emerges as a formidable ally. Unlike their traditional counterparts, deep learning-based watermarking algorithms harbor the potential to learn and adapt [7]. They encapsulate the capacity to intuitively morph in response to the unique attributes of each image and the evolving landscape of threats. This adaptability heralds a new epoch in watermarking—one characterized by enhanced robustness, imperceptibility, and the nuanced balancing of these cardinal attributes. As we traverse this trajectory, the integration of deep learning in watermarking is not just an incremental enhancement but a paradigmatic shift. It propels watermarking from a static, isolated, and case-specific discipline into a dynamic, adaptive, and holistic domain. This evolution is not only pivotal for the enhanced protection of digital media assets but also instrumental in the nuanced balancing of imperceptibility and robustness, ensuring that the integrity and aesthetic value of digital media are meticulously preserved.

The necessity of a survey focusing on deep learning-based image watermarking emanates from the rapid advancements and complexities ingrained in this burgeoning field. As the integration of deep learning in image watermarking has emerged as a pivotal focus, there is a requisite for a comprehensive, synthesized, and analytical review of the existing literature and methodologies. To this end, this paper presents a comprehensive survey of cutting-edge deep learning-based image watermarking techniques, serving as a reference for the state-of-the-art in deep learning-based image watermarking, summarizing key research directions and envisioning future studies in the domain.

Objectives and Distinctiveness of This Survey

We illustrate the objectives and distinctiveness of our survey by summarizing an overview of the topic concentrations of existing related survey papers in Table 1. Current surveys predominantly orient toward deep learning model architectures, diversified artificial intelligence methodologies, data hiding techniques, and prominent proposals. It should be noted that our review is intricately tailored to encapsulate the synopsis of works germane to deep learning-based image watermarking. Consequently, extended domains, including the watermarking of the deep learning models themselves [13], fall beyond the scope of our discussion.

Table 1. Summary of existing related surveys.

Methods	Concentration
Gupta and Kishore [14]	Summarizing various convolutional neural network model architectures used in deep learning-based image watermarking
Amrit and Singh [5]	Summarizing watermarking using artificial intelligence, machine learning, and deep learning
Zhang et al. [15]	Reviewing deep learning-based data hiding, classifying based on capacity, security, and robustness, and outlining three commonly used architectures
Byrnes et al. [16]	Surveying deep learning techniques for data hiding in watermarking and steganography, and categorizing them based on model architectures and noise injection methods
Singh et al. [17]	Reviewing the popular deep-learning model-based digital watermarking methods and summarizing/comparing contributions in the literature

In contrast to existing work, our survey focuses on deep learning-based image watermarking and provides a brief yet in-depth analysis, distinguished by three primary advantages. (1) We systematically categorize deep learning-based image watermarking into embedder–extractor, deep networks for feature transformation, and hybrid methods. This categorization, grounded in the distinct roles deep learning assumes in various studies,

aspires to offer technical insights and guidance. (2) We study representative methodologies and encapsulate the directions and challenges of research within each specified category, offering a coherent synopsis. (3) We extend our discussion to encompass a detailed exploration of prospective research avenues, delineating emerging frontiers in the domain of deep learning-based image watermarking.

Through the systematic analysis, critical research direction discussion, and prospective outlooks, one primary objective of our survey is to connect past research, present innovations, and future prospects, potentially propelling the field toward refined methodologies, enhanced effectiveness, and broader applicative horizons. The rest of this paper is structured as follows: Section 2 talks about the relevant preliminaries in conventional image watermarking, Section 3 categorizes the techniques and provides a survey of image watermarking based on deep learning, Section 4 explores potential research avenues for the future, and Section 5 presents our conclusion.

2. Preliminaries

2.1. Traditional Image Watermarking Components

Image watermarking entails the incorporation of watermark data within an image. This watermark, an encoded digital signal, is meticulously crafted to be inconspicuous to human vision yet readily identifiable and extractable via computational algorithms. As elaborated in Section 1, the application spectrum of image watermarking, delineated by the nature of the watermark information, spans copyright protection, authenticity verification, covert communication, and tampering detection, among others. Figure 1 succinctly encapsulates the components and steps inherent in the traditional paradigm of image watermarking.

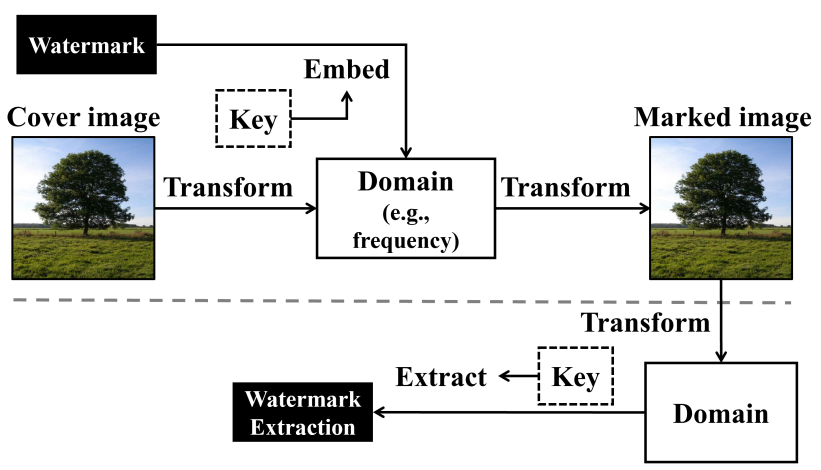

Figure 1. General components and steps of traditional image watermarking.

Embedding and extraction. In the embedding step, the watermark is integrated into the cover image via a watermarking algorithm. The overarching objective is to ensure the embedded watermark is robust and resistant to removal or alteration, whilst concurrently maintaining the visual integrity of the cover image. Various techniques exist for the watermark embedding, such as the modification of pixel values in the spatial domain [6], and the manipulation of coefficients within the frequency domain representation [18]. Post-embedding, the watermarked image is disseminated to the designated audience, potentially via online platforms. Authorized recipients are equipped to extract the embedded watermark utilizing a specialized extraction algorithm.

Key. Numerous image watermarking methodologies incorporate a key, denoting secret values instrumental in modulating the embedding and extraction processes of the watermark. Typically, this key is conjointly generated and disseminated among the content owner and authorized users. Its application within the watermarking procedure varies, contingent on the algorithmic design, predominantly aiming to augment security and

robustness. For instance, the key can be employed to govern the generation of a pseudo-random sequence integrated into the image for watermark embedding [18], or to designate the precise location of the embedded watermark within the image [6].

Watermark preprocessing. Watermark preprocessing can be instrumental in augmenting both security and robustness. One classic approach involves the encryption of the watermark using cryptographic techniques, such as AES [19] or RSA [20]. In this process, an encryption key is employed to convert the watermark into a ciphertext, enhancing the security of the embedded data. The decryption and, hence, the accessibility of the watermark, is contingent upon the application of the corresponding cryptographic key, ensuring the watermark remains impervious to unauthorized access and enhancing its applicability in high-security contexts. In addition to security, the watermark can be encoded using methodologies, such as the error correction code [21], facilitating the rectification of errors within the extracted watermark, and enhancing its robustness. The Reed–Solomon code [22] and convolutional code [23] are classic exemplary methodologies that infuse redundancy into the watermark. This inclusion of redundant data is strategically orchestrated to ameliorate errors encountered during watermark extraction, bolstering the accuracy and reliability of the extraction process, even amidst distortions.

2.2. Typical Metrics and Factors

Although there can be a large number of evaluation metrics and factors considered in image watermarking based on different applications, certain metrics are ubiquitously employed across both traditional and deep learning-based watermarking in the extant literature. In this context, we briefly discuss the typical factors of imperceptibility and robustness, which are integral to assessing the efficacy of watermarking techniques.

