



## A survey of Deep Neural Network watermarking techniques

This is a pre print version of the following article:

*Original:*

Li, Y., Wang, H., Barni, M. (2021). A survey of Deep Neural Network watermarking techniques. NEUROCOMPUTING, 461, 171-193 [10.1016/j.neucom.2021.07.051].

*Availability:*

This version is available <http://hdl.handle.net/11365/1204094> since 2022-04-28T11:00:45Z

*Published:*

DOI:10.1016/j.neucom.2021.07.051

*Terms of use:*

Open Access

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. Works made available under a Creative Commons license can be used according to the terms and conditions of said license.

For all terms of use and more information see the publisher's website.

(Article begins on next page)

# A survey of deep neural network watermarking techniques

Yue Li<sup>a</sup>, Hongxia Wang<sup>b,\*</sup> and Mauro Barni<sup>c</sup>

<sup>a</sup>*School of Information Science & Technology, Southwest Jiaotong University, 611756 Chengdu, China*

<sup>b</sup>*School of Cyber Science and Engineering, Sichuan University, 610065 Chengdu, China*

<sup>c</sup>*Dept. Information Engineering and Mathematics, University of Siena, 53100 Siena, ITALY*

## ARTICLE INFO

### Keywords:

Intellectual property protection

Deep Neural Networks

Watermarking

White box vs black box watermarking

Watermarking and DNN backdoors

## ABSTRACT

Protecting the Intellectual Property Rights (IPR) associated to Deep Neural Networks (DNNs) is a pressing need pushed by the high costs required to train such networks and the importance that DNNs are gaining in our society. Following its use for Multimedia (MM) IPR protection, digital watermarking has recently been considered as a mean to protect the IPR of DNNs. While DNN watermarking inherits some basic concepts and methods from MM watermarking, there are significant differences between the two application areas, calling for the adaptation of media watermarking techniques to the DNN scenario and the development of completely new methods. In this paper, we overview the most recent advances in DNN watermarking, by paying attention to cast it into the bulk of watermarking theory developed during the last two decades, while at the same time highlighting the new challenges and opportunities characterising DNN watermarking. Rather than trying to present a comprehensive description of all the methods proposed so far, we introduce a new taxonomy of DNN watermarking and present a few exemplary methods belonging to each class. We hope that this paper will inspire new research in this exciting area and will help researchers to focus on the most innovative and challenging problems in the field.

## 1. Introduction

Deep Neural Networks (DNNs) are increasingly deployed and commercialised in a wide variety of real-world scenarios due to the unprecedented performance they achieve. Training a DNN is a very expensive process that requires: (i) the availability of massive amounts of, often proprietary, data, capturing different scenarios within the target application; ii) extensive computational resources; iii) the assistance of Deep Learning (DL) experts to carefully fine-tune the network topology (e.g., the type and number of hidden layers), and correctly set the training hyper-parameters, like the learning rate, the batch size, etc. As a consequence, high-performance DNNs should be considered as the intellectual property (IP) of the model owner and be protected accordingly. Inspired by the use of classical watermarking techniques for the protection of property rights associated to multimedia contents, DNN watermarking is receiving increasing attention, and several works have been published leveraging on digital watermarking to address IP protection in the DL domain.


In the last two decades, watermarking technology has been applied to protect multimedia documents, and a vast body of literature has been developed as summarised in several books and surveys [6, 23, 61, 21, 69, 4]. Watermarked contents include audio, still images, video, graphics, text and several other kinds of media [53, 67, 59]. The watermark can be injected into the host document by adding to it a low-amplitude, often pseudo-random, signal, either directly in the sample domain or in a properly transformed domain. In the case of still images, for instance, the watermark can be

embedded in the spatial domain by adding a low amplitude spread spectrum signal or by substituting the least significant bits (LSB) of the pixel values [22, 58, 68]. In the case of transformed domain watermarking, several reversible transforms like the Discrete Cosine Transform (DCT) [7, 51], the Discrete Wavelet Transform (DWT) [30], or the Discrete Fourier Transform (DFT) [54, 74] are widely used to embed the watermark in a robust and imperceptible way. For audio watermarking, a number of time domain methods have been proposed, including the pioneering works in [8] and [33], and several other solutions borrowed from the image watermarking field and adapted to work with audio signals [3, 83, 40]. In the case of video watermarking, we can distinguish between techniques operating on the pixel values of video frames, possibly treating them as still images [65], and techniques working in the compressed domain. In the latter case, the watermarking algorithms are tightly tied to compression standards, like, for instance, H.264/AVC [48] and HEVC [72].

In all cases, the watermarking process exploits some forms of redundancy present in the host document, thanks to which the document can be modified without impairing its informative or perceptual meaning. A similar idea holds for the case of DNN watermarking. The very large number of parameters (the network weights) DNN models consist of, confers to the network a capability of analysing the input data that often exceeds the difficulty of the task the network is trained for, hence leaving many degrees of freedom in the choice of the exact weights. The weights, then, can be modified, or directly generated, in such a way to host the watermark.

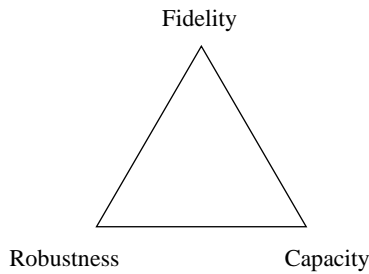
Despite the differences between watermarking techniques thought to work with different media and in different application scenarios, the requirements that any watermarking system must satisfy can be summarised by the so-called water-

\*Corresponding author

 liyue85900040@my.swjtu.edu.cn (Y. Li); hxwang@scu.edu.cn (H.

Wang); barni@dii.unisi.it (M. Barni)

ORCID(s):



**Figure 1:** The watermarking trade-off triangle

marking trade-off triangle shown in Figure 1. According to such a perspective, the capacity requirement, identifiable as the number of bits, or payload, conveyed by the watermark, conflicts with other two requirements, namely: i) fidelity, corresponding to watermark imperceptibility in the case of conventional media, and ii) robustness, that is the ability to recover the watermark even when the hosting document undergoes some modifications. DNN watermarking obeys to the same principles. Here fidelity refers to the capability of the watermarked network to accomplish the task it is thought for. Robustness is related to the possibility of correctly extracting the watermark from a slightly modified version of the network model (e.g. after fine tuning), while capacity indicates the size of the payload conveyed by the watermark.

A noticeable difference between DNN watermarking and multimedia watermarking is that in the case of DNNs, the injection of the watermark can not be carried out by directly modifying the weights of the model, since in this way it would be difficult to estimate the impact of the watermark on the performance of the network. On the contrary, watermark embedding is carried out during the training process, by properly modifying the loss function used in such a phase.

Another peculiarity of DNN watermarking is where the watermark can be read from. In a first case (referred to as *static* watermarking), the watermark can be read directly from the network weights, in a way that is similar to conventional multimedia watermarking techniques. In other cases, though, the effect of the watermark is to alter the *behaviour* of the watermarked model, when the network is fed with some specific inputs. In this case, hereafter referred to as *dynamic* watermarking, the watermark message is read by looking at the output of the model, or the values of the intermediate activation maps, in correspondence of properly crafted inputs. This marks a significant different with respect to classical multimedia watermarking, wherein only the static modality is possible.

By recognizing the similarities and dissimilarities between multimedia and DNN watermarking, the first goal of this paper, is to provide a taxonomy of DNN watermarking techniques based on a mix of conventional classification means and some brand new perspectives. In doing so, we pay great attention to list the essential requirements that a DNN watermarking scheme must satisfy and interpret them for the different classes of watermarking techniques. As a second goal, we review some of the most popular and best

performing DNN watermarking algorithms proposed so far, to give the reader a clear understanding of the practical challenges and opportunities of DNN watermarking.

The rest of the paper is organized as follows. In Section 2, we introduce a taxonomy of DNN watermarking comparing it with the conventional taxonomy used for multimedia watermarking. The main requirements of DNN watermarking are discussed in Section 3. In Section 4 and 5, we review, respectively, the main static and dynamic watermarking algorithm proposed so far. Section 6 is dedicated to the presentation of the most popular attacks against DNN watermarking algorithms. Eventually, in Section 7, we draw some conclusions and highlight some directions for future works.

## 2. A taxonomy of DNN watermarking

In this section, we present a taxonomy of DNN watermarking techniques. The taxonomy takes into account the similarities and dissimilarities between multimedia and DNN watermarking, some of which have already been outlined in the introduction. To start with, we review the most important classification criteria used to categorise classic watermarking algorithms, discussing their applicability to the DNN case. Later, we introduce some unique classification criteria based on the peculiar characteristics of DNN watermarking. Eventually, we discuss the relationship between the various watermarking classes.

### 2.1. Classical watermarking models

Classically, digital watermarking aims at embedding a certain message, called watermark, within a hosting multimedia content, often, but not only, for copyright protection. Watermark injection follows the general model depicted in Figure 2. The to-be-watermarked content  $A$  is possibly transformed so that the watermark is embedded in a different domain (e.g. the DCT or wavelet domain for still images)<sup>1</sup>, then the watermark is injected within the document, with the possible intervention of a watermarking key  $K$ . The watermarked content is then brought back into the original domain to obtain the watermarked document  $A_w$ . As simple as this scheme may seem, it contains some implicit assumptions that are not necessarily satisfied in the case of DNN watermarking: i) the existence of a to-be-watermarked content  $A$  and ii) the consequent possibility of measuring the distortion between the original and the watermarked contents,  $A$  and  $A_w$ . This is not necessarily the case with DNN watermarking, where the network weights, which are going to host the watermark, may not exist *per se*, since they are generated contextually to watermark embedding during the training phase. Moreover, the distortion introduced by the watermark can not be evaluated directly by measuring the difference between  $A$  and  $A_w$ . Rather, the impact of the watermark must be measured by evaluating the performance achieved by the watermarked model.

Note that watermark embedding (and later on watermark recovery) may require the knowledge of a secret key  $K$ . Such

<sup>1</sup>The direct and inverse transforms are optional steps.

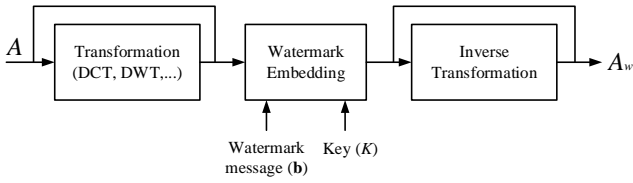


Figure 2: Overall view of classical watermark embedding.

a *key*, whose main goal is to introduce some secrecy within the watermarking system, is commonly used to parameterize the embedding process and make the recovery of the watermark impossible for non-authorized users who do not know  $K$ .

### 2.1.1. Multi-bit vs. zero-bit watermarking techniques

Depending on the exact content of the watermark message, and the way such a message is recovered from the host signal, we can distinguish between two main classes of algorithms: multi-bit and 0-bit watermarking. In the multi-bit case, the watermark message corresponds to a sequence of  $N$  bits  $\mathbf{b}$ . Such a sequence can be read from the watermarked content, as depicted in part (a) of Figure 3. In the 0-bit case, watermark extraction corresponds to a detection task, wherein the detector is asked to decide whether a known watermark is present in the analyzed content or not. In some applications, a mixture of the two functionalities is required. The detector must first determine if a watermark is present and, if so, identify which of the  $2^N$  messages is encoded. Such a detector would therefore have  $2^N + 1$  possible output values [21]. In the zero-bit case, only one possible watermark exists. In this case only  $2^0 + 1 = 2$  outputs are possible, thus justifying the *zero-bit watermarking* term.

In general, multi-bit watermarking offers more flexibility<sup>2</sup> and hence it can be used in a wider variety of applications, including fingerprinting [42], error concealment [2], source tracking [9], labelling etc. Zero-bit watermarking generally achieves a higher robustness and is widely adopted in copyright protection platforms, where the presence of the watermark is used as a flag to warn compliant devices that the piece of content they are dealing with is copyrighted material [29]. The distinction between multi-bit and zero-bit watermarking is also appropriate for DNN watermarking techniques.

### 2.1.2. Robust vs. fragile watermarking

Watermark robustness accounts for the capability of the hidden data to survive host signal manipulations, including non-malicious and malicious manipulations. Malicious manipulations precisely aim at damaging or removing the watermark, on the contrary, non-malicious manipulations indicate some unintentional or unavoidable processing that may perturb the hidden watermark, for instance, multimedia compression occurring during image or video transmission. *Robust watermarking* requires the watermark to be resistant

<sup>2</sup>It can be shown that a multi-bit watermarking algorithm can always be transformed into a zero-bit scheme [6].

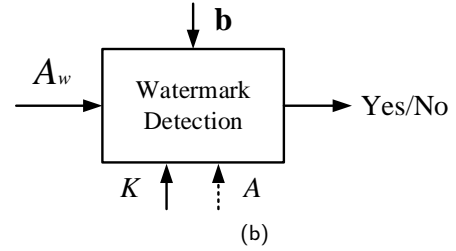
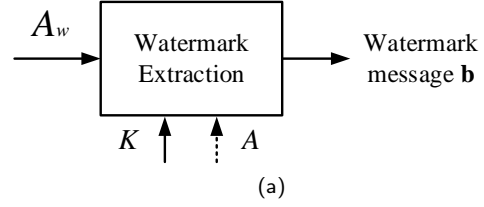


Figure 3: Multibit (a) vs zero-bit (b) watermarking

against non-malicious manipulations, whose application is implied by the normal use of the data, such as data compression, editing and cropping. A secure watermark, on the contrary, should survive also malicious manipulations. If the hidden watermark is lost or irremediably altered as soon as any modification is applied to the host content, the watermark is said to be fragile. Data authentication is the main application for fragile watermarking, because the alteration of the watermark can be taken as an evidence that the content has been tampered with.

With regard to DNN watermarking, most of the applications considered so far call for the adoption of a robust watermark, that can survive at least some common operations like fine-tuning or nodes pruning. Watermark security may also be required in specific applications wherein the presence of an adversary aiming at watermark removal can not be ruled out.

