Unforgeable Backdoor-based Watermarking for Deep Learning Models by Spectral Steganography

Paper ID: 4558

Abstract

Protecting deep learning models as intellectual properties is necessary for commercializing artificial intelligence products. Such protection is usually realized by backdoors when access to the pirated model is limited. Backdoor-based deep neural network watermarking are fragile to spoil attack and cannot handle overwriting without the assistance of extra top-level mechanisms. This paper presents a backdoor-based watermarking scheme, `Spectral-Sign`, whose security is built upon the secrecy of the training data. The proposed scheme generates backdoor triggers by steganography in the spectrum domain of the target deep model. A verification protocol is further designed for interactive proof of ownership. By bridging the gap between model security and data privacy, the proposed scheme is provably unforgeable and secure against the overwriting attack. Our work exploits the relationship between deep learning model protection and data privacy and paves the way for more practical and reliable machine learning model security.

Introduction

The development of deep learning facilitates the application of deep neural networks (DNN) in computer vision (Zhang et al. 2021b), natural language processing (Kordjamshidi, Pustejovsky, and Moens 2020), etc. The extensive deployment of deep learning models as services calls for standardized commercialization and protection of DNNs as intellectual properties (IP) (Ong et al. 2021).

Watermarking is a promising technique for DNN IP protection. An *owner* adopts a DNN watermarking scheme to embed owner-dependent information into the DNN. Then the owner can prove its ownership to a third-party *customer* by revealing the watermark under an ownership verification (OV) protocol. Such a protocol can be either centralized, where a trusted center publishes the legitimate proof or decentralized, where a community of distributed agents cooperates to verify the ownership (Li, Wang, and Liew 2021).

If the owner has full access to the pirated model then the identity information can be embedded into the DNN's parameters (Uchida et al. 2017; Liu, Weng, and Zhu 2021) or the statistics of its feature layers (Darvish Rouhani, Chen, and Koushanfar 2019). These *white-box* DNN watermarking schemes are unavailable in most real-world scenarios since the adversary is not obliged to submit its model for examination. Therefore, *black-box* DNN watermarking schemes, which are usually based on the backdoor (Wenger et al. 2021), are proposed to protect DNNs deployed as services.

To invoke a backdoor-based DNN watermarking scheme, the owner generates *backdoor triggers* and corresponding labels as its watermark. Then the DNN learns both the normal dataset and labeled triggers. Even with only the black-box access, the owner can evoke the backdoor as identity proof. Despite their availability, backdoor-based DNN watermarking schemes have some critical defects such as the vulnerability against the ambiguity attack, the overwriting attack, and the removal attack.

This paper proposes a novel backdoor-based DNN watermarking scheme, `Spectral-Sign`. It couples the backdoor with the training dataset by an auxiliary autoencoder. Owner-dependent watermarks are generated from an extra spectrum independent from that of normal samples. Such *spectral steganography* protects the secrecy of normal samples and the uniqueness of the ownership. In the ownership verification (OV) stage, the owner proves its ownership over a DNN to a customer under the assistance of a semi-honest shepherd. The contributions of this paper are threefold:

- We propose an unforgeable backdoor-based DNN watermarking scheme that is secure against overwriting.
- Our work indicates that the reliability of DNN IP protection can be built upon the privacy of data.
- Experiments justify the privilege of our proposal over other schemes regarding unforgeability and robustness.

Challenges and Motivations

To watermark a deep learning model M_{clean}, a DNN watermarking scheme embeds owner-dependent information, `key`, into the model. This process returns a watermarked model M_{WM} and a verification module `verify`, with which `key` can be correctly recognized from M_{WM}.

To leave watermarks in DNN for black-box verification, backdoor-based watermarking schemes generate different types of triggers and encode `key` into their labels. Stamped triggers are normal samples marked with an extra stamp. Random triggers are random noise released from a pseudorandom generator (Zhu et al. 2020). Out-of-dataset trig-

gers are samples outside the training dataset (Zhang et al. 2018). Wonder Filter triggers are images with out-of-range pixels (Li et al. 2019). Penetrative triggers (Li and Wang 2021) trained from surrogate autoencoders can penetrate adversarial filtering and ensure functionality.