Imperceptibility. The ability of the watermark to be embedded into the image data in a way that is invisible to human vision is referred to as imperceptibility. The imperceptibility helps ensure that the watermark does not interfere with the quality of the image. One most frequently applied evaluation metric is the peak signal-to-noise ratio ($PSNR$):

$$PSNR = 10 \times log_{10}\left(\frac{max(c)^2}{\frac{1}{\mathcal{RC}}\sum_{i=1}^{\mathcal{R}}\sum_{j=1}^{\mathcal{C}}(c_{ij} - m_{ij})^2}\right), \tag{1}$$

where $max(c)$ is the largest possible pixel value for the cover image c, which is 255 if we use 8 bits for each grayscale value, and $\mathcal{R}$ and $\mathcal{C}$ denote the height and width of images c and m.

Notably, apart from the extensively utilized $PSNR$, the structural similarity index measure ($SSIM$) [24] is also commonly employed to assess imperceptibility, incorporating evaluations of luminance, contrast, and structural disparities. An essential augmentation to visual imperceptibility is security [25]. This entails ensuring that the embedded watermark is not only invisible to the human eye but also resistant to detection through computational analysis, a criterion of paramount importance in secure watermarking applications, exemplified in domains like smart city planning and digital forensics.

Robustness. Robustness characterizes the watermark's resilience, denoting its capacity to be reliably extracted amidst attacks on the watermarked image, such as compression, filtering, or cropping. The assessment of robustness involves calculating the disparity between the extracted and original watermark post-attack. When the watermark is binary, the bit error rate (BER) serves as a prevalent metric, computed by dividing the number of erroneous bits by the total number of bits embedded. In instances where the watermark takes the form of a two-dimensional matrix, its resilience is often assessed via the normalized cross-correlation (NC), measuring the similarity between the original watermark w and the extracted watermark w':

$$NC = \frac{\sum_{i=1}^{\mathcal{H}}\sum_{j=1}^{\mathcal{L}}(w_{ij} \cdot w'_{ij})}{\sqrt{\sum_{i=1}^{\mathcal{H}}\sum_{j=1}^{\mathcal{L}}(w_{ij})^2}\sqrt{\sum_{i=1}^{\mathcal{H}}\sum_{j=1}^{\mathcal{L}}(w'_{ij})^2}}, \tag{2}$$

where $\mathcal{H}$ and $\mathcal{L}$ define the height and width of the watermark.

Blindness. Blind and semi-blind watermarking represent two prominent approaches in the domain of image watermarking. Blind watermarking is characterized by its ability to detect and extract watermarks without the necessity of referring to the original cover image nor the original watermark (while a key may be required). On the other hand, semi-blind watermarking, while also not requiring the original cover image for watermark extraction, requires the original watermark (and the key) for extraction.

Capacity. Capacity denotes the upper limit of watermark data that can be embedded into an image, giving rise to two primary types of image watermarking: (1) Zero-bit watermarking, which focuses on detecting the presence of a watermark in an image rather than extracting data, and (2) multi-bit watermarking, which involves embedding and extracting a watermark comprised of multiple bits. Zero-bit watermarking serves as a signature for authenticating image data, without the provision to incorporate additional embedded data. This form of watermarking is effective for verification purposes, ensuring the integrity and authenticity of the image content. On the other hand, multi-bit watermarking allows for the embedding of additional information within the image, facilitating a broader spectrum of applications including copyright protection, content annotation, and data tracking. However, the embedding of multiple bits may potentially impact the perceptual quality of the image, necessitating a careful balance between data capacity and image fidelity.

A distinct concept is zero watermarking, where no watermark is directly embedded [26,27]. Instead, a relationship is established and stored between the original content and the watermark (a master share). In case of disputes or verification, this relationship is used to demonstrate the presence of the watermark. This technique proffers advantages, such as eliminating the need to alter the image, thus preserving its original quality. Nonetheless, limitations exist, primarily due to its dependency on the uniqueness of image data, rendering it less effective for images with homogeneous content.

3. Comprehensive Survey of Deep Learning-Based Image Watermarking

This section discusses the integration of deep neural networks in contemporary deep learning-based image watermarking, explaining the adaptation of traditional watermarking processes within this advanced framework. The modern deep learning-based image watermarking approaches are architectured akin to their traditional counterparts, encompassing transformation, watermark embedding, and extraction phases. For a structured analysis, we classify current deep learning-based techniques into three distinct categories based on the roles of deep learning in various papers: (1) embedder–extractor joint training, (2) deep networks for feature transformation, and (3) hybrid methods. In order to provide a clear and concise overview of each category, Table 2 acts as a navigational guide. Each category is outlined, with its key features, requirements, and potential drawbacks listed.

Table 2. Overview of methods in each proposed category.

Methodology	Key Features/Requirements/Potential Drawbacks
Embedder–extractor Joint training	**Key features:** Automated watermarking schemes are learned from designated data through joint optimization of watermark embedding and extraction. **Requirements:** A robust dataset for training and an astute selection of noise levels in the noise layer to ensure robustness. **Potential drawbacks:** Efficacy may wane with inadequate training data or improper noise selection, emphasizing the necessity for a robust dataset and prudent noise level selection.

Table 2. *Cont.*

Methodology	Key Features/Requirements/Potential Drawbacks
Deep network as a feature transformation	**Key features:** Employment of deep networks for feature transformation, leveraging pre-trained models for adept feature extraction. **Requirements:** Pre-training of deep networks on tasks related to robustness within the domain, alongside a separate design of embedding and extraction in this feature space. **Potential drawbacks:** Robustness may be compromised if the feature transformation efficacy is subpar, and pre-trained networks might not adequately prioritize robustness.
Hybrid Methods	**Key features:** Fusion of classical watermarking with deep learning, harnessing strengths from both realms for enhanced watermarking. **Requirements:** Rigorous design and fine-tuning of both traditional watermarking schemes and deep learning models to ensure harmonious operation. **Potential drawbacks:** Increased design complexity and potential amplification of limitations inherited from both classical and deep learning-based techniques, necessitating a sagacious design strategy.

3.1. Embedder–Extractor Joint Training Methods

As depicted in Figure 2, methods within this category fundamentally involve the training of two core components: an embedder network, responsible for watermark integration into a cover image, and an extractor network, tasked with retrieving the embedded watermark from the marked image. Variations in design are present, with some iterations incorporating separate feature extraction networks within the embedder for the preliminary processing of the watermark and cover image. To enhance robustness, a noise module is typically positioned subsequent to the marked image. This module is instrumental in introducing and amalgamating noise into the marked image during training, equipping the extractor network with an augmented capacity to counteract disturbances.

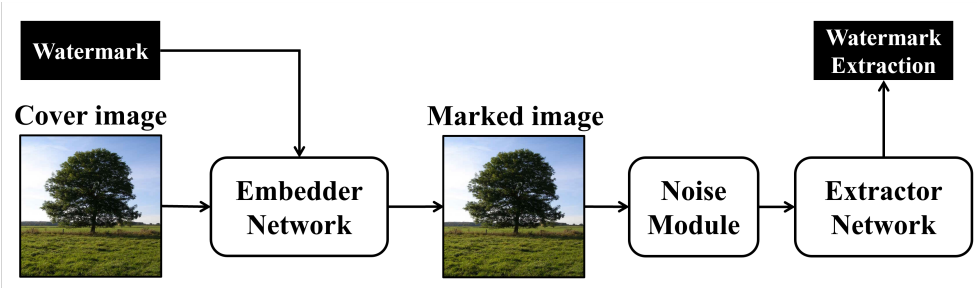

Figure 2. General process of the embedder–extraction joint training.

Typically, all components are jointly trained within a unified deep neural network framework, employing gradient descent as the optimization technique. The objective is to optimize a loss function, aiming to uphold imperceptibility while ensuring the effective extraction of the embedded watermark. Mathematically, a representative loss function for such joint training can be defined as follows:

$$l = f_1(c, m) + f_2(w, w'), \tag{3}$$

where c, m, w, w' represent the cover image, marked image, watermark, and extracted watermark, respectively. The function f_1 quantifies the visual disparity between c and m, aligning with the traditional criterion of achieving a high $PSNR$. Concurrently, f_2 assesses

the variance between w and w', fulfilling the conventional benchmarks of elevated BER or NC.