### 2.1.3. Blind vs. non-blind detection

From Figure 3, we can see that for both the multi-bit and the zero-bit cases, watermark recovery may require the knowledge of the non-watermarked content  $A$ . In classic watermarking theory, we say that a watermarking algorithm is *blind*<sup>3</sup> if the watermark is recovered without resorting to the comparison between the original non-marked content and the marked one. Conversely, *non-blind* watermarking algorithms use the original non-marked content to ease the extraction. As we already noted, in the case of DNN watermarking, the concept of original non-watermarked content does not apply, so the distinction between blind and non-blind watermarking does not make sense.

### 2.1.4. Informed watermarking

Informed embedding and informed coding are watermarking paradigms that have been proven to greatly improve the performance of a watermarking system [55]. Such paradigms stem from the interpretation of watermarking as a problem

<sup>3</sup>The term oblivious may also be used.

of channel coding with side information at the transmitter [24, 20]. In a nutshell, with informed coding each watermark message is associated to a pool of codewords (rather than to a single one), then informed watermark embedding is applied by choosing the codeword that results in the minimum distortion. Practical implementations of the informed coding paradigm include several popular watermarking schemes like Quantization Index Modulation (QIM) [11], Dither Modulation (DM) [10] and Scalar Costa's Scheme (SCS) [26]. Informed watermarking theory provides powerful means to improve the performance of watermarking algorithms, with particular reference to multi-bit watermarking, since its adoption permits to increase significantly the watermark payload without sacrificing its robustness. Nevertheless, such a theory has not been applied extensively to DNN watermarking. A DNN watermarking algorithm exploiting informed coding to increase the watermark payload is described in Section 4.

## 2.2. DNN watermarking models

Deep learning is a machine learning framework which automatically learns hierarchical data representation from training data without the need to resort to handcrafted feature representations [31]. DNNs are the most common learning architectures on which deep learning methods are based. To be specific, a DNN takes as input the content  $x \in \mathbb{R}^m$  to be processed in raw format (for instance an image or an audio signal), and maps it to the output via a parametric function,  $z = F_\theta(x)$ , where  $z \in [1, n]$ . The parametric function  $F_\theta(\cdot)$  is defined by the network architecture and the collective parameters of all the units composing it. Given the architecture, the network behaviour is determined by the values of the network parameters  $\theta$  (in most cases  $\theta$  corresponds to the network weights and offsets). Let  $D = \{x_i, z_i\}_{i=1}^T$  be the training data, where  $z_i \in [1, n]$  is the ground truth label for  $x_i$ . During the training phase, the network parameters are optimised to minimise a loss function expressing the difference between the predicted class labels and the ground truth labels. Currently, the most widely used approach for training a DNN is the back-propagation algorithm, whereby the network parameters are updated by propagating the gradient of the loss from the output layer through the entire network. DNN watermarking is based on the observation that the huge number of parameters  $\theta$  consists of, allows to slightly modify them, without that the performance of the network is degraded significantly. The parameters in  $\theta$ , then, can be used to encode additional information beyond what is required for the primary task of the network.

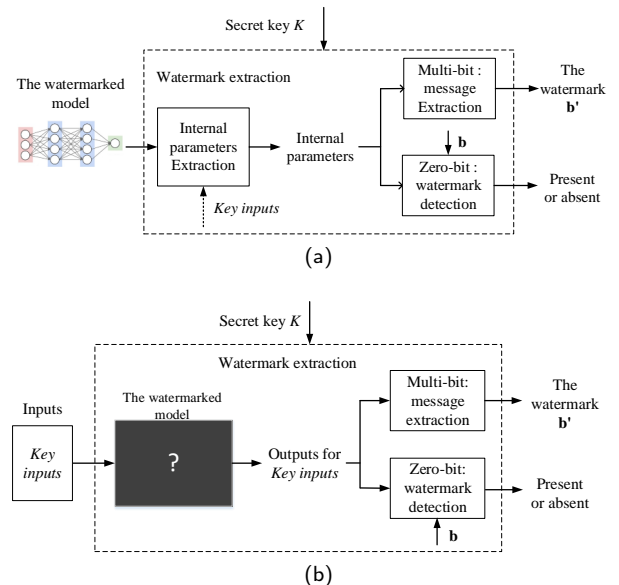
While the classical taxonomy of watermarking algorithms described in the previous section can be largely applied also to DNN watermarking, in the following, we introduce two new characterisations that are specific for DNN watermarking.

### 2.2.1. White-box vs. black-box DNN extraction

Based on the data accessible to the watermark extractor, we can divide DNN watermarking techniques into white-box and black-box techniques. As depicted in Figure 4a, if the internal parameters of the DNN models are available, we

say that watermark recovery is carried out in a white-box modality. The internal parameters may correspond directly to the model weights, or to the activation of the neurons in correspondence to specific inputs. In the case of multi-bit watermarking, the extractor, now called watermark decoder, extracts the concrete message bits the watermark consists of. For zero-bit watermarking, the watermark detector, must decide about the existence of a specific watermark. In some algorithms (e.g. [75, 56, 15, 63]), the key  $K$  is chosen independently from the original loss function  $E_0$  and the internal parameters  $\theta$ , during a key generation step.

With black-box watermarking (see Figure 4b), only the final output of the DNN is accessible. In this case, the watermark is recovered by querying the model and checking the output of the DNN in correspondence to a set of properly chosen inputs. During the entire decoding or detection process, the architecture and the internal parameters of the DNN model are totally blind to the decoder or detector. In other words, the only thing that can be controlled are the inputs used for querying the network and the outputs corresponding to the queries. Thus, the main target of black-box watermarking is to train the DNN model in such a way that it outputs particular labels  $z_i$  for certain inputs  $x_i$ 's. The inputs  $x_i$ 's may be secret or not. In the former case they play a role similar to a decoding/detection key.



**Figure 4:** White-box (a) vs black-box (b) DNN watermark recovery.

### 2.2.2. Static vs. dynamic DNN watermarking

A very important distinction of DNN watermarking techniques leads to the definition of static and dynamic watermarking. Static DNN watermarking methods, like those described in [75, 56, 15], embed the watermark into the weights of the DNN model. Such weights are determined during the training phase and assume fixed values that do not depend on the input of the network. With dynamic watermarking, instead, the watermark is associated to the behaviour

of the network in correspondence to specific inputs, sometimes called triggering inputs, or key inputs (see for instance [63, 45, 87]). Even in this case, the watermark is embedded by properly chosen the network weights, however the watermark is retrieved indirectly by looking at the impact of the watermark on the behaviour of the network. If the watermark is recovered by looking at the final output of the model (sometimes referred to as *DNN-output* watermarking), the watermark can be extracted in a black-box manner, since access to the internal status of the network is not required. When the watermark is associated to the activation map of the neurons in correspondence to certain inputs, as in [63], white-box extraction is required. The distinction between static and dynamic watermarking is illustrated in Figure 5.

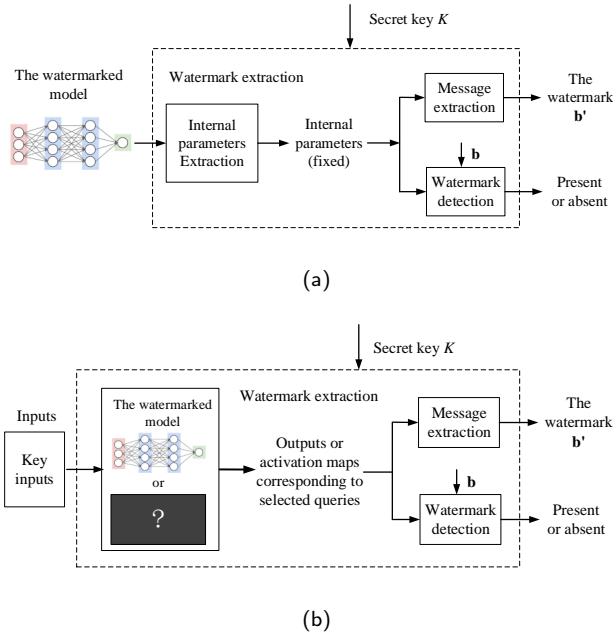


Figure 5: Static (a) vs Dynamic (b) DNN watermarking

### 2.3. Relationship among different DNN watermarking classes

Table 1 summarises the various perspectives that can be adopted to classify watermarking techniques and their applicability to conventional (multimedia) and DNN watermarking. The symbol  $\checkmark$  indicates that the corresponding classification criterion is suitable for a specific watermarking technique, while  $\times$  represents inappropriateness.

Analyzing the intersection among these classifications, we see that dynamic watermarking can be implemented both in white box (activation-map based watermarking) and in black-box modality (DNN-output watermarking). On the contrary, static watermarking can only be achieved in a white-box mode, since it requires the access to the internal parameters of the network. In other words, white box can be static and dynamic, but black-box can only be dynamic. Referring to prior researches, white-box methods have been used for both multi-bit, and zero-bit watermarking, while the black-box modality, due to its lower expressivity, is more often

Table 1

Taxonomy of classic and DNN watermarking.

	Multi/Zero-bit	Robust/Fragile	Blind/Non-blind
Classic	$\checkmark$	$\checkmark$	$\checkmark$
DNN	$\checkmark$	Robust only	$\times$
	Informed/Non-informed	White/Black-box	Static/Dynamic
Classic	$\checkmark$	$\times$	$\times$
DNN	$\checkmark$	$\checkmark$	$\checkmark$

used in conjunction with zero-bit watermarking.

## 3. Requirements

In this section, we pause to discuss the main requirements that DNN algorithms must satisfy. While many of them are inherited by classical multimedia watermarking, there are some important differences that are worth attention. The most common requirements for DNN watermarking are summarised in Table 2, and briefly discussed in the following.

### 3.1. Robustness

Together with security, robustness is one of the three corners of the watermarking trade-off triangle. It refers to the possibility of recovering the watermark from a perturbed version of the host content, be it a multimedia document or a DNN. Usually, it is required that the watermark survives at least the most common manipulations that the host content may undergo in its lifecycle. For DNNs, the two most common operations the watermark should be robust to are: fine-tuning and network pruning.

**Fine-tuning.** Fine tuning is a common operation related to transfer learning. It consists in retraining a model that was initially trained to solve a given task, so to adapt it to solve a new task (possibly related to the original one). Computationally, fine-tuning is by far less expensive than training a model from scratch, hence it is often applied by users to adapt a pre-trained model to their needs. Needless to say, when sufficient training data is not available for the user, model fine-tuning can help avoiding over-fitting problems, at least to a certain extent. Since fine-tuning alters the weights of the watermarked model, it is necessary to make sure that the watermark is robust against a moderate amount of fine-tuning.

**Network pruning.** This is a common strategy to simplify a complex DNN model to deploy it into low power or computationally weak devices like embedded systems or mobile devices. During pruning, the model weights whose absolute value is smaller than a threshold are cut-off to zero. The threshold is set in such a way that the accuracy of the model does not decrease significantly. Of course, network pruning changes the internal parameters of the model, hence it is necessary to make sure that the embedded watermark is

**Table 2**  
List of requirements for DNN watermarking.

Requirement	Description
Robustness	The embedded watermark should resist different kinds of processing.
Security	The watermark should be secure against intentional attacks from an unauthorized party.
Fidelity	The watermark embedding should not significantly affect the accuracy of the target DNN architectures.
Capacity	A multi-bit watermarking scheme should allow to embed as much information as possible into the host target DNN.
Integrity	The bit error rate should be zero (or negligible) for multibit watermarking and the false alarm and missed detection probabilities should be small for the zero-bit case.
Generality	The watermarking methodology can be applied to various DNN architectures and datasets.
Efficiency	The computational overhead of watermark embedding and extraction processes should be negligible.

resistant to this operation.

### 3.2. Security

This criteria deals with malicious manipulations, by assuming that an attacker, partially or fully aware of the watermarking algorithm, is present and attempts to destroy it, with the further goal of using the watermarked content illegally. For a secure DNN watermarking algorithm, the loss of the watermark should be obtainable only at the expense of a significant degradation on the quality of the host model. Until now, research has focused mainly on two kinds of intentional attacks, watermark overwriting and surrogate model attack.

**Watermark overwriting.** A third-party who is aware of the methodology used for DNN watermarking (but does not know the private key used to embed the watermark) may try to embed a new watermark in the model and overwrite the original one. An overwriting attack, then, inserts an additional watermark into the model to make the original watermark undetectable. Several degrees of security can be defined according to the knowledge the attacker has about the watermark. When the watermark is inserted directly in the network weights, for instance, more powerful attacks can be conceived if the attacker knows the exact layers the watermark is embedded in.

**Surrogate model attack.** In a surrogate model attack, the adversary tries to replicate the functionality of a target DNN, by feeding it with a series of requests and using the output provided by the network to train a surrogate model, which mimics the original network. During the attack, the adversary has no knowledge on the exact architecture of the model and limited access to the original training dataset. Ideally, if the network targeted by the attack is watermarked, the surrogate network trained by querying the watermarked model should also inherit the watermark, thus making it possible to recover the watermark from the surrogate model.

### 3.3. Fidelity

Fidelity represents the second corner of the watermarking trade-off triangle. It basically requires that the presence of the watermark does not degrade the *quality* of the watermarked object. In the case of DNNs, this means that the watermark should have a limited impact on the performance of the watermarked model  $F_{\theta'}$ . More specifically, the water-

marked model  $F_{\theta'}$  should guarantee a performance level that is as close as possible to that achieved by a model  $F_{\theta}$  trained on the same task without caring about the watermark. As we will see later on, in some cases, training the network by adding a watermark embedding term to the loss function can even be beneficial, since it reduces the risk of overfitting. In the following sections, fidelity will be measured in terms of the Test Error Rate (TER) achieved on the task the model is designed for.