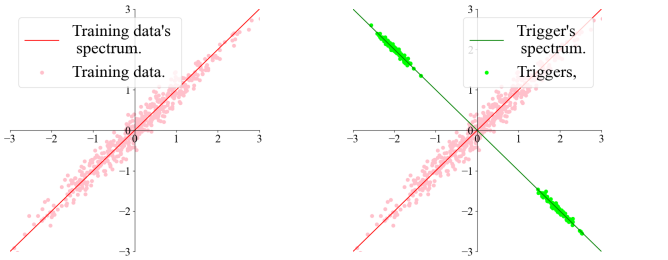
Existing backdoor-based schemes are confronted with several problems: (i) Some triggers cannot encode the owner's identity and the security of the ownership relies on the secrecy of the watermarking scheme. (ii) Triggers are homogeneous to some extend and can be reversed or cleaned by removal techniques such as Neural Cleanse (NC) (Wang et al. 2019) and Fine-Pruning (FP) (Liu, Dolan-Gavitt, and Garg 2018). NC distinguishes triggers from normal samples and mitigates them. FP prunes off inactive neurons within the DNN when processing normal samples, then fine-tunes the model. When the adversary has obtained some trigger patterns by reverse engineering (Hua, Zhang, and Suh 2018) or eavesdropping on the OV process, it can conduct the spoil attack by fitting the DNN on randomly shuffled labels for triggers. (iii) An adversary can conduct *overwriting* attack, i.e., embed its identification evidence $\texttt{key}_{\text{ADV}}$ into the DNN and declare the ownership. Such overwriting cannot be prevented, so the integrity of the ownership is at risk.

The reason behind these difficulties is that backdoor-based DNN watermarking schemes are mostly independent of the training dataset. If the watermarks are highly correlated with the training data, then: (i) The security of the ownership would depend on the secrecy of the data rather than that of the watermarking scheme; (ii) Removal attacks without full knowledge of the training data cannot accurately sabotage backdoor triggers; (iii) By proving its knowledge on the training dataset, the owner can defeat any adversary using overwriting.

Thence, assuming the secrecy of the training data, an adversary can neither forge its own watermarks nor differentiate backdoor triggers from normal samples and invalidate them. For most commercial DNN products, the secrecy of the dataset is crucial (Choquette-Choo et al. 2021) and can be achieved by homomorphic encryption (Li et al. 2020), differential privacy (Wei et al. 2020), etc.

We introduce the training dataset $\mathcal{D}$ for M_{clean} as the extra evidence in the watermarking scheme. The `verify` module now incorporates the statistics of a dataset as its additional input (the entire dataset is not exposed to the verifier for data protection). The adversary can conduct overwrite using its small dataset, $\mathcal{D}_{\text{ADV}}$, which is assumed to share a similar distribution with $\mathcal{D}$'s. Such an assumption holds for many real-world DNNs deployed as services, where the owner has collected abundant data while the adversary has only partial knowledge of the service and scanty data.

To comprehensively encode the training dataset without revealing it, we resort to the spectral domain. In this domain, identification information as triggers can be compactly embedded within extra dimensions. Without the knowledge on the training dataset, an adversary cannot estimate the spectral information and infer the patterns of triggers. These triggers are subject to a more complex and heterogeneous distribution, hence are more robust against adversarial tuning and removal attacks, this setting is illustrated in Fig. 1. Finally, by proving that its spectrum correctly encodes more data, the owner can dominate the adversary during an interactive OV process so the watermark is robust against the overwriting attack. Concretely, the probability that the genuine owner passes OV is higher than that of the adversary given the statistics of both parties' datasets, so the customer only needs to cross-validate both evidence without consulting other authorities for extra time-stamps.



(a) The original spectrum. (b) Inserting an extra dimension.

Figure 1: Generating triggers in the spectral domain.

The `Spectral-Sign` Scheme

We adopt an independent autoencoder, `AE` to encode the training dataset. The intermediate representations of $\mathcal{D}$ in `AE` form $\mathcal{D}$'s *spectrum*. To conceal this spectrum for privacy concerns and generate triggers alongside normal samples, a collection of vectors that are orthogonal to the original spectrum is added into `AE`. The OV proof is conducted under a semi-centralized protocol as shown in Fig. 2, where the customer obtains the proof without learning the data's spectrum. During the OV, the owner shows that: (i) It possesses data that is consistent with the DNN. (ii) It has injected backdoors into the model. Therefore, the OV consists of two branches:

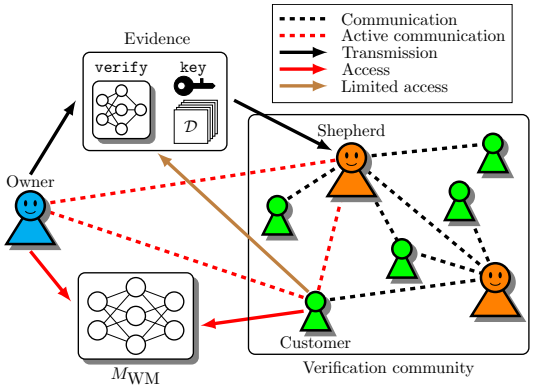


Figure 2: The semi-centralized verification protocol.