The elements within l present a trade-off: minimizing $f_1(c, m)$ to zero implies an exact resemblance between the marked and cover images. However, this may constrain the space available for watermark embedding, potentially leading to extraction inefficacy. To orchestrate a network capable of harmonizing this trade-off, a prevalent approach involves utilizing the gradients engendered by $f_2(w, w')$ to refine the weights of all components, effectuating back-propagation extending to both c and m. Conversely, gradients emanating from $f_1(c, m)$ are exclusively employed to optimize the embedder network's weights. Subsequent sections will unfold the intricate designs and distinct characteristics embodied by representative state-of-the-art methods within this joint training category.

The concept of joint training was first introduced in the HiDDeN (hiding data with deep networks) paper by Zhu et al. [7]. The authors aimed to integrate the embedder–extractor paradigm to streamline and unify the processes of image steganography and watermarking. Several innovative and practical designs are incorporated in HiDDeN. The embedder replicates the watermark, and its integration with the cover image occurs at the embedder's final layer to maintain the visual quality. Discriminator networks are deployed to determine the presence of watermarks in images, thereby ensuring that the embedder generates visually coherent marked images. Robustness is bolstered in HiDDeN through the introduction of a noise layer that incorporates noise-inducing operations, including dropout, Gaussian blur, and JPEG compression during the training phase. To navigate the challenges posed by the non-differentiability of the original JPEG incorporated within the noise layer, HiDDeN introduces a differentiable JPEG variant, ensuring the continuity of the gradient flow. Beyond HiDDeN, Zhang et al. [28] introduced universal deep hiding (UDH), another seminal work rooted in the joint embedder–extractor paradigm, applicable to both image watermarking and steganography. UDH is distinguished as one of the pioneering works to embed entire images as watermarks within embedder–decoder frameworks. It introduces an approach to watermark encoding that facilitates disentanglement during the extraction process. In this methodology, the encoded watermark is generated independently of the cover image and subsequently integrated, ensuring a systematic and efficient extraction process while preserving the integrity of the cover image.

A potential challenge associated with joint training for robust image watermarking stems from the necessity of the noise layer being differentiable to facilitate gradient flow. In addressing this issue, Liu et al. [29] proposed a two-stage training methodology. In the initial phase, both the embedder and extractor are collaboratively trained without any noise intervention, ensuring a seamless and undisturbed gradient flow. In the second stage, the embedder's parameters are fixed, rendering it non-adaptable to further training iterations. The focus is then channeled exclusively toward the training of the extractor. This bifurcated training approach allows the integration of non-differentiable noise layers into the extractor's training without compromising the effectiveness of the whole training process. This paper has tested its robustness against a spectrum of prevalent attacks, such as resizing, salt and pepper noise, dropout, crop-out, Gaussian blur, and JPEG compression.

Numerous endeavors aim to circumvent the challenge posed by the non-differentiable nature of the JPEG operation within the noise layer. Chen et al. [30] proposed the employment of simulation networks to emulate JPEG lossy compression, accommodating various quality factors. The model employs the max-pooling layer, convolution layer, and a noise mask to, respectively, represent the sampling, DCT, and quantization processes inherent in JPEG compression. In a related vein, Jia et al. [31] advocated for the incorporation of batches that amalgamate both actual and simulated JPEG compressions. Within the training's noise layer, each batch is configured to randomly incorporate either an actual JPEG compression layer, a differentiable simulation of a JPEG layer, or a layer without noise. In scenarios employing momentum-based optimization strategies, there is no strict requirement for the joint training of the embedder and extractor. However, the embedder is still tailored to generate high-quality images robust to JPEG compression, while the

extractor is engineered to retrieve features post-JPEG noise. Moreover, Zhang et al. [32] introduced a pseudo-differentiable methodology, designed to accommodate JPEG compression as a specialized noise variant. This approach features distinctive forward and backward paths during the training process. Notably, the backpropagation is structured to bypass the JPEG compression phase, thereby mitigating the impediments associated with non-differentiability.

Certain studies incorporate specialized noise into the noise layer to address special challenges in image watermarking. One intricate area involves extracting watermarks from images that have undergone resampling via a camera, which introduces multifarious noise types including JPEG artifacts, variations in lighting, and optical distortions. In response to this, Fang et al. [33] and Gu et al. [34] advanced approaches that incorporate a screen-shooting noise layer simulation. This adaptation enables the simulation of camera resampling noises like geometric distortions, optical bends, and RGB ripples within the training of deep learning-based image watermarking models, fostering a more robust system capable of counteracting these specific noise introductions.

Existing joint training paradigms necessitate the explicit identification and enumeration of training noise. Models tend to exhibit enhanced robustness to noises encountered during training than those not included. However, in real-world scenarios, anticipating and listing all potential noises can be impracticable. As such, a strand of research is dedicated to forging robust deep learning-based image watermarking models without prior noise knowledge. Zhong et al. [35] introduced an invariance layer designed to sieve out extraneous information during the watermark extraction phase. Within the training ambit, the Frobenius norm of the Jacobian matrix of the invariance layer's outputs with regard to its inputs is computed and employed as a regularization term. The dual objective of minimizing this term, alongside ensuring watermarking requisites (pertaining to the marked image's quality and watermark extraction efficacy), ensures the output of the invariance layer remains largely invariant to alterations in its input images, hence instilling robustness sans explicit noise enumeration. Furthermore, the embedding network employs multi-scale inception networks, facilitating an intricate fusion of the cover image and watermark. Another strategy to achieve robustness without resorting to manual noise layer introduction entails the deployment of an adversarial network, serving as an automated assailant. Luo et al. [36] illustrated this by amalgamating an adversarial network within their architecture, functioning as a noise module. During training, this adversarial entity, interfaced with the extractor, evolves in proficiency, adept at hampering watermark extraction. In counteraction, the extractor strives to mitigate the perturbations induced by the adversarial entity. A nuanced calibration of the training process, accentuating the fortification of the extractor against the adversarial network, culminates in a model characterized by enhanced robustness and reliability.

The joint embedder–extractor paradigm is as a notably effective approach within the existing body of literature. Enhanced performance has been a focal point, with innovations in architectural design and training methodologies spearheading advancements. Ahmadi et al. [37] enriched their noise layer with a variety of disturbances including Gaussian, white noise, random cropping, smoothing, and JPEG compression. Each training iteration involves a stochastic selection of one specific noise type, ensuring that each assault singularly influences the training loss. In another development, Plata et al. [38] unveiled a pre-processor termed a 'propagator', engineered to disseminate the watermark across the image's spatial domain. The researchers stratified assorted attacks and corroborated that integrating specific distortions during training augments robustness against an entire category of distortions. Echoing the two-stage training approach of Liu et al. [29] and mirroring the noise influences highlighted in HiDDeN [7], Zhang et al. [39] introduced a scheme accentuated by a multi-class discriminator connected to the noise-infused marked image. This innovation not only targets robustness but also amplifies the watermark's security within the marked image. Hao et al. [40], while aligning with the visual quality scrutiny embedded in HiDDeN [7], proposed the integration of a high-pass filter at the dis-

criminator's inception. This strategy nudges the watermark into the image's mid-frequency region, safeguarding visual quality given the amendable nature of high-frequency components. The loss computation accords amplified significance to the central region, resonating with the human visual system's focal inclination. For noise layer augmentation, additions encompass crop, crop-out, Gaussian blur, directional flips, and JPEG compression, painting a comprehensive spectrum of distortions. Xu et al. [41] employed a reversible neural network functioning dually as the embedder and extractor, a strategy that is consistent with the traditional, reversible nature of watermarking transformations. In a deviation, Mahapatra et al. [42] advocated for the computation and integration of the difference between the marked and cover images into the extractor to augment extraction quality, a technique that transitions away from the blind scheme archetype. Zhao et al. [43] introduced a factor into the embedder, modulating the watermark's intensity on the cover image, and employed a trained spatial attention feature map to optimize watermark positioning. Ying et al. [44] targeted an enhancement of embedding capacity, accommodating one to three watermark color images and employing a decoupling and revealing network tandem in the extraction phase. Their noise layer is fortified with cropping, scaling, Gaussian noise, JPEG, and Gaussian blur to simulate realistic distortions. In a novel structural approach, Fang et al. [45] introduced an extractor–embedder–extractor training architecture to bolster extractor efficiency. Their extractor transforms an image into a binary watermark sequence while the embedder crafts an image residual from both the original and decoded watermark, enhancing marked image creation and extraction efficacy. This methodology underscores decoder training, prioritizing extraction quality. Incorporating reinforcement learning, Mun et al. [46] enhanced the robustness of the embedder–extractor framework. They amalgamated convolutional neural networks (CNNs) [47] for embedding and extraction and a reinforcement learning actor for noise module operations. The actor, pivotal in selecting and integrating noise types and intensities into marked images, complements the embedder–extractor, which functions as an evaluative environment, assessing the actor's actions and training the extractor to counteract the induced noise efficaciously.