### 3.4. Capacity

Capacity is the third corner of the watermarking trade-off triangle. It is defined as the number of bits the watermark message consists of. As such, it can be referred only to multi-bit watermarking, since in the zero-bit case, the watermark does not convey any payload. While a large payload is a desirable feature of a watermarking algorithm, increasing the payload conflicts directly with the robustness requirement. The most straightforward way to increase the robustness, while also observing the fidelity criterion, in fact, is to spread the watermark over a large number of host samples (the network weights in the DNN case), thus rapidly exhausting the room available for watermark embedding. As an alternative, error correction coding can be used on top of the watermarking algorithm, to allow the recovery of some bits possibly lost because of the manipulation of the DNN. Even in this case, however, the room available to host the net payload decreases, thus raising again the necessity to find a tradeoff between the number of redundancy bits, which in turn determines the error correcting capability of the code, and the actual payload of the watermark.

### 3.5. Integrity

In the absence of model modifications, the extracted watermark  $\mathbf{b}'$  should be equal to  $\mathbf{b}$ . We use the Bit Error Rate (BER) to evaluate the integrity of a multi-bit algorithm. Ideally, in the absence of processing or attacks, the BER should be equal to zero. Classically, this is a characteristic achievable by the class of informed watermarking algorithms [6], while it can not be guaranteed by non-informed spread spectrum techniques [22]. Interestingly, several DNN watermarking algorithms directly generate the model weights in such a way that the watermark can be extracted without errors. This

resembles very closely the informed embedding paradigm [55], according to which embedding is achieved by applying a signal-dependent perturbation to the to-be-watermarked sequence. In this way, the embedding procedure results in a watermark which, in the absence of modifications, can be recovered with no errors.

The integrity requirement assumes a different meaning in the case of zero-bit watermarking. In this case, watermark recovery can be stated as a detection problem and as such is prone to two kinds of errors: false detection and missed detection<sup>4</sup>. The former refers to the probability that the presence of the watermark is detected in a non-watermarked model, while the second indicates the probability that the watermark can not be recovered from a watermarked network. As it is known from statistical detection theory [39], decreasing the false detection probability comes at the price of an increase of the missed detection probability (and viceversa), so a trade off must be looked for. In DNN watermarking, it is pretty easy to design the watermark embedding algorithm in such a way that the missed detection probability is equal to zero<sup>5</sup>, while it is impossible to guarantee that the false detection probability is equal to zero. The only possibility, then, is to carry out a statistical analysis to give an estimate of the false detection probability and design the watermarking algorithm so that such a probability is acceptably small.

### 3.6. Other Watermark Requirements

Based on the choice of a specific watermarking modality among those listed in Section 2.2, and on the envisaged application, other criteria may need to be satisfied.

*Generality:* To be effective, a DNN watermarking algorithm should be applicable to a wide variety of architectures carrying out different tasks. In this sense, the generality of an algorithm refers to its applicability to architectures and models other than those on which the algorithm was initially tested and developed. This also implies that the performance of the algorithm do not depend too heavily on the to-be-marked model.

*Efficiency:* Efficiency refers to the computational overhead needed to train the DNN by simultaneously teaching the model to carry out the target task and embedding the watermark. Training a DNN is always an expensive process, embedding the watermark into the DNN should not add an unaffordable extra burden to it.

## 4. Static Watermarking Algorithms

In this section, we describe some examples of static watermarking algorithms. The goal is to illustrate with practical examples, how the watermark can be embedded directly into the weights of the network model and later on extracted from them. An exhaustive list of static DNN watermarking algorithms, is provided at the end of the section.

<sup>4</sup>The terms false positive and false negative are also used.

<sup>5</sup>Once again this possibility is due to the adoption of an informed embedding strategy.

So far, we have discussed the main properties of DNN watermarking by referring to generic models with generic inputs. By considering that the great majority of the watermarking algorithms proposed until now have been applied to image classification networks, in the following (Sections 4, 5 and 6) we will always assume that the inputs of the DNN are images which must be classified into one of several possible classes.

### 4.1. Uchida et al. 's algorithm [75]

This is one of the first multi-bit techniques embedding the watermark message directly into the weights of the network model.

For a selected convolutional layer, let  $(s, s)$ ,  $d$ , and  $n$  represent, respectively, the kernel size of the filters, the depth of the input and the number of filters. Ignoring the bias term, the weights of the selected layer can be denoted by a tensor  $\mathbf{W} \in \mathbb{R}^{s \times s \times d \times n}$ . Before embedding the watermark into the weights, the tensor  $\mathbf{W}$  is *flattened* according to the following steps: i) calculate the mean of  $\mathbf{W}$  over the  $n$  filters, getting  $\bar{\mathbf{W}} \in \mathbb{R}^{s \times s \times d}$  with  $\bar{W}_{ijk} = \frac{1}{n} \sum_{h=1}^n W_{ijkh}$ , in order to eliminate the effect of the order of the filters; ii) flatten  $\bar{\mathbf{W}}$  producing a vector  $\mathbf{w} \in \mathbb{R}^v$  with  $v = s \times s \times d$ . Watermark embedding is achieved by training the DNN with an additional loss term ensuring that the watermark bits can be correctly extracted from  $\mathbf{w}$ . Specifically, the loss function  $E(\mathbf{w})$  used to train the network and simultaneously embed the watermark into  $\mathbf{w}$  is given by:

$$E(\mathbf{w}) = E_0(\mathbf{w}) + \lambda E_R(\mathbf{w}) \quad (1)$$

where  $E_0(\mathbf{w})$  represents the original loss function of the network (ensuring a good behavior with regard to the classification task),  $E_R(\mathbf{w})$  is the regularization term added to ensure correct watermark decoding, and  $\lambda$  is a parameter adjusting the tradeoff between the original loss term and the regularization term. As we said, the goal of  $E_R(\mathbf{w})$  is to make sure that the watermark is extracted from  $\mathbf{w}$  with no errors. Specifically,  $E_R(\mathbf{w})$  is given by

$$E_R(\mathbf{w}) = - \sum_{j=1}^l (b_j \log(y_j) + (1 - b_j) \log(1 - y_j)) \quad (2)$$

where  $b_j$  is the  $j$ -th bit of the watermark message (whose length is equal to  $l$ ), and  $y_j$  is the corresponding bit extracted by the watermark decoder. In particular, we have:

$$y_j = \sigma \left( \sum_i X_{ji} w_i \right) \quad (3)$$

where  $\sigma(\cdot)$  is the sigmoid function defined as<sup>6</sup>:

$$\sigma(x) = \frac{1}{1 + e^{-\gamma x}} \quad (4)$$

<sup>6</sup>We introduced the parameter  $\gamma$  for sake of generality; in [75], we have  $\gamma = 1$ .



Here  $\mathbf{X}$  is a spreading matrix, somewhat playing the role of a watermarking key, whose  $j$ -th row ( $X_j$ ) is responsible of spreading the  $j$ -th bit onto a pseudorandom direction. In particular, in [75],  $\mathbf{X}$  is built by randomly drawing its elements according to a zero mean, unitary variance, Gaussian distribution  $N(0, 1)$ . We observe that with this choice all the bits in  $\mathbf{b}$  are cast into the same weights, however, they can be recovered with no error due to the statistical orthogonality of the rows of  $\mathbf{X}$ .

Watermark retrieval is pretty simple, since it consists in computing the projection of  $\mathbf{w}$  onto each  $X_j$ , and thresholding the projection at 0, that is:

$$\hat{b}_j = \begin{cases} 1 & \sum_i X_{ji} w_i \geq 0, \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

By training a wide residual network (WRN [84]) on the CIFAR-10 dataset (whose accuracy when trained without watermark corresponds to a test error rate equal to 8.04%), Uchida et al.'s demonstrated that watermark embedding can be achieved without impacting significantly the classification accuracy, as reported in Table 3. Interestingly, due to the use of the spreading matrix  $\mathbf{X}$ , it is possible to embed a payload which is larger than the number of weights available for embedding. The number of weights for the convolutional levels 2, 3 and 4, in fact, are 576, 1152, and 2304, however the maximum embeddable payload with BER = 0 is, respectively, 1024, 2048 and 4096. As it can be seen from the table, for a given convolutional layer, the payload is limited by the *Integrity* requirement. When the payload exceeds a certain level, in fact, it is no more possible to embed the watermark ensuring a zero BER.

The robustness of Uchida et al.'s watermarking algorithm is exemplified in Tables 4 and 5. With regard to fine tuning (Table 4), the watermarked model was fine-tuned for 20 epochs on the same CIFAR-10 dataset. As it can be seen, the test error does not change significantly, and the watermark BER remains zero. Table 5, reports robustness results against pruning. Specifically, we randomly set to zero a percentage  $p\%$  of the trainable parameters of the embedding layer. The results in the table prove the good robustness of Uchida et al.'s algorithm against pruning. On the other hand, overwriting the watermark with additional 256-bits (using a different spreading matrix) resulted in a large BER, thus showing the weakness of the algorithm against an intentional overwriting attack.

#### 4.2. ST-DM DNN watermarking [47]

Uchida's algorithm is designed according to a classical spread spectrum strategy [22]. Given that the weights are generated directly in such a way that the watermark is recovered correctly, the embedding procedure obeys the informed embedding paradigm [55]. In contrast, no attempt is made to exploit informed coding to increase the payload or diminish the impact of watermarking on the accuracy of the watermarked model. An example of the use of informed coding for DNN watermarking is given in [47]. The algorithm proposed in such a work relies on a new loss function defined

**Table 3**

TER and BER for different convolutional layers with various payloads, for Uchida et al.'s algorithm [75].

Payloads (bit)	Conv 2		Conv 3		Conv 4	
	TER(%)	BER(%)	TER(%)	BER(%)	TER(%)	BER(%)
256	7.97	0	7.98	0	7.92	0
512	8.47	0	8.22	0	7.84	0
1024	8.43	0	8.12	0	7.84	0
2048	8.36	28.34	8.93	0	7.75	0
4096	8.25	29.12	8.46	26.17	8.60	0

**Table 4**

TER and BER for robustness against fine-tuning for Uchida et al.'s algorithm [75].

Embedded layer	Payload (bit)	TER (%)	TER after attack (%)	BER after attack (%)
Conv 2	256	7.97	8.11	0
	1024	7.98	7.58	0
Conv 3	256	8.12	8.17	0
	1024	8.12	8.17	0
Conv 4	256	7.75	7.90	0
	4096	8.60	8.26	0

**Table 5**

TER and BER for robustness against parameter pruning for Uchida et al.'s algorithm [75].

Embedded layer	Payload (bit)	$p$	TER (%)	TER after attack (%)	BER after attack (%)
Conv 2	256	10%	7.97	10.23	0
	1024	10%	7.98	8.49	0
Conv 3	256	10%	8.22	8.80	0
	1024	10%	8.22	8.80	0
Conv 4	256	20%	7.75	8.73	0
	4096	10%	8.60	9.69	0

as follows:

$$E_{\text{ST-DM}}(\mathbf{w}) = E_0(\mathbf{w}) + \lambda E_{\text{ST-DM}}(\mathbf{w}) \quad (6)$$

with the new regularization term  $E_{\text{ST-DM}}$  given by:

$$E_{\text{ST-DM}}(\mathbf{w}) = - \sum_{j=1}^l (b_j \log(y_j) + (1 - b_j) \log(1 - y_j)), \quad (7)$$

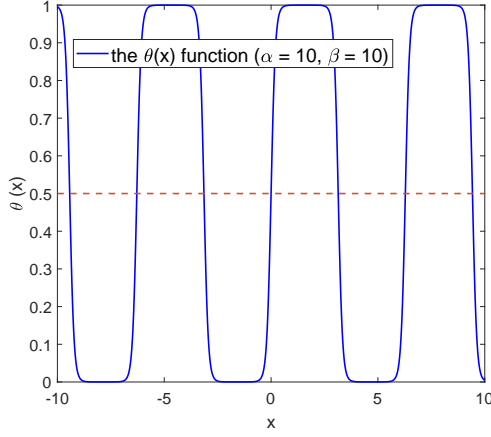
where  $\mathbf{y}$  is obtained by approximating the decoding function employed by ST-DM. To see how, let us consider two uniform interleaved quantizers  $\mathcal{Q}_0$  and  $\mathcal{Q}_1$  associated, respectively, to bit 0 and 1, and defined by the following codebooks

$$\mathcal{U}_0 = \{k\Delta, k \in \mathbb{Z}\} \quad (8)$$

$$\mathcal{U}_1 = \{k\Delta + \Delta/2, k \in \mathbb{Z}\}. \quad (9)$$

where  $\Delta$  is the quantization step. Given a watermarked sample  $w_m$ , Dither-Modulation (DM) decoding works by looking for the entry in  $\mathcal{U}_0 \cup \mathcal{U}_1$  closest to  $w_m$  and see if such an entry belongs to  $\mathcal{U}_0$  or  $\mathcal{U}_1$ , which is equivalent to applying the following decoding function:

$$\phi_{DM}(w) = \arg \min_{b=0,1} (\min_{u_k \in \mathcal{U}_b} |w_m - u_k|). \quad (10)$$



**Figure 6:** Behavior of the function  $\theta(x)$  in [47].

DM watermarking follows directly from the above formulation, and is achieved by quantizing the host sample  $w$  either with  $\mathcal{Q}_0$  (to embed a 0 bit) or  $\mathcal{Q}_1$  (to embed a 1 bit). ST-DM watermarking is achieved by applying DN watermarking to the projection of a sequence of host samples on a spreading direction.