- The owner submits normal samples that can be correctly reconstructed by the autoencoder and processed by the DNN. The customer verifies that they belong to a subspectrum within the autoencoder.
- The owner submits triggers that correctly invoke backdoor responses. The customer verifies that triggers have

components from another independent spectrum.

Spectral steganography within the autoencoder

AE contains an encoder Enc, two fully connected linear layers, and a decoder Dec. The output of a sample $\mathbf{x}$ is:

$$\texttt{AE}(\mathbf{x}) = \texttt{Dec}(\mathbf{W}_2^{\mathrm{T}} \cdot \mathbf{W}_1 \cdot \texttt{Enc}(\mathbf{x})).$$

The two linear layers form a principal component analysis module. Specifically, $\mathbf{W}_1$'s rows are the eigenvectors of $\mathcal{D}$ after the mapping of Enc (modulo linear combinations), i.e., its spectrum. The intermediate representation of $\mathbf{x}$:

$$h(\mathbf{x}) = \mathbf{W}_1 \cdot \texttt{Enc}(\mathbf{x})$$

is the projection of $\texttt{Enc}(\mathbf{x})$ on the spectrum. The dimensionality of $h(\mathbf{x})$ is H. AE is firstly trained to reconstruct $\mathcal{D}$ by minimizing the following loss:

$$\begin{aligned}\mathcal{L}_{\text{reconstruct}}(\texttt{Enc}, \texttt{Dec}, \mathbf{W}_1, \mathbf{W}_2) = &\sum_{\mathbf{x}\in\mathcal{D}} \|\mathbf{x} - \texttt{AE}(\mathbf{x})\|_2^2 \\ &+ \lambda_1 \cdot \sum_{\mathbf{x}\in\mathcal{D}} \|h(\mathbf{x})\|_2^2.\end{aligned} \tag{1}$$

Properly tuned AE ensures that no less than $(1-\gamma_1)$ samples' classification labels remain invariant after its processing, i.e., $M_{\text{clean}}(\mathbf{x}) = M_{\text{clean}}(\texttt{AE}(\mathbf{x}))$.

An extra spectrum reserved for backdoor watermarks is inserted into AE. As demonstrated by Fig. 3, this is done by expanding the linear transformations:

$$\hat{\mathbf{W}}_1 = \mathbf{P}\begin{pmatrix}\mathbf{W}_1 \\ \mathbf{W}_1'\end{pmatrix}, \hat{\mathbf{W}}_2 = \mathbf{P}^{-\mathrm{T}}\begin{pmatrix}\mathbf{W}_2 \\ \mathbf{W}_2'\end{pmatrix},$$

$$\hat{h}(\mathbf{x}) = \hat{\mathbf{W}}_1 \cdot \texttt{Enc}(\mathbf{x}), \hat{h}_Q(\mathbf{x}) = \mathbf{W}_1' \cdot \texttt{Enc}(\mathbf{x}),$$

in which $\mathbf{W}_i'$ contains Q vectors that are orthogonal to those principal directions in $\mathbf{W}_i$'s rows (i=1,2) and $\mathbf{P}$ is a random permutation matrix. This spectrum is generated by Schmidt orthogonalization. The projection of a sample $\mathbf{x}$ on this extra

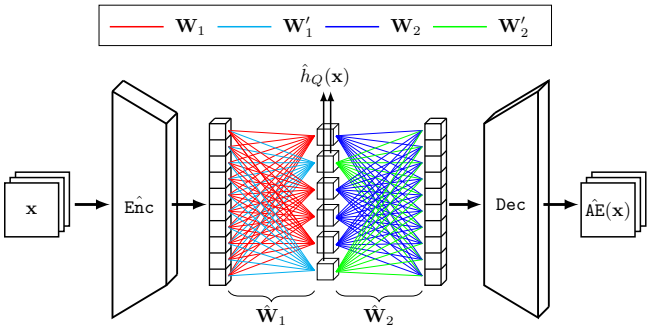


Figure 3: The autoencoder architecture of spectral steganography, AE.

spectrum under the encoder $\texttt{Enc}(\cdot)$ is denoted by $\hat{h}_Q(\mathbf{x})$. The norm of this projection, which we denote as the *residual norm* of $\mathbf{x}$, is uniformly small, i.e., $\|\hat{h}_Q(\mathbf{x})\|_2^2 \leq \delta$ holds for at least $(1-\gamma_2)\cdot|\mathcal{D}|$ samples. Since the eigenvectors corresponding to the H largest eigenvalues of the genuine spectrum have been included in $\mathbf{W}_1$, it can be expected that:

$$\delta < \lambda_{H+1}^2 \cdot Q, \tag{2}$$

where λ_{H+1} is the (H+1)-th largest eigenvalue in $\mathcal{D}$'s spectrum.