Multimedia has also emerged as a significant focus in the field of deep learning-based image watermarking. Das and Zhong [48], for instance, developed a novel method to embed audio watermarks into cover images, accompanied by a network specifically engineered to determine the fidelity of the extracted audio watermark to its original form. In the domain of document images, Ge et al. [49] proposed a technique enriched with multiple skip connections within the embedder, ensuring the preservation of intricate details in both watermark and cover images. The robustness of the watermarked document images is fortified through the integration of Dropout, crop-out, Gaussian blur, Gaussian noise, resizing, and JPEG techniques in the noise layer. Liao et al. [50] expanded the embedder–extractor watermarking horizon to GIF animations. Their approach entails the employment of three-dimensional deep neural networks, transforming a single watermark into a three-dimensional feature for integration with GIFs. Discriminator networks are employed to evaluate the watermarking efficacy, focusing on maintaining the imperceptibility of the watermarked GIFs, and ensuring visual quality while securing embedded data.

Since the introduction of the embedder–extractor joint training concept (a natural extension of a traditional watermarking paradigm) with HiDDeN [7], a burgeoning body of literature has meticulously expanded upon this initial idea, giving rise to a diverse array of methodologies. As researchers have delved deeper into this domain, an intricate landscape of challenges and corresponding solutions has emerged, each contributing a unique perspective to the overarching narrative of deep learning-based image watermarking. These contributions, marked by their innovative approaches to overcoming specific hurdles, underscore the dynamic and evolving nature of this field. We have cataloged these varied challenges and solutions, presenting a comprehensive summary in Table 3.

Table 3. Summary of the challenges and representative solutions in the embedder–extractor image watermarking.

Challenges	Representative Solutions
The noise layer needs to be differentiable	Performing a two-stage training scheme
The non-differentiable nature and low-performance issues of JPEG	Including differentiable JPEG simulations in the noise layer
Special challenging noises such as the camera resampling	Including simulated camera distortions in the noise layer
Models have more robustness to trained noises than those not included in the noise layer	Developing strategies that do not require noise lists during the training, e.g., an invariance layer that sieves out extraneous information, or an adversarial network to automatically attack the extractor
Aiming at enhanced and improved overall model performance	Introducing innovative architectures and training paradigms aligns with the nuanced processes of embedding, extraction, and feature transformation, mirroring the strategic design inherent in conventional watermarking
Including multimedia for cover images while maintaining high performance	Designing special neural networks to process the multi-modal features and robustness

3.2. Methods Using Deep Networks for Feature Transformation

As illustrated in Figure 3, watermarking procedures within this category predominantly utilize deep neural networks for feature transformations. Both cover and marked images undergo transformations facilitated by these networks, leading to the creation of distinct feature spaces. Subsequent watermark embedding and extraction are executed within these defined spaces. A common expectation is the robustness of the transformed domains, implying that even minor alterations to the marked images should yield consistent or nearly identical feature values.

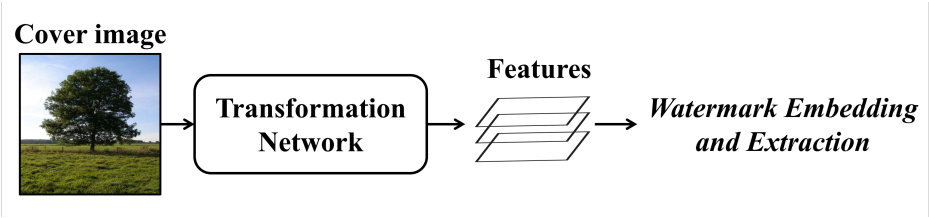

Figure 3. General process of the deep networks for feature transformation.

Numerous methods have adopted the concept of deep networks for feature transformation in the context of deep learning-based zero watermarking. Fierro et al. [26] employed CNNs to extract features from cover images, which were then integrated with a permuted binary watermark sequence via an exclusive or XOR operation to create a master share. The same CNN processes a test image to extract features, which are subsequently XORed with the master share to extract the watermark. An appropriate key can ensure the identification of the watermark. He et al. [51] extended this foundational approach by adding fully connected layers to draw shallow features from various convolutional layers, enhancing the master share creation process. Their introduction of a shrinkage module facilitates the automatic learning of soft thresholding for each feature channel, enhancing feature extraction precision. To optimize the feature space, they focused on eliminating redundancy by learning inter and intra-feature weights and incorporated a noise layer

during feature training to increase robustness. Han et al. [52] enhanced this methodology by introducing a chaotic encryption algorithm to encrypt the watermark before the XOR operation, enhancing security. They also adopted the Swin Transformer [53] to generate features for master share creation, achieving a feature space that is invariant to geometric distortions and enhances the robustness of the watermarking process.

Another research direction in this category involves employing pre-trained deep neural networks, wherein the training of input data is performed to yield the intended marked images. In this scenario, the pre-trained weights remain static, and the alterations in the input are driven toward achieving specific objectives. The resultant marked image is visually analogous to the original cover image, yet reveals the embedded watermark upon undergoing feature extraction by the deep network. Vukotic et al. [54] illustrated this by implementing pre-trained CNN and adaptively modifying the input cover image through gradient descent. The dual-faceted loss function encompasses a term that minimizes the perceptual discrepancy between the cover and marked images and another that ensures watermark detectability via a dot product operation, expressed as $\varphi(m)^T \cdot k$. Here, $\varphi(m)$ denotes the marked image's feature extraction through the deep network and k is a predetermined key, facilitating the detection of watermark presence. Expanding on this, Fernandez et al. [55] introduced the capability of multi-bit extraction by assigning distinct keys to each bit of the binary watermark sequence. Contrary to the utilization of convolutional networks pre-trained for classification, they employed networks that had undergone self-supervised learning. This strategic adoption confers a distinct advantage, as the feature spaces derived from self-supervised learning are characterized by augmented robustness, thereby enhancing the effectiveness and reliability of the watermark extraction process.

The utilization of deep networks for feature transformation represents a nascent avenue in the domain of image watermarking. This approach deviates from the more intuitive embedder–extractor joint training model, which seamlessly aligns with traditional image watermarking paradigms by encapsulating both embedding and extraction processes. Consequently, the academic literature on this innovative method is relatively sparse. Nonetheless, this emerging methodology paves the way for captivating research trajectories, offering fresh perspectives and approaches in the field. To provide a consolidated overview, we have collated the prevailing challenges and representative solutions in Table 4.

Table 4. Summary of the challenges and representative solutions in the image watermarking using deep networks for feature transformation.

Challenges	Representative Solutions
How to utilize the fitting ability of deep learning to extract the cover image feature (the master share) in zero watermarking	Applying off-the-shelf CNNs or Transformers and designing extended branches of these architectures
How to choose appropriate deep learning models for image watermarking given their different feature extraction abilities and purposes	Adopting pre-trained CNNs or the models in self-supervised learning
How to design separate embedding and extraction schemes, given a deep learning feature extractor	To obtain a marked image, fix the pre-trained model, and update the input image with the gradient (produced by a loss ensuring the imperceptibility and extraction integrity)

3.3. Hybrid Methods

Methods encompassed in this category exhibit a fusion of deep learning techniques and traditional calculations associated with image watermarking. Such an integration implies a symbiotic relationship where the strengths of one approach compensate for the weaknesses

of the other, resulting in enhanced efficiency and effectiveness. The design paradigms and operational frameworks of these methods can be diverse, exhibiting a wide range of structural and functional variations. In these hybrid systems, deep learning typically plays a pivotal role in watermark extraction. The complex and intricate architectures of deep learning models offer enhanced capacity for fitting complex functions, and these models are adept at uncovering intricate patterns and correlations within the watermarked images, thereby facilitating the efficient and accurate extraction of embedded watermarks. The conventional image watermarking calculations, on the other hand, lend stability, reliability, and a degree of interpretability to the process. They serve as a solid foundation upon which the deep learning models can build.