In [47], watermark decoding is achieved by applying  $\phi_{DM}$  to the projection of  $\mathbf{w}$  (defined as in the previous section) over the directions determined by the rows of the pseudo-random matrix  $\mathbf{X}$ , that is:

$$y_j = \phi_{DM} \left( \sum_i X_{ji} w_i \right). \quad (11)$$

Due to the non-continuous nature of  $\phi_{DM}$ , the direct use of  $\phi_{DM}$  in Eq. (7) would make it difficult the application of back-propagation to train the network, hence  $\phi_{DM}$  is approximated with a smooth function  $\theta()$ , defined as:

$$\theta(x) = \frac{e^{\alpha \sin \beta x}}{1 + e^{\alpha \sin \beta x}}, \quad (12)$$

where  $\alpha$  determines the smoothness of  $\theta$  and  $\beta$  its periodicity, directly related to the quantization step defining the ST-DM watermarking algorithm. An example of the function  $\theta(x)$  for  $\alpha = \beta = 10$  is shown in Fig. 6. Finally, the  $y_j$ 's in Eq. (7) are computed as  $y_j = \theta(\sum_i X_{ji} w_i)$ , and watermark extraction is carried out as follows:

$$\hat{b}_j = \begin{cases} 1 & \theta(\sum_i X_{ji} w_i) \geq 0.5, \\ 0 & \text{otherwise.} \end{cases} \quad (13)$$

The results of some experiments showcasing the performance of ST-DM DNN watermarking are reported in Tables 6 and 7 obtained by using the same setting as in Tables 4 and 5, with  $\alpha = \beta = 10$ . Generally speaking, ST-DM watermarking allows a larger capacity than conventional spread spectrum, with a similar level of robustness.

### 4.3. DeepMarks.

In [15], Chen et al. exploit Uchida's algorithm to implement a traitor tracing watermarking system with anti-collision

**Table 6**

TER and BER for different convolutional layers with various payloads for Li et al.'s algorithm [47].

Payloads (bit)	Conv 2		Conv 3		Conv 4	
	TER(%)	BER(%)	TER(%)	BER(%)	TER(%)	BER(%)
256	8.20	0	8.15	0	8.20	0
512	7.75	0	8.07	0	8.57	0
1024	7.63	0	8.12	0	7.84	0
1200	7.96	0	8.31	0	8.06	0
2048	8.36	5	7.85	0	8.03	0
2400	8.82	15.08	8.22	0	8.31	0
4096	8.35	23.14	8.35	13.84	7.64	0
4800	8.29	24.33	8.26	14.06	8.65	0

**Table 7**

TER and BER for robustness against fine-tuning for Li et al.'s algorithm [47].

Embedded layer	Payload (bit)	TER (%)	TER after attack (%)	BER after attack (%)
Conv 2	256	8.20	8.02	0
	1024	7.79	7.65	0
Conv 3	256	8.15	8.05	0
	4096	8.20	7.93	0
Conv 4	256	8.20	7.93	0
	4096	7.64	7.43	0

capabilities. In a traitor tracing application, a different code is assigned to each different user of the network. Then, a watermark bearing the user's code is embedded within the network prior to its release to the user. If a non-authorized copy of the model is found, the content of the watermark allows to identify the user who illegally distributed it. To avoid that a subset of the users form a coalition, to produce a new model where all the watermarks are *mixed* together hence making it impossible to identify the users who redistributed the model illegally, Anti Collision Codes (ACC) may be used to identify the users [73].

In DeepMarks, the watermark bits are generated by using a  $(v, k, 1)$ -Balanced Incomplete Block Design (BIBD) code, which can be represented by its corresponding incidence matrix  $C_{v \times n}$  where  $n = (v^2 - v)/(k^2 - k)$  and each elements of  $C_{v \times n}$ :

$$c_{ij} = \begin{cases} 1 & \text{if } i - \text{th value occurs in } j - \text{th block} \\ 0 & \text{otherwise.} \end{cases} \quad (14)$$

By letting the bit complement of the columns of the incidence matrix  $C_{v \times n}$  represent the code-vectors, the resulting  $(v, k, 1)$ -BIBD code can identify coalitions of up to  $(k - 1)$  members chosen among  $n$  users [73].

### 4.4. Tartaglione et al. [27]

As an example of a static zero-bit watermarking system, we describe the method recently proposed by Tartaglione et al. in [27]. By leveraging on the superior robustness of zero-bit watermarking systems, and by adopting a properly designed loss function explicitly thought to increase the robustness of the watermark, such a system achieves a remarkable robustness against fine tuning and weights quantisation.

In contrast to the schemes described so far, in this case the watermarked coefficients are set before the training procedure starts and are not modified during training. To make the watermarked weights undistinguishable from the others (for security reasons), and by assuming that the weights of the to-be-watermarked architectures are initialised following a gaussian distributions ( $\mathbf{w} \sim \mathcal{N}(\mu, \sigma)$ ), the watermarked weights are generated as follows. Let  $b_i \in [0, 1]$  be the watermark coefficient associated to the  $j$ -th weight  $w_j$  (the choice of the weights carrying the watermark is made based on a pseudorandom number generator distributing the watermark across the entire model), the watermarked weights are generated as:

$$w_j = 2\sigma(b_i - 0.5) + \mu \quad \forall w_j \in \mathbf{w}_b, \quad (15)$$

where  $\mathbf{w}_b$  is a vector with all the weights associated to the watermark element  $b$ . To make sure that the watermarked weights are not modified during the training process, the watermarked weights are labelled and left unchanged during the gradient-based update rule, that is:

$$w_j = w_j - [1 - \text{ind}(w_j)]\eta \frac{\partial L}{\partial w_j} \quad (16)$$

where  $\eta$  is the learning rate and  $\text{ind}(w_j)$  is the indicator function identifying the watermarked weights:

$$\text{ind}(w_j) = \begin{cases} 1 & \text{if } w_j \in \mathbf{w}_b \\ 0 & \text{otherwise.} \end{cases} \quad (17)$$

Watermark detection is achieved by inverting equation of Eq. (15):

$$\hat{b}_i = \frac{w_j - \mu}{2\sigma} + 0.5 \quad (18)$$

Robustness against fine tuning is obtained by adding a term to the training loss function forcing the watermarked network to be highly sensitive to even small changes of the watermarked weights. In this way, during fine tuning, such weights are not modified (or modified only slightly) hence limiting the effect of fine tuning on the watermark. To do so,  $R$  perturbed versions of the DNN are generated. The weights of such networks are perturbed by adding to them a noise term generated according to a normal distribution:

$$w_j^r = w_j + \text{ind}(w_j)\Delta w_j^r, \quad (19)$$

where  $w_j^r$  ( $r = 1 \dots R$ ) are the weights of the perturbed networks and  $\Delta w_j^r$  is the noise term added to perturb them. The gradient of the loss function of the perturbed networks with respect to the perturbed weights is then computed and averaged across all the networks:

$$g_i = \frac{1}{R} \sum_{r=1}^R \frac{\partial L}{\partial w_j^r}. \quad (20)$$

Finally, the to-be-watermarked network is trained in such a way to maximise the sensitivity of the model to the watermarked weights. Specifically, the update rule of the DNN

**Table 8**

Performances of the watermarking system proposed in [27].  $R = 0$  refers to a network trained without enforcing watermark robustness.

Dataset	Architecture	Epochs	R	TER(%)
MNIST	LeNet5	100	0	0.78
			1	0.83
			4	0.82
			16	0.85
CIFAR-10	ResNet-32	350	0	7.14
			2	7.08
			4	7.29

**Table 9**

Fine-tuning attack on LeNet5 trained with different  $R$  values according to the method described in [27].

R values	Epochs	Person correlation(%)
0	10	99.986
	30	99.951
	50	99.900
4	10	99.999
	30	99.998
	50	99.997
16	10	99.999
	30	99.999
	50	99.998

weights is defined by:

$$w_j = w_j - [1 - \text{ind}(w_j)] \left[ \eta \frac{\partial L}{\partial w_j} - \gamma g_j \right], \quad (21)$$

where  $\gamma > 0$  is a properly selected hyper-parameter.

An excerpt of the experimental results reported in [27] is given in Table 8, where 0.4% of the model weights are used for conveying the watermark. According to the results, the influence of the watermark on the TER is negligible. The Pearson correlation between the original and extracted watermark is used to measure the robustness against fine tuning. As reported in Table 9, this method shows good robustness against fine-tuning.

A summary of the static watermarking algorithms described so far is given in Table 10, including their classification according to the taxonomy introduced in Sect. 2 and their main properties.

#### 4.5. Other works

Many other static watermarking schemes have been developed starting from the basic scheme by Uchida et al. [75].

Wang et al. [77], for instance, propose to embed the watermark into early converging weights by bringing in an additional independent neural network to embed and extract the watermark.

In [78], the undetectability of Uchida et al's watermark is studied, showing that the existence of the watermark can be easily detected by analyzing the statistical distribution of the weights. A solution to solve this problem is proposed in

**Table 10**  
Summary of the main static watermarking algorithms described in Sect. 4.

Algorithm	White/Black-box	Multi/ Zero-bit	Methodology	Key generation	Payload (bits)	Robustness and Security
Uchida et al. [75]	White-box	Multi-bit	A regularization term is added to the loss function to embed the watermark into the model weights according to the SS principle.	SS matrix drawn from a standard normal distribution.	conv2:1024; conv3:2048; conv4:4096	Moderate robustness against fine-tuning and pruning
Li et al. [47]	White-box	Multi-bit	As Uchida's scheme with ST-DM-like regularization term.	SS matrix drawn from a standard normal distribution.	conv2:1200; conv3:2400; conv4:4800	Moderate robustness against fine-tuning and pruning
DeepMarks [15]	White-box	Multi-bit	As Uchida's scheme with anti-collusion codebooks	Normally distributed SS matrix, plus anti-collusion codebooks	Basically the same as [75]	Moderate robustness against fine-tuning and pruning, Collusion attack
Tartaglione et al. [27]	White-box	Zero-bit	The watermarked weights are frozen during the training procedure. The loss function is designed so to maximize the sensitivity of the network to changes of the watermarked weights.	A pseudorandom number generator is used to select the weights to be watermarked.	Zero-bit	Good robustness against fine-tuning and weights quantization

[80], by relying on a generative adversarial network architecture (GAN). The watermarked model acts as the generator, while the discriminator is in charge of detecting the watermarked weights. The discriminator provides its feedback to the generator, which in turns is encouraged to embed the watermark in such a way that the watermarked weights are statistically similar to the non-watermarked ones. A further development of the scheme described in [80] (called RIGA - **Robust white-box GAN**) is presented in [79] together with a detailed theoretical and experimental analysis.

Another interesting improvement of [75] has been proposed in [19], based on the observation that the adoption of the Adam optimiser introduces a dramatic variation on the histogram distribution of the weights after watermarking, which can be easily detected by the adversaries. To solve this problem, the authors propose to use an orthonormal projection matrix, and to include in the projection also the gradients of the weights. Then the Adam optimiser is run on the projected weights using the projected gradients.

Kuribayashi et al. [44], propose a method that embeds the watermark into the selected weights via fine-tuning. After sampling the weights from the fully-connected layer, the original weights are substituted by the watermarked sampled weights by means of a quantization-based method (such as DM-QIM). To limit the impact that the watermark has on the accuracy of the network, both the fully-connected layer and the convolutional blocks ahead of the watermarked fully-connected layer are fine-tuned.

The DeepMarks scheme has been extended in [13] for protecting the rights of devices providers, and in [12] for IPR protection of Automatic Speech Recognition (ASR) systems.

## 5. Dynamic Watermarking

As we already noted in Section 2.2.2, with dynamic methods, the watermark is associated to the behaviour of the DNN in correspondence to a set of properly selected inputs. If the inputs activating the watermark are kept secret, their role is

analogous to a watermarking key. For this reason, in the following we will refer to them as key-inputs, or key-images in the case of DNN dedicated to image analysis. Due to their capability to *trigger* the desired dynamic watermarking behaviour, the key-inputs are sometimes referred to as watermark triggers. The key-inputs and the corresponding labels form a so-called *key input-label pair*.

The choice of the inputs activating the behaviour associated to the watermark is a crucial one. An important distinction can be made between systems wherein the key inputs are *entangled* with the task the network is intended to solve and *non-entangled* ones [37]. In the former case, the key-inputs are chosen within the class of inputs the network has been designed to work on. In the latter case, they correspond to outlier inputs that are not expected to be fed to the model in normal operative conditions. The use of entangled key inputs, forces the model to learn features which are jointly used to analyse both the normal and the key inputs. In this way, it is more difficult for an adversary to separate the watermark and the training task, hence resulting in a more robust watermark. On the contrary, in the non-entangled case, the impact of the watermark on model accuracy is lower, since the network may learn distinct features for watermark encoding and the primary task it is intended for, but, arguably, the resulting watermark will be weaker and easily removed by fine tuning.

Depending on the network layer whose behaviour is modified by the watermark presence, we can classify dynamic watermarking methods in two main classes, *activation-map* watermarking and *DNN-output* watermarking. In the former case, the behaviour induced by the watermark is observed at the intermediate layers of the network, in the latter, the watermark presence can be revealed by observing only the output of the network. Since the activation maps are internal parameters of the DNN models, activation-map methods belong to the class of white-box watermarking. On the contrary, DNN-output algorithms belong to the black-box category.

### 5.1. White-box dynamic watermarking of activation maps: the DeepSigns algorithm [63]

As an example of white-box dynamic watermarking based on activation maps, we describe the DeepSigns method proposed in [63]. As a matter of fact, DeepSigns can be applied in both white-box and black-box setting, in this section we focus on the white box version.

In DeepSigns, the behaviour of the watermarked activation map is defined by forcing the distribution of the map to follow a Gaussian Mixture Model (GMM) for which the mean values of the Gaussian probability density functions (pdfs) satisfy certain conditions which in turn determine the embedded bits. The desired behaviour is enforced by adding two additional terms to the loss function used for training:

$$L_{DS} = L_0 + \lambda_1 L_1 + \lambda_2 L_2, \quad (22)$$

where  $L_0$  is the original loss function used to train the non-watermarked models, and  $\lambda_1$  and  $\lambda_2$  are weights used to balance the importance of the two additional terms of the loss.

The goal of the term  $L_1$  is to let the distribution of the activation map hosting the watermark to be as close as possible to the desired distribution. Different Gaussian models are enforced for inputs belonging to different input classes, so to make sure that the network retains its classification capabilities. To be specific the term  $L_1$  is defined as:

$$L_1 = \sum_{i \in T} \|\mu_i^i - f_i^i(x, \theta)\|_2^2 - \sum_{i \in T, j \notin T} \|\mu_i^i - \mu_j^j\|_2^2, \quad (23)$$

where  $f_i^i(x, \theta)$  is the activation map corresponding to the input sample  $x$  belonging to class  $i$  at the  $l^{th}$  layer,  $T$  is the set of target Gaussian classes selected to carry the watermark, and  $\mu_i^i$  denotes the mean value of the  $i^{th}$  Gaussian distribution at layer  $l$ . In the following, we indicate with  $s$  the number of Gaussian models (input classes) chosen to host the watermark. The goal of the second term in Eq. (23) is to ensure that the Gaussian models associated to different classes are far apart, hence resulting in better classification performance.

To understand the form of the term  $L_2$ , let us consider the procedure whereby the watermark is read from the activation map. Let  $b^{s \times N}$  be a matrix with the watermark bits (the  $i^{th}$  row of the matrix contains the bits embedded in the pdf of class  $i$ ) and let  $\mu_l^{s \times M}$  represent the mean values of the corresponding Gaussian distributions. We first multiply  $\mu_l^{s \times M}$  by a spreading matrix  $\mathbf{A}$  whose role is analogous to that of  $\mathbf{X}$  in Uchida et al.'s algorithm, then we pass the results through a sigmoid function. The watermark bits are computed by comparing the output of the sigmoid function with a threshold equal to 0.5. In formulas:

$$G_\sigma^{s \times N} = \sigma(\mu_l^{s \times M} \cdot A^{M \times N})$$

$$\hat{b}^{s \times N} = \begin{cases} 1 & G_\sigma^{s \times N} \geq 0.5, \\ 0 & \text{otherwise.} \end{cases} \quad (24)$$

where  $M$  is the size of the activation map in the selected layer, that is the number of nodes of the fully-connected layer and  $N$  indicates the length of the watermark sequence embedded within each Gaussian model.

The loss term  $L_2$  is built in such a way to guarantee that the bits extracted as detailed above correspond to the correct bits. Such a goal is reached by letting  $L_2$  correspond to the cross-entropy between the desired bit matrix  $b^{s \times N}$  and  $G^{s \times N}$ :

$$L_2 = - \sum_{j=1}^N \sum_{k=1}^s (b^{kj} \ln(G_\sigma^{kj}) + (1 - b^{kj}) \ln(1 - G_\sigma^{kj})). \quad (25)$$

The term  $L_2$  is computed only on the key images, since the watermark can be read only when the network is fed with such images. In [63], the key images are chosen as a subset of the input training data belonging to the selected classes (entangled watermark triggers). For example, for a network trained on the MNIST dataset, the selected class to carry watermark corresponds to a specific digit, e.g. the zero digit. Thus, the key images are a subset of randomly chosen images of the zero-digit class.

The experimental results reported in [63], prove that the DeepSigns watermark provides good performances with regard to fidelity and robustness. In contrast, the watermark payload is quite limited as shown in Table 11. To ensure a zero BER, DeepSigns can allow up to 64 bits for MNIST and up to 128 bits for CIFAR 10-CNN and CIFAR 10-WRN.

### 5.2. Black-box dynamic watermarking

With black-box dynamic watermarking, the watermark is encoded in the relationship between a selected set of key inputs and the corresponding outputs. For such a relationship to be able to characterise the watermark and to avoid interfering with the intended behaviour of the model, it is necessary that the key inputs are carefully chosen.

#### 5.2.1. Watermarking and DNN backdoors

Generally speaking, teaching the network to behave in a *special* way in correspondence to a small set of selected inputs, while working as usual on the other inputs, is analogous to embedding a backdoor within the network, hence drawing a strong relationship between dynamic DNN watermarking and DNN backdoors [1]. In a backdoor attack [49], the attacker corrupts the training phase to induce a classification error, or any other erroneous behaviour, at test time. Test time errors, however, only occur in the presence of a triggering event corresponding to a properly crafted input. In this way, the backdoored network continues working as expected for regular inputs, and the malicious behaviour is activated only when the attacker decides to do so by feeding the network with a triggering input. With black-box dynamic watermarking, the activation of the backdoor does not correspond to a misbehaviour of the network, on the contrary it is instrumental for revealing the presence of the watermark and read its content.

Black-box dynamic watermarking methods can be clas-

**Table 11**

Performance achieved by DeepSigns on different marked models and datasets [63]. 64C3(1) indicates a convolutional layer with 64 output channels and  $3 \times 3$  filters applied with a stride of 1, MP2(1) denotes a max-pooling layer over regions of size  $2 \times 2$  and stride of 1, and 512FC is a fully-connected layer with 512 output neurons. In all cases BER = 0.

DNN Model Type	DNN Model Architecture	Dataset	Baseline TER	TER of Marked Model	Payload
MLP (Multi-Layer Perceptron)	784-512FC-512FC-10FC	MINST	1.46%	1.87%	4
CNN	$3 \times 32 \times 32$ -32C3(1)-32C3(1)-MP2(1)-64C3(1)-64C3(1)-MP2(1)-512FC-10FC	CIFAR10	21.53%	19.3%	4
WideResNet	Please refer to [84]	CIFAR10	8.58%	7.98%	128

sified according to the way the key-images<sup>7</sup> (or trigger images if we adopt a backdoor perspective) are chosen. In a first category of methods, the key images are chosen within a set of existing images, either belonging to the class of images the network is supposed to classify (entangled key-images) or not. In the following we will refer to this kind of key images as *natural key images*. By interpreting watermarking as backdoor injection, we can say that in this class of methods, the events triggering the watermark behaviour are the images themselves. In a second case, the key-images are hand-crafted so to satisfy certain conditions. From a backdoor perspective, this corresponds to embed within the key images a specific triggering pattern, that may assume the form of a hidden signal or even a visible pattern.

In the following, we describe some different black-box watermarking methods characterised by different ways of choosing the key-images. With few exceptions (see [16]), most black-box schemes belong to the class of zero-bit watermarking.

### 5.2.2. Natural key images

This branch of algorithms choose the key images without manipulating them. For this kind of methods, it's vital to find a proper way to identify a set of key images for which meeting the requirements listed in Section 3 is not too difficult. Some example of methods belonging to this category among those proposed so far are explicit described in the following.

**Yossi et al. [1]** were the first to suggest the use of backdoors for DNN watermarking. Specifically, a set of non-entangled key images is first chosen, then the labels associated to the key-images are sampled randomly from all the input classes excluding original labels predicted by the non-watermarked network. Two approaches are investigated for watermarking. According to the first approach, the network is first trained without key images, then a second phase of training is applied by using the key images as well. The second approach consists in training the network from scratch by

<sup>7</sup>In this section we always refer to key-images since most works proposed so far focus on the watermarking of DNN with images inputs.

**Table 12**

Classification accuracy of [1] on STL-10 dataset and the key images, after fine-tuning from either CIFAR-10 or CIFAR-100 classifiers.

	Test set accuracy (%)	Key image-label pairs accuracy (%)
CIFAR10 → STL10	81.87	72
CIFAR100 → STL10	77.3	62.0

also including the key images. To detect the presence of the watermark (the methods described in [1] is a zero-bit algorithm), the labels predicted on the key-images are compared with the desired one: for non watermarked models the agreement will generally be very low, while for the watermarked models most of the key-images will be assigned the correct label. The detection threshold is set based on the false positive probability estimated by modelling the probability that a non-watermarked network fed with a key image outputs the correct key label with a binomial distribution.

The validity of the system has been demonstrated by applying it to a ResNet-18 model trained on CIFAR10, CIFAR 100 and ImageNet datasets. 100 non-entangled key images with an abstract content have been used with key labels randomly chosen from the labels of the training set. The results show that the watermark obtained in this way achieves a good robustness against a fine-tuning attack, as reported in Table 12. For the fine-tuning experiments, the models were fine-tuned for 20 epochs on the on STL-10 datasets [18], starting from well-trained watermarked models trained on CIFAR-10 and CIFAR-100 datasets. Although the key image-label pairs accuracy decreases, the presence of the watermark can still be revealed with good accuracy.

**Rouhani et al.** As mention above, two versions of the DeepSigns watermarking algorithm have been proposed in [63]: one operating at the activation map level, and one based on the DNN-output. The latter is a black-box zero-bit watermarking method for which the key image-label pairs are chosen among the images that are misclassified by the original unmarked model. First a non-watermarked model is trained,

then the model is fine-tuned on misclassified key image-label pairs. The retrained model should have exact prediction accuracy for the chosen key images. The final key image-label pairs are the intersection of the key inputs that are correctly predicted by the marked model and falsely predicted by the unmarked model, so to reduce the false positive probability to a minimum.

**DAWN [71]** The DAWN (Dynamic Adversarial Watermarking of Neural Networks) system introduced in [71] is explicitly designed to counter the surrogate model attack described in Sect. 3.2. The goal of DAWN is to force the adversaries to learn the key image-label pairs association while training the surrogate model. Unlike other watermarking schemes, DAWN does not work when the to-be-protected model is trained. On the contrary it deploys an additional component within the API whereby the model is accessible by the users. More specifically, by referring to the overall block diagram shown in Figure 7, when the API receives the queries  $D_A$  from a client (who might be a latent adversary  $\mathcal{A}$ ), DAWN randomly selects a bunch of inputs  $x_n$  as key images and replies to them with incorrect predictions  $B(x_n) \neq F_v(x_n)$ , where  $F_v(x)$  is the original prediction of the protected model. When  $\mathcal{A}$  uses the inputs  $D_A$  and the corresponding predictions returned by DAWN to train a surrogate model  $F_A$ , the model learns the key image-label pairs hence embedding the watermark into the trained model.

More in details, when the DAWN module receives the queries from a user, it marks a fraction  $r_w$  of the inputs ( $r_w \times |D_A|$ ), chosen as follows, as key images:

$$M(x) = \begin{cases} 1 & \text{if } \text{HMAC}(K_w, x)[0, 127] < r_w \times 2^{128}, \\ 0 & \text{otherwise} \end{cases} \quad (26)$$

where  $M(x) = 1$  indicates that the query input  $x$  is marked as a key image,  $K_w$  is a model-specific secret key and SHA-256 is the hash function used to compute  $\text{HMAC}(K_w, x)$ . After receiving the predicted labels from the DNN model, DAWN scrambles the input images and their labels by means of a pseudo-random permutation fed with a secret key  $K_\pi$ :

$$B(x) = \pi(K_\pi, F_v(x)), \quad (27)$$

where  $F_v(x)$  indicate the correct predictions, and  $\pi$  is a random shuffling function. Later on, the marked images and their permuted labels are used by DAWN as key image-label pairs.

To detect if a given watermark  $(T, \hat{B}(T))$  is embedded in a suspect model  $F'$ , the detector computes the fraction of queries for which  $\hat{B}(T)$  and  $F'(T)$  do not match:

$$L(T, \hat{B}(T), F') = \frac{1}{|T|} \sum_{x \in T} (F'(x) \neq \hat{B}(x)) \quad (28)$$

if  $L(T, \hat{B}(T), F')$  is lower than a threshold  $e$ , the watermark presence is detected. Given that each user receives a fraction  $r_w$  of incorrect predictions, to avoid that DAWN impairs the performance of the protected model, the value of  $r_w$  must

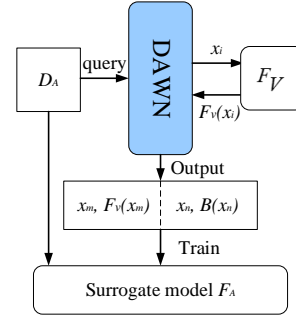


Figure 7: Overview of DAWN [71].

Table 13

DAWN accuracy: test (*test*) and watermark (*wm*) accuracy of surrogate models  $F_A$ .

Model	Best $Acc_{wm}$			Best $Acc_{test}$		
	<i>wm</i>	<i>test</i>	epochs	<i>wm</i>	<i>test</i>	epochs
MNIST-3L	99%	89%	210	98%	96%	290
MNIST-5L	99%	88%	215	99%	94%	365
GTSRB-5L	98%	89%	105	98%	90%	200
CIFAR10-9L	93%	78%	110	92%	79%	105

Table 14

Effectiveness of DAWN against PRADA surrogate model attack. Baseline gives the test accuracy  $Acc_{test}$  of the victim  $F_v$  and surrogate model  $F_A$  trained without DAWN.  $F_A$  with DAWN provides  $Acc_{test}$  and watermark accuracy  $Acc_{wm}$  of  $F_A$  when DAWN protects  $F_v$  from PRADA attack.

Model	Baseline $Acc_{test}$		$F_A$ with DAWN	
	$F_v$	$F_A$	$Acc_{test}$	$Acc_{wm}$
MNIST-5L	98.71%	95%	78.93%	100.00%
GTSRB-5L	91.50%	61.00%	61.43%	98.23%
CIFAR10-9L	84.53%	60.03%	60.95%	71.17%

be kept as small as possible, yet large enough to make the watermark detection reliable enough. In [71], the rate  $r_w$  is set to 0.5% and 250 images are selected as key-images.

Several experiments on different kinds of surrogate models are presented in [71] to evaluate the effectiveness of DAWN. An excerpt of the results are reported in Table 13, where  $Acc_{test}$  and  $Acc_{wm}$  represent the *test accuracy* and *watermark accuracy* of the surrogate model. The training epochs reaching best  $Acc_{wm}$  (optimal for the detector) or best  $Acc_{test}$  (optimal for  $\mathcal{A}$ ) are also reported in Table 13. As it can be seen, whenever the surrogate model achieves good performance on the test set, the detection accuracy of the watermark is also good. DAWN can also resist the surrogate attack called PRADA [38] (see Sect. 6 for further details on PRADA attack) as shown in Table 14.

### 5.2.3. Hand-crafted key images

For the methods belonging to this class, the key images are handcrafted in such a way that forcing specific outputs when they are fed into the network does not compromise the good behaviour of the model. It should also be highly unlikely that a non-watermark network provides the correct

output by chance (low false alarm probability). This means that the output of the network in correspondence of the key images is not linked too tightly to the *normal* behaviour of the network.