Kandi et al. [56] employed two convolutional autoencoders to reconstruct a cover image individually. The distinctions between the autoencoder-reconstructed images and the original cover image are integral to their approach. The first autoencoder's reconstruction denotes bit zeros in a binary watermark, while the second represents bit ones. In a different context, Ferdowsi et al. [57] tailored a technique specifically for Internet of Things (IoT) applications, utilizing classic spread spectrum for watermark embedding, wherein a key pseudo-noise sequence augments the original signal. The innovation lies in mapping features like spectral flatness, mean, variance, skewness, and kurtosis of the cover image to bit streams, serving as the watermark, enhancing security against eavesdropping attacks by eschewing predefined bit streams. Li et al. [58] introduced a method where pre-processed grayscale watermark images are integrated into the DCT blocks of cover images. The extraction of these embedded watermarks is facilitated by training CNN, establishing a bridge between conventional and neural approaches. Mellimi et al. [59] advocated for embedding watermarks into the lifting wavelet domain [60] of cover images. They introduced noise into the marked image and deployed a deep neural network as an extractor, exemplifying the robustness of the infused noise. Zhu et al. [61] innovated a technique amalgamating key point detection with deep learning-based image watermarking. Utilizing SURF [62], they delineated scale-invariant embedding regions, placing normalized binary watermarks at their centers in the Y color channels. This aggregated data are routed through an embedding network, yielding the marked Y channel. An extractor network, fine-tuned through training, facilitates watermark retrieval. Robustness is amplified by the incorporation of perturbations during the training phase, underscoring an enhanced resilience to various forms of distortions and manipulations. Chack et al. [63] introduced a hybrid methodology that intertwines traditional watermarking, CNN, and evolutionary optimization. This multifaceted approach embeds an Arnold-transformed watermark into the DCT domain, employs Harris Hawks optimization to fine-tune the embedding strength, and relies on a CNN to uncover the embedded watermark. In a separate study, Fang et al. [64] presented a deep template-based image watermarking mechanism. The embedding process in their approach encodes the watermark using established techniques, employs an auxiliary locating template to manipulate a pseudo-random Gaussian noise pattern, and integrates the watermark into the red and blue color channels of a cover image. Extraction is facilitated by two deep neural networks; the initial network extracts and accentuates features, while the subsequent network classifies the watermark bit patterns. Kim et al. [65] presented another nuanced technique utilizing templates for watermarking images. Their strategy involves segmenting a cover image into distinct patches, earmarking specific patches for watermark embedding, and others for housing a predefined template. Watermark insertion is executed in the curvelet domain via quantization index modulation, while the template undergoes processing by a dedicated generation network before integration into the cover image. The marked image is derived by assembling the various embedded patches. Extraction is facilitated by a template extraction network that unveils the embedded template, which is subsequently juxtaposed against the original via a template-matching network. This comparison process facilitates the identification of potential geometric distortions inflicted upon the marked image. Chen et al. [66] prioritized the development of a mechanism for authenticating watermark systems via deep learning. Specifically, their framework is adept

at discerning the accuracy of watermark extractions from medical images. Their innovative approach involves simulating a variety of watermark distortions and compiling a labeled dataset. This dataset then undergoes training on a neural network designed to validate the integrity of extractions derived from potentially marked images, thus bridging the gap between watermark verification and deep learning methodologies.

The integration of deep learning and traditional image watermarking has given rise to a plethora of methodologies, each characterized by its distinct approach and underlying principles. Despite the diversity inherent in these hybrid methods, it is noteworthy that they tend to encounter a set of common challenges and have consequently adopted prevailing solutions to mitigate these issues effectively. These challenges largely stem from the complex interplay between the adaptive, data-driven nature of deep learning and the algorithmic, rule-based structure of traditional watermarking. Addressing these issues necessitates a nuanced approach that is sensitive to the strengths and limitations inherent in both paradigms. In Table 5, we have compiled a summary of typical challenges and their corresponding solutions.

Table 5. Summary of the challenges and representative solutions in hybrid methods.

Challenges	Representative Solutions
Determining the optimal role of deep learning in hybrid watermarking frameworks	Employing deep learning to enhance watermark extraction processes
The integration of deep learning and traditional watermarking techniques often results in augmented complexity	Crafting modular and scalable architectures facilitate seamless integration and interoperability between both methodologies
Refining the synergy between embedding and extraction processes is essential, given the distinct strengths and weaknesses inherent to each approach	Utilizing deep learning for its adaptability and learning prowess, complemented by leveraging the proven properties of traditional algorithms

4. Discussion of Potential Future Directions

The proliferation of proposals concerning deep learning-based image watermarking has instigated our comprehensive survey aimed at bridging historical, contemporary, and prospective research. Figure 4 encapsulates the prevailing trends delineated in Section 3 and extrapolates future investigative trajectories. The contemporary focus gravitates toward the intricacies of noise layer differentiation, diversity in noise types, enhancements in architecture and training paradigms, and the strategic integration of deep learning within image watermarking. This survey underscores a spectrum of untapped research avenues that transcend traditional frameworks. These emergent perspectives are poised to foster considerable innovations in this domain, skillfully navigating the complex interplay of robustness, imperceptibility, capacity, and security requisite in the dynamic realm of digital media and communications. The remainder of this section discusses our proposals for potential research directions for the future.

Robustness toward unforeseen noise. Deep learning-based image watermarking models exhibit distinct robustness to various types of noise, a characteristic intricately linked to the specific noise types they are trained on. This variance in robustness is prominently observed when contrasting the model's performance against trained and untrained noise types. Trained noise types refer to those the model has been explicitly exposed to during the training phase, allowing it to develop specialized mechanisms to counteract their effects. Consequently, the model's efficacy in watermark extraction remains largely stable when encountered with these familiar noise types. Conversely, untrained noise types introduce an element of unpredictability. Since the model lacks prior exposure and adaptive development against these noise types, its performance can potentially be

compromised. This differential in robustness underscores the critical importance of a comprehensive training regimen that encompasses a diverse array of noise types to bolster the model's generalization capabilities. Future research could focus on enhancing model adaptability and robustness against untrained noise types, perhaps through the integration of online learning methods [67], or meta-learning [68] strategies that equip the model to swiftly acclimatize to unfamiliar noise environments.

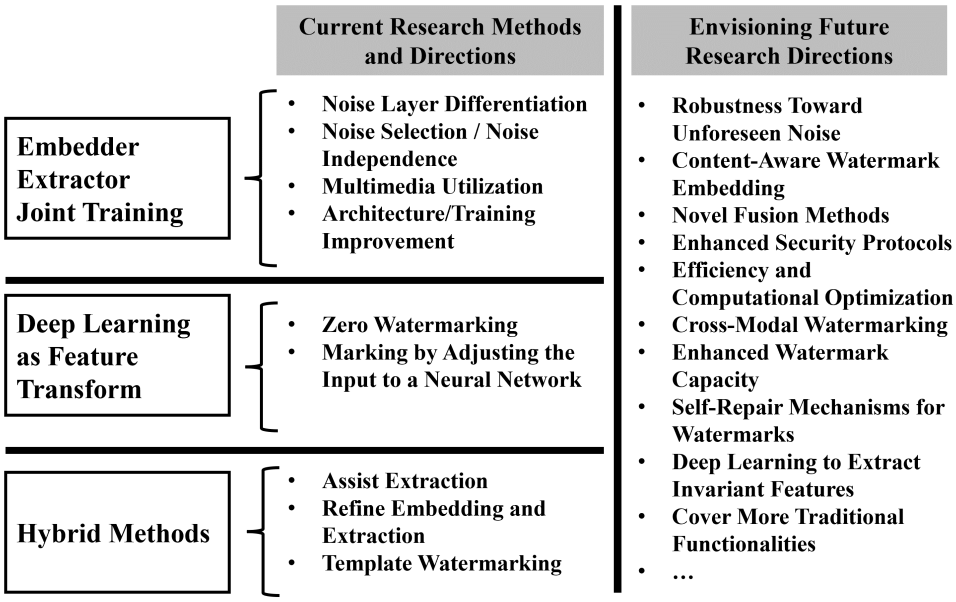

Figure 4. Summarizing and envisioning research directions in deep learning-based image watermarking.