**Merrer et al. [45]** The method proposed in [45] by Merrer et al. embeds the watermark by fine-tuning a pre-trained model so that the boundary of the classification region assumes a desired shape. More specifically, the desired shape is obtained by *stitching* it around a set of inputs corresponding to a set of adversarial examples computed on the pre-trained model.

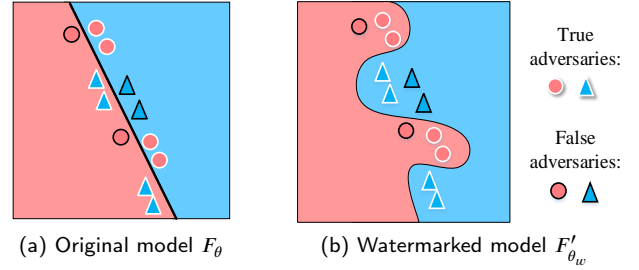
To fix the ideas, let us assume that we aim at watermarking a network implementing a binary classifier. Let  $F_\theta$  be the pre-trained model. To create the watermarked model  $F'_{\theta_w}$ , the pretrained model is fine-tuned so to change the way it classifies some selected key-images close to the decision boundary. In order to preserve the accuracy of the model, the key-images correspond to a set of adversarial examples for which  $F_\theta$  makes a wrong decision<sup>8</sup>. To start with,  $F_\theta$  is attacked by generating a set of adversarially perturbed images. In [45], the IFGSM algorithm [32] is used, however other adversarial attacks could be used as well. Given a set of correctly classified images,  $\{x_1 \dots x_n\}$  the corresponding adversarial examples are generated as:

$$x_i^* = x_i + \varepsilon \cdot \text{sign}(\nabla_{F_\theta}(x_i)), \quad (29)$$

where  $\nabla$  indicates the gradient of  $F$  with respect to the input  $x_i$  and  $\varepsilon$  determines the strength of the attack. When  $\varepsilon$  is chosen properly the adversarial attack succeeds ( $F_\theta$  classifies the adversarial images wrongly) and the corresponding images are referred to as *true adversaries*. For some of the images  $\varepsilon$  is chosen in such a way that the attack fails. Such images, for which  $F_\theta$  still provides a correct classification, are called *false adversaries*. The true and false adversaries represent the key images of the watermark. As a next step,  $F_\theta$  is fine-tuned on the key images, until they are all classified correctly. In so doing the boundary of the decision region is stitched around the key-images as depicted in Figure 8.

In the watermark detection phase, the model is queried with the key images and a statistical null-hypothesis test is carried out. A non-watermarked model will produce the two possible outputs of the classification with the same probability (given the proximity of the adversarial images to the decision boundary and given that the number of true and false images is the same), while a watermarked model is going to classify all (most) of them correctly. Assuming a binomial distribution for the classification errors produced by the non-watermarked models on the key-images, the presence of the watermark is revealed when the number of misclassified key-images is lower than a threshold set by fixing the false alarm probability.

<sup>8</sup>Adversarial examples are slightly perturbed versions of the input images for which the network makes a wrong decision. As shown in several papers [70, 60], adversarial images are ubiquitously present in all deep architectures, so we can assume that it is always possible to generate them.

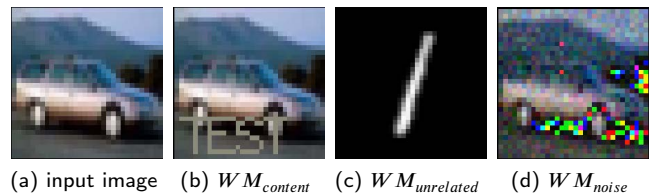


**Figure 8:** Stitching the boundary of the classification region around key images as proposed [45].

The results reported in [45] on three architectures CNN (Convolutional Neural Network), MLP (Multi-Layer Perceptron) and IRNN (Integral Recurrent Neural Network), in the context of MNIST digit recognition task, demonstrate the effectiveness of this approach, even if the parameters used for the generation of the adversarial examples must be carefully tuned.

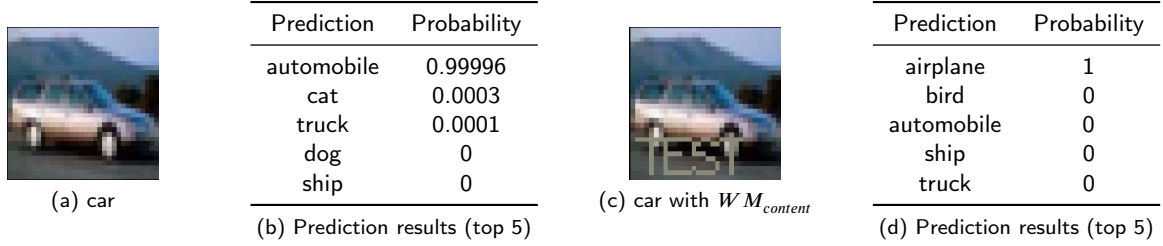
**Zhang et al. [87]** The method presented in [87] follows closely the watermarking through backdooring paradigm. The key-images are generated by superimposing to some of the training images a visible pattern (watermark triggering pattern), unrelated to the host image. The images with the pattern are then re-labeled by changing their original true class and used to train the watermarked vector to output the chosen label in the presence of the watermark triggering pattern. Three different ways of generating the triggering pattern are proposed, as exemplified in Fig. 9. To verify the presence of the watermark, the owner feeds the key-images into the DNN and verifies if the response matches with the desired key labels.

A snapshot of the performance achieved by Zhang et al.'s method is given in Figure 10 and Table 15 showing, respectively, the output of the watermarked model in correspondence of a key-image, and the accuracy of the original and the watermarked models on two standard tasks. The results of some experiments to measure the robustness of the watermark against model pruning and fine tuning are discussed in [87]. Although the three methods to generate the key-images achieve good performance also in the presence of pruning, the key images generated by noise addition are quite sensitive to fine-tuning.



**Figure 9:** Key images generated by the method in [87].





**Figure 10:** A case of watermark detection for the method by Zhang et al. [87].

**Table 15**

Testing accuracy of different tasks in [87].

(a) MNIST			
CleanModel	$WM_{content}$	$WM_{unrelated}$	$WM_{noise}$
99.28%	99.46%	99.43%	99.41%
(b) CIFAR-10			
CleanModel	$WM_{content}$	$WM_{unrelated}$	$WM_{noise}$
78.6%	78.41%	78.12%	78.49%

**Guo et al. [34]** In [34], the key images are also generated by adding to them a triggering pattern, however such a pattern is an invisible one and can be considered as a way to *sign* a subset of the images in the training set. The signed images are assigned a set of predefined (possibly random) labels. The DNN is first trained on non-signed images, then fine tuning is applied to teach the network to classify the signed images as desired. Due to the invisibility of the signature, a non-watermarked model will continue classifying the signed images as the pre-trained model, while a marked model will recognise the presence of the watermark and behave accordingly. As a possible instantiation of the above scheme, in [34] it is proposed to sign the key by adding the output of a pseudorandom bit sequence to random pixel locations. The amplitude of the signature is determined in such a way to be invisible but large enough to be detected by the network, namely:

$$x' = x + \alpha m, \quad (30)$$

where  $x'$  is the key-image obtained by signing  $x$ ,  $m$  is the additive signature and  $\alpha$  a weight determining its strength. The optimum value of  $\alpha$  is determined by performing a binary search between a minimum and a maximum value,  $\alpha_{min}$  and  $\alpha_{max}$ . According to the results shown in Table 16, Guo et al's method achieves a good performance regarding fidelity and low false positive rate.

An overview of the watermarking method proposed in [34] is given in Figure 11.

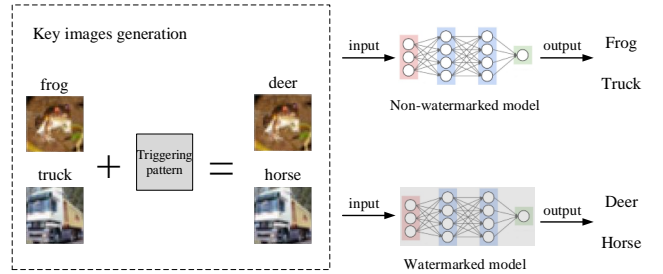
### 5.3. Dynamic watermarking of image-processing networks

All the DNN-output dynamic watermarking methods described so far focus on classification networks. More re-

**Table 16**

Classification accuracy of Guo et al.'s method [34] on different models and datasets. The classification results are obtained from regular training set ( $D^{train}$ ), test set ( $D^{est}$ ) and training set with key images ( $D_{am}^{train}$ ).

Dataset	Model	$D^{train}$	$D^{est}$	$D_{am}^{train}$
MNIST	LeNet	99.17	98.99	0.10
	LeNet <sup>WMK</sup>	98.41	98.48	98.38
CIFAR-10	VGG	99.97	93.07	0.0060
	VGG <sup>WMK</sup>	99.96	92.86	99.94
	ResNet	100	94.53	0.022
	ResNet <sup>WMK</sup>	99.99	94.25	99.98
	DenseNet	100	94.73	0.022
	DenseNet <sup>WMK</sup>	99.98	94.23	99.97



**Figure 11:** Overview of the method proposed by Guo et al. [34].

cently, researchers have applied the dynamic watermarking paradigm to image processing networks. This is the case, for instance, of GAN watermarking, denoising CNN, autoencoders and so on. In this case, the behaviour of the watermarked model is no longer represented by the labels output in correspondence of the key-images, but by the content of the images produced by the network (possibly, but not necessarily, in correspondence of key input images). Interestingly, this kind of dynamic watermarking is tightly related to classical image watermarking, since the DNN-watermark ultimately boils down to the embedding of a specific signature into the images produced by the network. Stated in another way, a watermarked DNN can be recognised by the fact that all the images it produces are watermarked<sup>9</sup>.

An example of dynamic watermarking applied to an im-

<sup>9</sup>Of course, it is also possible that the output of the DNN is watermarked only in correspondence of a predefined set of key input images.

age processing network is given in the following.

**Zhang et al. [86]** The basic idea put forward in [86] is to train the to-be-protected network (say  $F$ ) in such a way that all the images it produces contain a watermark. When a surrogate model  $SM$  is trained by feeding it with the images produced by  $F$ , the surrogate model imitates  $F$  and it implicitly learns to produce images containing the watermark, hence proving that  $SM$  was obtained by *copying* the behaviour of  $F$ . As a matter of fact, in the actual implementation described in [86], the watermark is not embedded by  $F$  in concomitance with the processing task it is supposed carry out. Rather, the watermark is embedded by a second network cascaded to  $F$ , implementing a conventional media watermarking task. In addition, a watermark recovery network is trained making sure that a null watermark is extracted from non-watermarked images and to ensure robustness against the watermark degradation introduced due to the unavoidable differences between the original and the surrogate models.

The overall pipeline of Zhang et al.'s method is depicted in Figure 12. At the output of the image processing network  $F$ , a watermark embedding subnetwork  $H$  is trained to embed the chosen watermark into the images produced by  $F$ . To ensure the invisibility of the watermark,  $H$  is trained in adversarial way, by making sure that the watermarked images can not be distinguished from the non-watermarked ones by a discriminator  $D$ . At the same time, a watermark extractor  $R$ , is trained in such way to extract the correct watermark from the images produced by  $H$  and a null watermark for the other images. Moreover (see bottom part of the Figure 12), a local surrogate model  $SM$  is built and the watermark extractor trained in such a way to be able to recover the watermark also from the images produced by the surrogate model.

For the experimental validation described in [86], the UNet [62] architecture was used to implement  $H$  and  $SM$ , while the extractor  $R$ , whose output has a different size than the input, was implemented by CEILNet [28]. Finally a PatchGAN [36] was adopted for the discriminator. As for the processing task, image deraining [85] and Chest X-ray image debone [81] were considered. With regard to the quality of the watermarked images the average PSNR was 39.98 and 47.89 for derain and debone respectively. The experiments also show a good capability to extract the watermark from the images produced by a surrogate model, with a normalised correlation coefficient close to 1. Such a capability is mostly due to the choice of training the watermark extractor also on images produced by the surrogate model.

A summary of the dynamic watermarking algorithms detailed so far is reported in Table 17, together with their main characteristics.

#### 5.4. Other works

Other dynamic watermarking algorithms have been, and continues to be, developed, in addition to those described above. Here give a brief overview of some of the latest al-

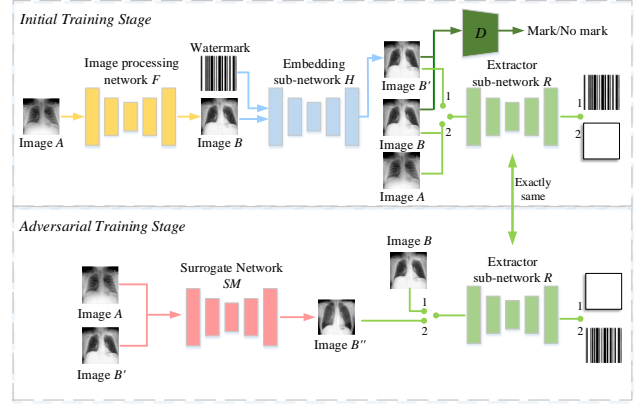


Figure 12: Overall pipeline of Zhang et al.'s method [86].

gorithms that have been proposed.

In [57], Namba et al. chose images from training set as key images and assign them key labels different from the original predictions. To reinforce the robustness over parameter pruning, they exponentially weight each parameter of the pre-trained model, employing the weighted parameters to retrain the target model with key image-label pairs.

Following the watermarking-through-backdooring paradigm, in [35, 46, 37, 50, 5] the key images are generated by attaching a triggering pattern to a subset of images chosen from the training set. To decrease the false positive rate, Guo et al. [35] deployed a differential evolution to optimize the position of either a visible [87] or an invisible pattern triggering pattern [34]. An approach similar to [87] was proposed in [50] where natural images are combined with a *logo* and the model is trained to predict them into a specific class. To keep the key images as close as possible to the original ones, an autoencoder is used whose discriminator is trained to distinguish between original images and key images. The triggering pattern used by Li et al. in [46] is a white-black pixels pattern converted from a binary sequence, where image pixels under white pixels are changed to a large positive value while black pixels are mapped into small negative value. To prevent an adversary from removing the trigger. Jia et al. [37] adopted a soft nearest neighbor loss to find the optimized location of the patched invisible pattern within the key images, whereby the key images are entangled with the regular training images and activate the same network neurons. In this way, attempting to remove the trigger is likely to result in a significant loss of performance. Atli et al. [5] expanded this class of approaches to federated learning.

The use of adversarial examples to generate the images is also adopted in [88], [52] and [16]. In particular, Lukas et al. [52] exploit a property called conferrability of adversarial examples, according to which adversarial examples are transferrable only from a source model to its surrogates, but not to independently trained models. Thanks to the exploitation of this property, Lukas et al. report very good performance on against surrogate model attacks.

Chen et al. proposed the first multi-bit watermarking al-

**Table 17**  
Summary of the dynamic watermarking algorithms described in Sections 5.1 through 5.3.

Algorithm	White/Black-box	Multi/ Zero-bit	Principle methods	Key generation	Payloads(bit)	Robustness and Security
DeepSigns [63] (Activation map)	White-box	Multi-bit	Embeds an arbitrary N-bit string into the probability density function of the activation maps	Input key images, normally distributed projection matrix	MNIST: 64 / CIFAR10: 128	Moderate robustness against fine-tuning and pruning
DeepSigns [63] (Output layer)	Black-box	Zero-bit	Train pairs of random selected images and labels as key image-label pairs.	Random key images alien to classification task	X	Against fine-tuning, pruning and watermark overwriting.
Yossi et al. [1]	Black-box	Zero-bit	Select random images to inject a backdoor into the target model	Random key images selected by the embedder	X	Model fine-tuning
Szyller et al. [71]	Black-box	Zero-bit	Deployed at the input and output of the target model's API, DAWN is an additional component that dynamically embeds watermarks in responses to queries made by an API client.	Compute HMAC( $K_w, x$ ) using SHA-256, where $K_w$ is the generated secret key and $x$ is an input of target model.	X	Regular surrogate model attack and surrogate attack PRADA [38]
Merrer et al. [45]	Black-box	Zero-bit	Leverage on adversarial attacks to tweak the decision boundary of the target model	Key image-label pairs are generated by crafting adversarial samples	X	Parameter pruning, Overwriting via adversarial fine-tuning
Zhang et al. [87]	Black-box	Zero-bit	Backdoor-based method with visible triggering patterns	Add visible patterns to original training images, or directly use unrelated images as key images.	X	Model fine-tuning, Parameter pruning, Model inversion attack
Guo et al. [34]	Black-box	Zero-bit	Backdoor-based method with invisible triggering signatures	Pseudorandom generation of the triggering signature and the location of signed pixels	X	Model fine-tuning
Zhang et al. (image processing networks) [86]	Black-box	Zero-bit	Use a sub-network to embed the watermark into the images produced by the network	X	X	Surrogate model attack

gorithm (called BlackMarks) applicable in a black-box scenario [16]. The purpose of BlackMarks is to design a model-dependent watermarking scheme that maps the class predictions to binary bits. The key images are created by deploying target adversarial attacks on the original training images.

As [86], Wu et al. [82] also consider the watermarking of image transformation networks wherein the inputs and outputs are both images. However, instead of splitting the transformation and the watermarking tasks, the watermark is embedded into the output images by the transformation network itself (an independent network to extract the watermark is presented as well).

Finally, Zhao et al. [89] considered the IPR of Graph Neural Network (GNN), in which an Erdos-Renyi random graph with random node feature vectors and labels is generated as a trigger key to train the to-be-protected GNN.

## 6. Specific Attacks against DNN Watermarking Algorithms

In the last column of Table 10 and Table 17, we listed some attacks that algorithms can resist to. However, most of them are operations commonly carried out on DNN, without the explicit goal of removing the watermark. This is the case, for instance of fine tuning and model pruning. Especially, with the rapidly growing attention attracted by DNN watermarking, researchers have started developing specific attacks deliberately thought to remove the watermark from a network, without impairing the network itself. In other

words, the security of the watermark against deliberate attacks is gaining attention with respect to the initial focus on robustness.

Wang et al. [78] and Shafieinejad et al. [64] independently proposed an attack that first detects the presence of a watermark within the network weights, then removes it<sup>10</sup>. Their approaches rely on the observation (also shared by [19]) that watermark embedding increases the variance of the weights thus allowing to distinguish a watermarked model from a non-watermarked one, and also to estimate the length of the watermark. In the following we briefly outline the attack described in [78] and the theoretical analysis justifying it.

According to Uchida et al's algorithm, watermark embedding is achieved by adding a regularization term to the loss function used for training, as stated in Eq. (2). The ultimate effect of such a term is to increase the projection of the weights onto the rows of the embedding matrix  $\mathbf{X}$  in correspondence of bits equal to 1 and decreasing it towards negative values, when the to-be-embedded bits are equal to 0. In both cases, watermark embedding aims at increasing the absolute value of the projection, namely  $|\sum_i X_{ji} w_i|$ . By observing that the rows of  $\mathbf{X}$  are independently distributed according to an  $N(0, 1)$  distribution, we can ignore the subscript  $j$  given that all the rows will share the same statistics, and rewrite the magnitude of the projection as  $|\sum_i x_i w_i|$ ,

<sup>10</sup>Focusing directly on the weights of the model this class of attacks applies specifically to static watermarking systems.

where  $x_i$  is a generic sequence of independent and identically distributed normal random variable with zero mean and unit variance. The magnitude of the projection can be upper bounded by using Cauchy-Schwarz inequality, yielding

$$\left| \sum_i x_i w_i \right|^2 \leq \sum_i x_i^2 \sum_i w_i^2 \approx N \sum_i w_i^2, \quad (31)$$

where  $N$  is the number of columns of the matrix  $\mathbf{X}$ , and where the last approximation is valid (by the law of large numbers) whenever  $N$  is large enough. To simplify the network, the average of  $w_i$  is usually set to 0, thus qualifying the last term in (31) as the variance of  $\mathbf{w}$ . It is clear from the above discussion<sup>11</sup> that the addition of the watermarking regularization terms causes an increase of the variance of the network weights. In [78] it is also shown that standard deviation of the weights scales approximately linearly with the watermark dimension of the watermark  $N$ , hence making it possible for the attacker to detect the presence of the watermark and estimate its length. Such a knowledge is then used to overwrite the existing watermark with a new one, this making the original watermark unreadable.

Selection of key images is a necessary step to implement any dynamic watermarking algorithm. Shafieinejad et al. [66] pointed out a possible flaw of key images selection. Their observations relies on the fact that, in order to limit the impact of the watermark on the effectiveness of the network, the key images are often chosen outside the kind of images the network has been designed to work on (non entangled key-images). As a consequence, when the network is fine-tuned with even a small of data, the watermark is easily removed. An attacker may also improve the effectiveness of a fine-tuning attack, by properly designing the fine-tuning process as suggested by Chen et al. in [17], where a learning rate schedule that favours watermark forgetting is used, together with elastic weight consolidation [41]. As an additional observation, we note that all watermarking algorithms designed following backdoor principles can be attacked by using one of the many existing backdoor defences (see for instance [76, 43, 14]), even without knowing the key images.

A major challenge of any watermarking algorithm, is surviving surrogate model attacks. In addition to the basic attack described in Sect. 3.2. Several refined, more powerful, surrogate model attacks have been developed. Such refined versions, usually try to optimise the two main steps each surrogate model attack consists of: hyperparameters selection (including the learning rate and the number of training epochs) and generation of the queries to be fed to the targeted model. In [38] Jutti et al. introduced a powerful surrogate model attack named PRADA (PROtecting Against DNN model stealing Attacks) consisting of several, so called, *duplication rounds*. Each round consists of three steps: firstly, the target model  $F$  is queried with inputs called *seed samples*; then, the surrogate model  $F_A$  is trained by exploiting the predictions provided by  $F$ ; at last, new synthetic query samples are crafted based on  $F_A$ , according to one

of two possible strategies: *Jacobian-based* adversarial examples and *Random* generation. Jacobian-based synthetic sample generation relies on adversarial examples crafted by using one of the many available methods to generate them (for instance I-FGSM [32] and MI-FGSM [25]), while with Random generation the new samples are obtained by perturbing the color channels of the input images.

## 7. Final remarks and suggestion for future research

The demand of methods for protecting the Intellectual Property Rights associated to DNNs has pushed researchers to develop a new class of algorithms to embed a watermark into DNN models. In a few years, several techniques have been proposed, sometimes based on naive arguments and sometimes by relying on solid theoretical bases. By looking at the methods proposed so far, it is evident that watermarking a DNN model presents some peculiarities that must be taken into account when trying to apply general multimedia watermarking principles to DNNs. For this reason, we have started our review by highlighting the main differences and similarities with classical multimedia watermarking, and introducing a new specific taxonomy explicitly thought to highlight the characteristics of the various DNN watermarking algorithms proposed so far. Then we have reviewed the main algorithms available for each watermarking category, pointing out the main advantages and drawbacks characterising the various approaches.

By the light of the properties and limits of the watermarking algorithms proposed so far, the most important challenges researchers are going to face with in the years can be outlined as follows.

1. Most of the solutions proposed so far employ a spread spectrum approach to develop a multi-bit watermarking scheme. The extension to multi-bit watermarking is mostly limited to static watermarking algorithms. Designing a high-capacity multibit watermarking algorithm with at least some robustness against the most common DNN manipulations has not been addressed properly yet. A closely related question regards the capacity of DNN watermarks. How many bits can be reliably hidden within a DNN model consisting of a certain number of parameters and thought to solve a given task? Is there a difference, on this respect, between static and dynamic schemes?
2. Robustness against fine tuning, model pruning and, even more, transfer learning, is one of the most difficult challenges researchers will need to face with. The great majority of the watermarks proposed so far, most noticeably multibit methods, are very weak against fine tuning and overwriting attacks. In addition, no scheme has been proven to be robust against retraining for transfer learning. While it is pretty obvious that channel coding may help to increase the robustness of DNN watermarking, the way channel coding should be incorporated within the embedding process

<sup>11</sup>A more detailed analysis is provided in [78].

during the training phase is not clear. Also unclear, it is the kind of channel codes that fits better the DNN-watermarking scenario.

3. Security against deliberate attacks is another area that requires more investigation. If DNN watermarking is going to be used in security critical applications, like IPR protection, the presence of an informed adversary explicitly aiming at watermark removal can not be ruled out, and will have to be taken into account before watermarked-based DNN protection is deployed in real world applications.
4. Dynamic watermarking offers a bunch of brand new opportunities that were not available in the multimedia case, however several questions need to be answered to clarify the potentialities of dynamic (vs static) watermarking, including: i) how many triggering inputs can we define without affecting the capability of the network to solve the problem it is designed for? ii) What is the impact of fine-tuning, retraining, pruning, on the behaviour of the network in correspondence to the watermark triggering inputs? iii) Is it preferable that the triggering inputs are chosen in the vicinity of standard inputs or should they be alien to the task the network is asked to solve?
5. Multimedia watermarking was deeply influenced by the development of a rigorous theoretical framework to cast the watermarking problem in. The discovery that watermarking could be modelled as a problem of channel coding with side information [24] led to the development of the powerful class of side-informed watermarking algorithms. Does a similar model apply to the DNN case? We believe that the development of a rigorous theory of DNN watermarking would be strongly beneficial, and would speed up the advances in the field.

Given the urgent need of suitable means to protect DNNs from misuse, we expect that the above challenges will be the subject to an intense research that will occupy the agenda of researchers for the years to come.