Content-aware watermark embedding. A predominant focus of deep learning-based image watermarking has been accorded to static images. This static orientation potentially undermines the watermark's efficacy, given that optimal embedding strategies can significantly vary across different content types and dynamic scenarios. A transition toward content-aware watermark embedding techniques has the potential to redress this imbalance. This approach, conceptualized to be inherently adaptive, is envisaged to utilize sophisticated algorithms capable of analysis and adaptation to the unique attributes of each image or media sequence. For instance, CNNs or similar deep learning architectures could be trained to discern intricate patterns and variances in visual content, enabling the model to adapt watermark placement and intensity based on different image regions. This would ensure that the watermark is not only imperceptible but also robust against various attacks, establishing a harmonious balance between visibility and security, and marking a significant stride in the advancement of image watermarking technologies.

Novel fusion methods. Investigating innovative algorithms and techniques for the fusion of watermarks within cover images is crucial. A meticulous investigation into innovative embedding strategies is pivotal to determine a harmonious blend that ensures both the visual integrity of the cover image and the resilience of the watermark. Current methods mainly apply additive fusion and concatenation. Additive fusion integrates the watermark into the cover image by additive amalgamation, and concatenation involves the direct attachment of watermark features to the cover image. Future works can focus on embedding algorithms to ensure that watermarks are intricately woven into the cover images, balancing perceptual transparency and robustness against removal or attacks. One prospective method can be cross-attention [69], which can leverage the attention mechanism to selectively focus on specific features of the cover image during the embedding process, ensuring a dynamic and adaptive incorporation of the watermark.

Enhanced security protocols. The imperatives of security, privacy, and integrity are being redefined by the sophistication of adversarial attacks. Consequently, the integration of innovative security protocols is not just desirable, but essential. A compelling research

trajectory could involve the synthesis of cutting-edge cryptographic algorithms with deep learning, an amalgamation promising enhanced watermark protection. The incorporation of blockchain technology presents another frontier, offering decentralized, immutable, and transparent platforms for watermarking data transactions and validations. These multifaceted, integrative approaches are predicated on a nuanced understanding of both deep learning intricacies and the dynamics of contemporary cryptographic paradigms. As we forge ahead, the synthesis of these technologies could engender a new epoch of resilience, privacy, and security in deep learning-based image watermarking.

Efficiency and computational optimization. The dynamic landscape of deep learning-based image watermarking is increasingly underscored by the imperative to balance computational efficiency and processing power, especially for real-time applications. There lies a complex interplay between ensuring robust watermarking and the computational load, where an optimal middle ground is sought to ensure efficiency without compromising performance. In this context, the conceptualization and development of lightweight architectures and algorithms embody a critical focal point of future research trajectories. One promising avenue involves the integration of quantization and pruning techniques [70] within the deep image watermarking models, aiming to reduce the model size while preserving the watermarking efficacy. Furthermore, the exploration of knowledge distillation could facilitate the training of compact models that inherit the performance characteristics of larger, more complex models, thereby ensuring efficiency and efficacy in tandem.

Cross-modal watermarking. In the evolving sphere of deep learning-based image watermarking, cross-modal watermarking emerges as a frontier that offers unprecedented opportunities and challenges. It signifies the confluence of diverse media types, extending the watermarking paradigm beyond its traditional confines, and fostering a multi-dimensional approach to content protection and authentication. Embedding watermarks in images that can be subsequently extracted from audio or video entails a complex interplay of algorithms and technologies, necessitating innovation and adaptability. One methodological prospect could involve the integration of transformer-based models, renowned for their capability to handle varied data types and complexities. Such models can be designed to embed intricate watermark patterns in images, with complementary algorithms tailored for the extraction of these patterns from audio or video formats. A synchronization protocol, ensuring the congruence of embedding and extraction processes across different media, would be integral to this approach.

Self-repair mechanisms for watermarks. The integration of self-repair mechanisms in deep learning-based image watermarking presents an avant-garde approach to enhance the robustness and sustainability of watermarks amidst distortions or attacks. A watermark endowed with self-repair capabilities can significantly augment the reliability of information authentication and integrity verification processes. This concept aligns with the notion of regenerative embedding patterns that maintain their integrity even when subjected to complex distortions or malicious interventions. Algorithmically, this could be achieved through the incorporation of redundant encoding schemes, where the critical information is dispersed within the watermark in a manner that allows for reconstruction from partial data. Error correction codes [21] and machine learning models adept in pattern recognition and restoration [71] can be synergized to enhance the watermark's robustness. By employing neural networks trained to identify and rectify distortions, the watermark's intrinsic characteristics can be preserved.

Deep learning to extract invariant features. Current advancements in deep learning-based watermarking leverage pre-trained neural networks, specifically honed through self-supervised learning, to bolster robustness against noise within the transformed domain. Such advancements are anchored on the premise that various self-supervised networks, especially those that employ joint feature-embedding and contrastive learning methodologies [72,73], are effective in mitigating the effects of various types of noise. These networks ensure that multiple augmentations of a single image yield identical feature representations. Nevertheless, contemporary contrastive learning is predominantly oriented toward

evaluating the representational efficacy of the learned space [74]. The metric for assessing this efficacy hinges on the network's performance in tasks encompassing classification, segmentation, and low-shot learning. Invariant feature training is an ancillary aspect, not the central focus of these learning paradigms [75]. Given this context, the direct application of pre-trained self-supervised neural networks to image watermarking can be impractical, primarily because these networks often neglect to consider ubiquitous distortions for image watermarking like perspective transformations. Consequently, there exists a notable research void warranting exploration—formulating specialized self-supervised neural networks expressly tailored for image watermarking applications. These networks would be instrumental in confronting geometric distortions, including but not limited to, rotations and perspective alterations, underscoring a pivotal frontier for ensuing inquiries.

Cover more traditional functionalities. For deep learning-based image watermarking, a pronounced gap exists in adequately addressing traditional imperatives such as tamper detection. Contemporary methodologies predominantly concentrate on robustness, imperceptibility, and capacity, often sidelining the quintessential aspect of detecting alterations or manipulations in the watermarked images. Classical watermarking techniques have showcased efficacy in this domain, enabling the identification of unauthorized modifications with reasonable accuracy. Incorporating advanced deep learning architectures could potentially elevate the precision and reliability of tamper detection. One plausible approach involves the integration of CNNs trained to discern subtle alterations in the watermarked images, leveraging their capacity for feature extraction and pattern recognition. Another avenue could be the exploration of recurrent neural networks (RNNs) to analyze sequences of image data for temporal alterations, offering insights into the progression of tampering efforts.

The primary focus of this paper resides in a comprehensive examination of image watermarking with deep learning, yet it is acknowledged that there exists a spectrum of compelling research areas that, albeit unexplored in this treatise, hold significant relevance and intrigue. Instances of such topics include the embedding of watermarks within deep learning models to bolster their protection, as highlighted by Uchida et al. [76] and Guo and Potkonjak [77]. Another noteworthy area is the study of watermarking neural networks with watermarked images as input, explored in the work of Wu et al. [78]. Additionally, the fascinating area of deploying attacks on neural networks using watermarked images is explored in the studies conducted by Jiang et al. [79], and Apostolidis and Papakostas [80]. Each of these areas presents a rich vein of inquiry that complements the broader landscape of image watermarking research.

5. Conclusions

In this paper, we delved into the nuanced realm of image watermarking, a technique characterized by the subtle integration and retrieval of watermarks within a cover image. The motivation for this investigation is spurred by the growing synergy between image watermarking and deep learning—a field renowned for its adeptness at unraveling intricate patterns and representations. This study stands as a comprehensive exploration, not merely retracing the trajectories of extant methodologies but exploring the historical context, current innovations, and future prospects of deep learning in image watermarking.