## References

- [1] Adi, Y., Baum, C., Cisse, M., Pinkas, B., Keshet, J., 2018. Turning your weakness into a strength: Watermarking deep neural networks by backdooring, in: 27th USENIX Security Symposium, pp. 1615–1631.
- [2] Adsumilli, C.B., Farias, M.C., Mitra, S.K., Carli, M., 2005. A robust error concealment technique using data hiding for image and video transmission over lossy channels. *IEEE Transactions on Circuits and Systems for Video Technology* 15, 1394–1406.
- [3] Arnold, M., Chen, X.M., Baum, P., Gries, U., Doerr, G., 2013. A phase-based audio watermarking system robust to acoustic path propagation. *IEEE Transactions on Information Forensics and Security* 9, 411–425.
- [4] Arnold, M.K., Schmucker, M., Wolthusen, S.D., 2003. *Techniques and applications of digital watermarking and content protection*. Artech House.
- [5] Atli, B.G., Xia, Y., Marchal, S., Asokan, N., 2020. Waffle: Watermarking in federated learning. *arXiv preprint arXiv:2008.07298*.
- [6] Barni, M., Bartolini, F., 2004. *Watermarking systems engineering: enabling digital assets security and other applications*. Crc Press.
- [7] Barni, M., Bartolini, F., Cappellini, V., Piva, A., 1998. A dct-domain system for robust image watermarking. *Signal processing* 66, 357–372.
- [8] Bassia, P., Pitas, I., Nikolaidis, N., 2001. Robust audio watermarking in the time domain. *IEEE Transactions on multimedia* 3, 232–241.
- [9] Celik, M.U., Lemma, A.N., Katzenbeisser, S., van der Veen, M., 2008. Lookup-table-based secure client-side embedding for spread-spectrum watermarks. *IEEE Transactions on Information Forensics and Security* 3, 475–487.
- [10] Chen, B., Wornell, G., 1998. Digital watermarking and information embedding using dither modulation, in: *IEEE Second Workshop on Multimedia Signal Processing*. doi:10.1109/MMSP.1998.738946.
- [11] Chen, B., Wornell, G., 2001. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Transactions on Information Theory* 47, 1423–1443. doi:10.1109/18.923725.
- [12] Chen, H., Darvish, B., Koushanfar, F., 2020. Specmark: A spectral watermarking framework for ip protection of speech recognition systems, in: *Proceedings of Interspeech 2020*, pp. 2312–2316.
- [13] Chen, H., Fu, C., Rouhani, B.D., Zhao, J., Koushanfar, F., 2019a. Deepattest: an end-to-end attestation framework for deep neural networks, in: *2019 ACM/IEEE 46th Annual International Symposium on Computer Architecture (ISCA)*, IEEE. pp. 487–498.
- [14] Chen, H., Fu, C., Zhao, J., Koushanfar, F., 2019b. DeepInspect: A Black-box Trojan Detection and Mitigation Framework for Deep Neural Networks, in: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 4658–4664.
- [15] Chen, H., Rouhani, B.D., Fu, C., Zhao, J., Koushanfar, F., 2019c. Deepmarks: A secure fingerprinting framework for digital rights management of deep learning models, in: *Proceedings of the 2019 International Conference on Multimedia Retrieval*, pp. 105–113.
- [16] Chen, H., Rouhani, B.D., Koushanfar, F., 2019d. Blackmarks: Black-box multibit watermarking for deep neural networks. *arXiv preprint arXiv:1904.00344*.
- [17] Chen, X., Wang, W., Ding, Y., Bender, C., Jia, R., Li, B., Song, D., 2019e. Leveraging unlabeled data for watermark removal of deep neural networks, in: *ICML workshop on Security and Privacy of Machine Learning*.
- [18] Coates, A., Ng, A., Lee, H., 2011. An analysis of single-layer networks in unsupervised feature learning, in: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223.
- [19] Cortiñas-Lorenzo, B., Pérez-González, F., 2020. Adam and the ants: On the influence of the optimization algorithm on the detectability of dnn watermarks. *Entropy* 22, 1379.
- [20] Costa, M., 1983. Writing on dirty paper. *IEEE Transactions on Information Theory* 29, 439–441. doi:10.1109/TIT.1983.1056659.
- [21] Cox, I., Miller, M., Bloom, J., Fridrich, J., Kalker, T., 2007. *Digital watermarking and steganography*. Morgan kaufmann.
- [22] Cox, I.J., Kilian, J., Leighton, F.T., Shamon, T., 1997. Secure spread spectrum watermarking for multimedia. *IEEE transactions on image processing* 6, 1673–1687.
- [23] Cox, I.J., Miller, M.L., Bloom, J.A., Honsinger, C., 2002. *Digital watermarking*. volume 53. Springer.
- [24] Cox, I.J., Miller, M.L., McKellips, A.L., 1999. Watermarking as communications with side information, in: *Proceedings of the IEEE*, pp. 1127–1141.
- [25] Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J., 2018. Boosting adversarial attacks with momentum, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193.
- [26] Eggers, J.J., Bauml, R., Tzschoppe, R., Girod, B., 2003. Scalar costa scheme for information embedding. *IEEE Transactions on Signal Processing* 51, 1003–1019. doi:10.1109/tsp.2003.809366.
- [27] Enzo, T., Marco, G., Davide, C., Marco, B., 2020. Delving in the loss landscape to embed robust watermarks into neural networks, in: *Pro-*

- ceedings of the 2020 International Conference on Pattern Recognition (ICPR).
- [28] Fan, Q., Yang, J., Hua, G., Chen, B., Wipf, D., 2017. A generic deep architecture for single image reflection removal and image smoothing, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3238–3247.
- [29] Furon, T., 2007. A constructive and unifying framework for zero-bit watermarking. *IEEE Transactions on Information Forensics and Security* 2, 149–163.
- [30] Ganic, E., Eskicioglu, A.M., 2004. Robust dwt-svd domain image watermarking: embedding data in all frequencies, in: Proceedings of the 2004 Workshop on Multimedia and Security, pp. 166–174.
- [31] Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT press.
- [32] Goodfellow, I.J., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples, in: 2015 International Conference on Learning Representations (ICLR).
- [33] Gruhl, D., Lu, A., Bender, W., 1996. Echo hiding, in: International Workshop on Information Hiding, Springer. pp. 295–315.
- [34] Guo, J., Potkonjak, M., 2018. Watermarking deep neural networks for embedded systems, in: 2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), IEEE. pp. 1–8.
- [35] Guo, J., Potkonjak, M., 2019. Evolutionary trigger set generation for dnn black-box watermarking. arXiv preprint arXiv:1906.04411 .
- [36] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134.
- [37] Jia, H., Choquette-Choo, C.A., Papernot, N., 2021. Entangled watermarks as a defense against model extraction, in: 30th USENIX Security Symposium.
- [38] Juuti, M., Szyller, S., Marchal, S., Asokan, N., 2019. Prada: protecting against dnn model stealing attacks, in: 2019 IEEE European Symposium on Security and Privacy (EuroS&P), IEEE. pp. 512–527.
- [39] Kay, S.M., 1993. Fundamentals of statistical signal processing. Prentice Hall PTR.
- [40] Khaldi, K., Boudraa, A.O., 2012. Audio watermarking via emd. *IEEE transactions on audio, speech, and language processing* 21, 675–680.
- [41] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al., 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 114, 3521–3526.
- [42] Kirovski, D., Malvar, H., Yacobi, Y., 2002. Multimedia content screening using a dual watermarking and fingerprinting system, in: Proceedings of the tenth ACM international conference on Multimedia, pp. 372–381.
- [43] Kolouri, S., Saha, A., Pirsiavash, H., Hoffmann, H., 2020. Universal Litmus Patterns: Revealing Backdoor Attacks in CNNs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 301–310.
- [44] Kuribayashi, M., Tanaka, T., Funabiki, N., 2020. Deepwatermark: Embedding watermark into dnn model, in: 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE. pp. 1340–1346.
- [45] Le Merrer, E., Perez, P., Trédan, G., 2019. Adversarial frontier stitching for remote neural network watermarking. *Neural Computing and Applications* 32, 9233–9244. doi:10.1007/s00521-019-04434-z.
- [46] Li, H., Willson, E., Zheng, H., Zhao, B.Y., 2019a. Persistent and unforgeable watermarks for deep neural networks. arXiv preprint arXiv:1910.01226 .
- [47] Li, Y., Tondi, B., Barni, M., 2020a. Spread-transform dither modulation watermarking of deep neural network. arXiv preprint arXiv:2012.14171 .
- [48] Li, Y., Wang, H.X., 2019. Robust H. 264/AVC video watermarking without intra distortion drift. *Multimedia Tools and Applications* 78, 8535–8557.
- [49] Li, Y., Wu, B., Jiang, Y., Li, Z., Xia, S.T., 2020b. Backdoor learning: A survey. arXiv preprint arXiv:2007.08745 .
- [50] Li, Z., Hu, C., Zhang, Y., Guo, S., 2019b. How to prove your model belongs to you: a blind-watermark based framework to protect intellectual property of dnn, in: Proceedings of the 35th Annual Computer Security Applications Conference, pp. 126–137.
- [51] Lin, S.D., Shie, S.C., Guo, J.Y., 2010. Improving the robustness of dct-based image watermarking against jpeg compression. *Computer Standards & Interfaces* 32, 54–60.
- [52] Lukas, N., Zhang, Y., Kerschbaum, F., 2021. Deep neural network fingerprinting by conferrable adversarial examples, in: 2021 International Conference on Learning Representations (ICLR).
- [53] Luo, M., Bors, A.G., 2011. Surface-preserving robust watermarking of 3-d shapes. *IEEE Transactions on Image Processing* 20, 2813–2826.
- [54] M. Barni, F. Bartolini, A.P., 2001. Improved wavelet-based watermarking through pixel-wise masking. *IEEE Transactions on Image Processing* 10, 783–791.
- [55] Miller, M.L., Doërr, G.J., Cox, I.J., 2004. Applying informed coding and embedding to design a robust high-capacity watermark. *IEEE Transactions on image processing* 13, 792–807.
- [56] Nagai, Y., Uchida, Y., Sakazawa, S., Satoh, S., 2018. Digital watermarking for deep neural networks. *International Journal of Multimedia Information Retrieval* 7, 3–16. doi:10.1007/s13735-018-0147-1.
- [57] Namba, R., Sakuma, J., 2019. Robust watermarking of neural network with exponential weighting, in: Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security, pp. 228–240.
- [58] Nikolaidis, N., Pitas, I., 1998. Robust image watermarking in the spatial domain. *Signal processing* 66, 385–403.
- [59] Ohbuchi, R., Ueda, H., Endoh, S., 2002. Robust watermarking of vector digital maps, in: Proceedings. IEEE International Conference on Multimedia and Expo, IEEE. pp. 577–580.
- [60] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A., 2016. The limitations of deep learning in adversarial settings, in: 2016 IEEE European symposium on security and privacy (EuroS&P), IEEE. pp. 372–387.
- [61] Podilchuk, C.I., Delp, E.J., 2001. Digital watermarking: algorithms and applications. *IEEE signal processing Magazine* 18, 33–46.
- [62] Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.
- [63] Rouhani, B.D., Chen, H., Koushanfar, F., 2019. Deepsigns: an end-to-end watermarking framework for protecting the ownership of deep neural networks, in: The 24th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS). ACM.
- [64] Sakazawa, S., Myodo, E., Tasaka, K., Yanagihara, H., 2019. Visual decoding of hidden watermark in trained deep neural network, in: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE. pp. 371–374.
- [65] Seo, J.S., Yoo, C.D., 2006. Image watermarking based on invariant regions of scale-space representation. *IEEE Transactions on Signal Processing* 54, 1537–1549.
- [66] Shafieinejad, M., Wang, J., Lukas, N., Li, X., Kerschbaum, F., 2019. On the robustness of the backdoor-based watermarking in deep neural networks. arXiv preprint arXiv:1906.07745 .
- [67] Shehab, M., Bertino, E., Ghafoor, A., 2007. Watermarking relational databases using optimization-based techniques. *IEEE transactions on Knowledge and Data Engineering* 20, 116–129.
- [68] Singh, A.K., Sharma, N., Dave, M., Mohan, A., 2012. A novel technique for digital image watermarking in spatial domain, in: 2012 2nd IEEE International Conference on Parallel, Distributed and Grid Computing, IEEE. pp. 497–501.
- [69] Singh, N., Jain, M., Sharma, S., 2013. A survey of digital watermarking techniques. *International Journal of Modern Communication Technologies and Research* 1, 265852.
- [70] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Good-

- fellow, I., Fergus, R., 2014. Intriguing properties of neural networks, in: 2014 International Conference on Learning Representations (ICLR).
- [71] Szyller, S., Atli, B.G., Marchal, S., Asokan, N., 2019. Dawn: Dynamic adversarial watermarking of neural networks. arXiv preprint arXiv:1906.00830 .
- [72] Tew, Y., Wong, K., 2014. Information hiding in HEVC standard using adaptive coding block size decision, in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 5502–5506.
- [73] Trappe, W., Wu, M., Wang, Z.J., Liu, K.R., 2003. Anti-collusion fingerprinting for multimedia. IEEE Transactions on Signal Processing 51, 1069–1087.
- [74] Tsui, T.K., Zhang, X., Androustos, D., 2008. Color image watermarking using multidimensional fourier transforms. IEEE Transactions on Information Forensics and Security 3, 16–28. doi:10.1109/TIFS.2007.916275.
- [75] Uchida, Y., Nagai, Y., Sakazawa, S., Satoh, S., 2017. Embedding watermarks into deep neural networks, in: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pp. 269–277.
- [76] Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., Zhao, B.Y., 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks, in: 2019 IEEE Symposium on Security and Privacy (SP), IEEE. pp. 707–723.
- [77] Wang, J., Wu, H., Zhang, X., Yao, Y., 2020. Watermarking in deep neural networks via error back-propagation. Electronic Imaging 2020, 22–1.
- [78] Wang, T., Kerschbaum, F., 2019a. Attacks on digital watermarks for deep neural networks, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 2622–2626.
- [79] Wang, T., Kerschbaum, F., 2019b. Riga: Covert and robust white-box watermarking of deep neural networks. arXiv preprint arXiv:1910.14268 .
- [80] Wang, T., Kerschbaum, F., 2019c. Robust and undetectable white-box watermarks for deep neural networks. arXiv preprint arXiv:1910.14268 .
- [81] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2097–2106.
- [82] Wu, H., Liu, G., Yao, Y., Zhang, X., 2020. Watermarking neural networks with watermarked images. IEEE Transactions on Circuits and Systems for Video Technology .
- [83] Xiang, Y., Natgunanathan, I., Guo, S., Zhou, W., Nahavandi, S., 2014. Patchwork-based audio watermarking method robust to desynchronization attacks. IEEE/ACM Transactions on Audio, Speech, and Language Processing 22, 1413–1423.
- [84] Zagoruyko, S., Komodakis, N., 2016. Wide residual networks, in: Proceedings of the British Machine Vision Conference 2016, British Machine Vision Association. doi:10.5244/c.30.87.
- [85] Zhang, H., Patel, V.M., 2018. Density-aware single image de-raining using a multi-stream dense network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 695–704.
- [86] Zhang, J., Chen, D., Liao, J., Fang, H., Zhang, W., Zhou, W., Cui, H., Yu, N., 2020. Model watermarking for image processing networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence (AAAI). pp. 12805–12812. doi:10.1609/aaai.v34i07.6976.
- [87] Zhang, J., Gu, Z., Jang, J., Wu, H., Stoecklin, M.P., Huang, H., Molloy, I., 2018. Protecting intellectual property of deep neural networks with watermarking, in: Proceedings of the 2018 on Asia Conference on Computer and Communications Security, pp. 159–172.
- [88] Zhao, J., Hu, Q., Liu, G., Ma, X., Chen, F., Hassan, M.M., 2020a. Afa: Adversarial fingerprinting authentication for deep neural networks. Computer Communications 150, 488–497.
- [89] Zhao, X., Wu, H., Zhang, X., 2020b. Watermarking graph neural networks by random graphs. arXiv preprint arXiv:2011.00512 .