Distinctive in its approach, this survey illuminates the landscape of deep learning-based image watermarking, marked by its precision and depth of analysis. It offers three primary contributions to the scholarly discourse. First, we introduce a systematic classification that segments deep learning-based image watermarking into three core categories: embedder–extractor, deep networks for feature transformation, and hybrid methods. This refined categorization is premised on the diverse roles that deep learning occupies in related studies and is crafted to infuse clarity and direction into ongoing research. Secondly, we examine emblematic methodologies and encapsulate the multifaceted directions and challenges that each category embodies. This aims to provide readers with a consolidated, insightful overview, distilled from a plethora of diverse yet interconnected research. Finally,

our analysis expands to unravel prospective research trajectories, mapping out uncharted territories and emergent themes in the field of deep learning-based image watermarking.

Author Contributions: Conceptualization, X.Z.; methodology, X.Z., A.D., F.A. and A.T.; writing—review and editing, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Science Foundation (NSF) grant number 2104267.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Number of Photos (2023): Statistics, Facts, & Predictions. Available online: https://photutorial.com/photos-statistics/ (accessed on 1 September 2023).
2. Cox, I.; Miller, M.; Bloom, J.; Fridrich, J.; Kalker, T. *Digital Watermarking and Steganography*; Morgan Kaufmann : Burlington, MA, USA, 2007.
3. Barni, M.; Bartolini, F. *Watermarking Systems Engineering: Enabling Digital Assets security and Other Applications*; CRC Press: Boca Raton, FL, USA, 2004.
4. Mastorakis, S.; Zhong, X.; Huang, P.C.; Tourani, R. Dlwiot: Deep learning-based watermarking for authorized iot onboarding. In Proceedings of the 2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 9–12 January 2021; pp. 1–7.
5. Amrit, P.; Singh, A.K. Survey on watermarking methods in the artificial intelligence domain and beyond. *Comput. Commun.* **2022**, *188*, 52–65. [CrossRef]
6. Shih, F.Y. *Digital Watermarking and Steganography: Fundamentals and Techniques*; CRC Press: Boca Raton, FL, USA, 2017.
7. Zhu, J.; Kaplan, R.; Johnson, J.; Fei-Fei, L. Hidden: Hiding data with deep networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 657–672.
8. Baluja, S. Hiding images in plain sight: Deep steganography. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
9. Shih, F.Y.; Zhong, X. High-capacity multiple regions of interest watermarking for medical images. *Inf. Sci.* **2016**, *367*, 648–659. [CrossRef]
10. Chen, B.; Wornell, G.W. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Trans. Inf. Theory* **2001**, *47*, 1423–1443. [CrossRef]
11. Kang, X.; Huang, J.; Zeng, W. Efficient general print-scanning resilient data hiding based on uniform log-polar mapping. *IEEE Trans. Inf. Forensics Secur.* **2010**, *5*, 1–12. [CrossRef]
12. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
13. Li, Y.; Wang, H.; Barni, M. A survey of deep neural network watermarking techniques. *Neurocomputing* **2021**, *461*, 171–193. [CrossRef]
14. Gupta, M.; Kishore, R.R. A survey of watermarking technique using deep neural network architecture. In Proceedings of the 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 19–20 February 2021; pp. 630–635.
15. Zhang, C.; Lin, C.; Benz, P.; Chen, K.; Zhang, W.; Kweon, I.S. A brief survey on deep learning based data hiding. *arXiv* **2021**, arXiv:2103.01607.
16. Byrnes, O.; La, W.; Wang, H.; Ma, C.; Xue, M.; Wu, Q. Data hiding with deep learning: A survey unifying digital watermarking and steganography. *arXiv* **2021**, arXiv:2107.09287.
17. Singh, H.K.; Singh, A.K. Comprehensive review of watermarking techniques in deep-learning environments. *J. Electron. Imaging* **2023**, *32*, 031804. [CrossRef]
18. Cox, I.J.; Kilian, J.; Leighton, F.T.; Shamoon, T. Secure spread spectrum watermarking for multimedia. *IEEE Trans. Image Process.* **1997**, *6*, 1673–1687. [CrossRef]
19. Selent, D. Advanced encryption standard. *Rivier Acad. J.* **2010**, *6*, 1–14.
20. Zhou, X.; Tang, X. Research and implementation of RSA algorithm for encryption and decryption. In Proceedings of the 2011 6th International Forum on Strategic Technology, Harbin, China, 22–24 August 2011, Volume 2, pp. 1118–1121.
21. Peterson, W.W.; Weldon, E.J. *Error-Correcting Codes*; MIT Press: Cambridge, MA, USA, 1972.
22. Wicker, S.B.; Bhargava, V.K. *Reed-Solomon Codes and Their Applications*; John Wiley & Sons: Hoboken, NJ, USA, 1999.
23. Johannesson, R.; Zigangirov, K.S. *Fundamentals of Convolutional Coding*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
24. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]

25. Begum, M.; Uddin, M.S. Digital image watermarking techniques: A review. *Information* **2020**, *11*, 110. [CrossRef]
26. Fierro-Radilla, A.; Nakano-Miyatake, M.; Cedillo-Hernandez, M.; Cleofas-Sanchez, L.; Perez-Meana, H. A robust image zero-watermarking using convolutional neural networks. In Proceedings of the 2019 7th International Workshop on Biometrics and Forensics (IWBF), Cancun, Mexico, 2–3 May 2019; pp. 1–5.
27. Dong, F.; Li, J.; Bhatti, U.A.; Liu, J.; Chen, Y.W.; Li, D. Robust Zero Watermarking Algorithm for Medical Images Based on Improved NasNet-Mobile and DCT. *Electronics* **2023**, *12*, 3444. [CrossRef]
28. Zhang, C.; Benz, P.; Karjauv, A.; Sun, G.; Kweon, I.S. UDH: Universal deep hiding for steganography, watermarking, and light field messaging. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Online, 6–12 December 2020; pp. 10223–10234.
29. Liu, Y.; Guo, M.; Zhang, J.; Zhu, Y.; Xie, X. A novel two-stage separable deep learning framework for practical blind watermarking. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1509–1517.
30. Chen, B.; Wu, Y.; Coatrieux, G.; Chen, X.; Zheng, Y. JSNet: A simulation network of JPEG lossy compression and restoration for robust image watermarking against JPEG attack. *Comput. Vis. Image Underst.* **2020**, *197*, 103015. [CrossRef]
31. Jia, Z.; Fang, H.; Zhang, W. MBRS: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression. In Proceedings of the 29th ACM International Conference on Multimedia, Online, 20–24 October 2021; pp. 41–49.
32. Zhang, C.; Karjauv, A.; Benz, P.; Kweon, I.S. Towards robust deep hiding under non-differentiable distortions for practical blind watermarking. In Proceedings of the 29th ACM International Conference on Multimedia, Online, 20–24 October 2021; pp. 5158–5166.
33. Fang, H.; Jia, Z.; Ma, Z.; Chang, E.C.; Zhang, W. PIMoG: An Effective Screen-shooting Noise-Layer Simulation for Deep-Learning-Based Watermarking Network. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 2267–2275.
34. Gu, W.; Chang, C.C.; Bai, Y.; Fan, Y.; Tao, L.; Li, L. Anti-Screenshot Watermarking Algorithm for Archival Image Based on Deep Learning Model. *Entropy* **2023**, *25*, 288. [CrossRef]
35. Zhong, X.; Huang, P.C.; Mastorakis, S.; Shih, F.Y. An automated and robust image watermarking scheme based on deep neural networks. *IEEE Trans. Multimed.* **2020**, *23*, 1951–1961. [CrossRef]
36. Luo, X.; Zhan, R.; Chang, H.; Yang, F.; Milanfar, P. Distortion agnostic deep watermarking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 13548–13557.
37. Ahmadi, M.; Norouzi, A.; Karimi, N.; Samavi, S.; Emami, A. ReDMark: Framework for residual diffusion watermarking based on deep networks. *Expert Syst. Appl.* **2020**, *146*, 113157. [CrossRef]
38. Plata, M.; Syga, P. Robust spatial-spread deep neural image watermarking. In Proceedings of the 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Guangzhou, China, 29 December–1 January 2020; pp. 62–70.
39. Zhang, L.; Li, W.; Ye, H. A blind watermarking system based on deep learning model. In Proceedings of the 2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Shenyang, China, 20–22 October 2021; pp. 1208–1213.
40. Hao, K.; Feng, G.; Zhang, X. Robust image watermarking based on generative adversarial network. *China Commun.* **2020**, *17*, 131–140. [CrossRef]
41. Xu, H.B.; Wang, R.; Wei, J.; Lu, S.P. A Compact Neural Network-based Algorithm for Robust Image Watermarking. *arXiv* **2021**, arXiv:2112.13491.
42. Mahapatra, D.; Amrit, P.; Singh, O.P.; Singh, A.K.; Agrawal, A.K. Autoencoder-convolutional neural network-based embedding and extraction model for image watermarking. *J. Electron. Imaging* **2023**, *32*, 021604. [CrossRef]
43. Zhao, Y.; Wang, C.; Zhou, X.; Qin, Z. DARI-Mark: Deep Learning and Attention Network for Robust Image Watermarking. *Mathematics* **2022**, *11*, 209. [CrossRef]
44. Ying, Q.; Zhou, H.; Zeng, X.; Xu, H.; Qian, Z.; Zhang, X. Hiding Images into Images with Real-world Robustness. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 111–115.
45. Fang, H.; Jia, Z.; Qiu, Y.; Zhang, J.; Zhang, W.; Chang, E.C. De-END: Decoder-driven Watermarking Network. *arXiv* **2022**, arXiv:2206.13032. .
46. Mun, S.M.; Nam, S.H.; Jang, H.; Kim, D.; Lee, H.K. Finding robust domain from attacks: A learning framework for blind watermarking. *Neurocomputing* **2019**, *337*, 191–202. [CrossRef]
47. Aloysius, N.; Geetha, M. A review on deep convolutional neural networks. In Proceedings of the 2017 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 6–8 April 2017; pp. 0588–0592.
48. Das, A.; Zhong, X. A Deep Learning-based Audio-in-Image Watermarking Scheme. In Proceedings of the 2021 International Conference on Visual Communications and Image Processing (VCIP), Munich, Germany, 5–8 December 2021; pp. 1–5.
49. Ge, S.; Xia, Z.; Fei, J.; Tong, Y.; Weng, J.; Li, M. A robust document image watermarking scheme using deep neural network. *Multimed. Tools Appl.* **2023**, *82*, 38589–38612. [CrossRef]
50. Liao, X.; Peng, J.; Cao, Y. GIFMarking: The robust watermarking for animated GIF based deep learning. *J. Vis. Commun. Image Represent.* **2021**, *79*, 103244. [CrossRef]
51. He, L.; He, Z.; Luo, T.; Song, Y. Shrinkage and Redundant Feature Elimination Network-Based Robust Image Zero-Watermarking. *Symmetry* **2023**, *15*, 964. [CrossRef]

52. Han, B.; Wang, H.; Qiao, D.; Xu, J.; Yan, T. Application of Zero-Watermarking Scheme Based on Swin Transformer for Securing the Metaverse Healthcare Data. *IEEE J. Biomed. Health Inform.* 2023, *Early Access*. [CrossRef] [PubMed]
53. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
54. Vukotić, V.; Chappelier, V.; Furon, T. Are classification deep neural networks good for blind image watermarking? *Entropy* **2020**, *22*, 198. [CrossRef] [PubMed]
55. Fernandez, P.; Sablayrolles, A.; Furon, T.; Jégou, H.; Douze, M. Watermarking images in self-supervised latent spaces. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 3054–3058.
56. Kandi, H.; Mishra, D.; Gorthi, S.R.S. Exploring the learning capabilities of convolutional neural networks for robust image watermarking. *Comput. Secur.* **2017**, *65*, 247–268. [CrossRef]
57. Ferdowsi, A.; Saad, W. Deep learning-based dynamic watermarking for secure signal authentication in the Internet of Things. In Proceedings of the 2018 IEEE International Conference on Communications (ICC), IEEE, Kansas City, MO, USA, 20–24 May 2018; pp. 1–6.
58. Li, D.; Deng, L.; Gupta, B.B.; Wang, H.; Choi, C. A novel CNN based security guaranteed image watermarking generation scenario for smart city applications. *Inf. Sci.* **2019**, *479*, 432–447. [CrossRef]
59. Mellimi, S.; Rajput, V.; Ansari, I.A.; Ahn, C.W. A fast and efficient image watermarking scheme based on deep neural network. *Pattern Recognit. Lett.* **2021**, *151*, 222–228. [CrossRef]
60. Sweldens, W. Lifting scheme: A new philosophy in biorthogonal wavelet constructions. In Proceedings of the Wavelet Applications in Signal and Image Processing III, SPIE, San Diego, CA, USA, 12–14 July 1995; Volume 2569, pp. 68–79.
61. Zhu, L.; Wen, X.; Mo, L.; Ma, J.; Wang, D. Robust location-secured high-definition image watermarking based on key-point detection and deep learning. *Optik* **2021**, *248*, 168194. [CrossRef]
62. Bay, H.; Tuytelaars, T.; Van Gool, L. Surf: Speeded up robust features. In Proceedings of the Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; Proceedings, Part I 9; Springer: Berlin/Heidelberg, Germany, 2006; pp. 404–417.
63. Chacko, A.; Chacko, S. Deep learning-based robust medical image watermarking exploiting DCT and Harris hawks optimization. *Int. J. Intell. Syst.* **2022**, *37*, 4810–4844. [CrossRef]
64. Fang, H.; Chen, D.; Huang, Q.; Zhang, J.; Ma, Z.; Zhang, W.; Yu, N. Deep template-based watermarking. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 1436–1451. [CrossRef]
65. Kim, W.H.; Kang, J.; Mun, S.M.; Hou, J.U. Convolutional neural network architecture for recovering watermark synchronization. *Sensors* **2020**, *20*, 5427. [CrossRef]
66. Chen, Y.P.; Fan, T.Y.; Chao, H.C. Wmnet: A lossless watermarking technique using deep learning for medical image authentication. *Electronics* **2021**, *10*, 932. [CrossRef]
67. Hoi, S.C.; Sahoo, D.; Lu, J.; Zhao, P. Online learning: A comprehensive survey. *Neurocomputing* **2021**, *459*, 249–289. [CrossRef]
68. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 1126–1135.
69. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017, Volume 30.
70. Kim, J. Quantization Robust Pruning With Knowledge Distillation. *IEEE Access* **2023**, *11*, 26419–26426. [CrossRef]
71. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.
72. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, PMLR, Online, 13–18 July 2020; pp. 1597–1607.
73. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9650–9660.
74. Da Costa, V.G.T.; Fini, E.; Nabi, M.; Sebe, N.; Ricci, E. Solo-learn: A library of self-supervised methods for visual representation learning. *J. Mach. Learn. Res.* **2022**, *23*, 2521–2526.
75. Bardes, A.; Ponce, J.; LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv* **2021**, arXiv:2105.04906.
76. Uchida, Y.; Nagai, Y.; Sakazawa, S.; Satoh, S. Embedding watermarks into deep neural networks. In Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, Bucharest, Romania, 6–9 June 2017; pp. 269–277.
77. Guo, J.; Potkonjak, M. Watermarking deep neural networks for embedded systems. In Proceedings of the 2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), San Diego, CA, USA, 5–8 November 2018; pp. 1–8.
78. Wu, H.; Liu, G.; Yao, Y.; Zhang, X. Watermarking neural networks with watermarked images. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 2591–2601. [CrossRef]

79. Jiang, H.; Yang, J.; Hua, G.; Li, L.; Wang, Y.; Tu, S.; Xia, S. Fawa: Fast adversarial watermark attack. *IEEE Trans. Comput.* **2021**, *early access*. [CrossRef]
80. Apostolidis, K.D.; Papakostas, G.A. Digital watermarking as an adversarial attack on medical image analysis with deep learning. *J. Imaging* **2022**, *8*, 155. [CrossRef] [PubMed]