To generate a trigger T, a random *index* $\mathbf{u}$ is propagated through the decoder:

$$\mathbf{u} \leftarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_{H+Q}), T(\mathbf{u}) = \texttt{Dec}\left(\hat{\mathbf{W}}_2^{\mathrm{T}} \cdot \mathbf{u}\right).$$

To incorporate the owner's key, the label $y(\mathbf{u})$ for the trigger $T(\mathbf{u})$ is assigned:

$$y(\mathbf{u}) = f(\mathbf{u}, \texttt{key}),$$

where f is a hash function accessible for any party.

Having generated the triggers and their corresponding labels, the owner fine-tunes its DNN, which has been trained on $\mathcal{D}$ by minimizing:

$$\mathcal{L}(M_{\text{clean}}) = \sum_{(\mathbf{x}, y(\mathbf{x}))\in\mathcal{D}} l(y(\mathbf{x}), M_{\text{clean}}(\mathbf{x})),$$

in which $l(\cdot,\cdot)$ is the task-dependent loss, into the watermarked model by minimizing:

$$\begin{aligned}\mathcal{L}(M_{\text{WM}}) = &\sum_{\mathbf{u}\in\mathcal{U}} l(y(\mathbf{u}), M_{\text{WM}}(T(\mathbf{u}))) \\ &+ \lambda_2 \cdot \|M_{\text{clean}} - M_{\text{WM}}\|_p^2,\end{aligned} \tag{3}$$

where $\|\cdot\|_p$ is the norm of the DNN's numerical parameters.

Not knowing $\mathcal{D}$, it is hard to identity the H characteristic eigenvectors from all H+Q candidates in $\hat{\mathbf{W}}_1$'s row space. Therefore, an adversary can neither generate normal samples from AE nor distinguish a normal sample from a trigger. Yet implicitly exposing the entire spectrum remains a threat to data privacy and an adversary obtaining $\hat{\texttt{AE}}$ might pirate the model's ownership by claiming it to be its evidence. We *mask* the spectrum so the OV succeeds only with the owner's authorization.

The OV protocol

To prove its ownership over a DNN M to a customer, the owner requests the assistance of a shepherd. The entire OV protocol proceeds as Algo. 1, where the involved cryptological protocols are detailed in Appendix A. In cases where two parties passing individual verification simultaneously claim to be the owner, the OV protocol operates as Fig. 4 and Algo. 2 to break the tie.

Security analysis

Correctness The correctness and the security against overwriting (unforgeability) of Spectral-Sign are established in the following theorems.

Theorem 1: (Correctness) Algo. 1 as an implementation of verify satisfies the following condition for correct (Li, Wang, and Liew 2021) identity retrieval:

$$\Pr\{\texttt{verify}(M_{\text{WM}}, \texttt{key}|\mathcal{S}(\mathcal{D})) = 1\} \geq 1 - \epsilon(N),$$

where $\mathcal{S}(\mathcal{D})$ is the masked AE and samples delivered by the oblivious transfer protocol, ϵ is negligible in the security parameter N.

Algorithm 1: The OV protocol for `Spectral-Sign`.

Participants: The owner, the shepherd, and the customer.
Require: An oblivious transfer protocol, the SMPC protocol for matrix multiplication.
Output: The ownership proof result.

1: The owner submits the threshold parameters δ, γ_1, γ_2, and its identity `key` to the customer.
2: The owner generates an invertable matrix $\mathbf{Q}_1$, the shepherd generates an invertable matrix $\mathbf{Q}_2$.
3: The owner masks its $\hat{\mathtt{AE}}$ into $\hat{\mathtt{AE}}_{\mathrm{O}}$ with $\hat{\mathbf{W}}_{1,\mathrm{O}} = \mathbf{Q}_1 \cdot \hat{\mathbf{W}}_1$ and $\hat{\mathbf{W}}_{2,\mathrm{O}} = \mathbf{Q}_1^{-\mathrm{T}} \cdot \hat{\mathbf{W}}_2$ and submits $\hat{\mathtt{AE}}_{\mathrm{O}}$ to the shepherd.
4: The shepherd masks $\hat{\mathtt{AE}}_{\mathrm{O}}$ into $\hat{\mathtt{AE}}_{\mathrm{S}}$ with $\hat{\mathbf{W}}_{1,\mathrm{S}} = \mathbf{Q}_2 \cdot \hat{\mathbf{W}}_{1,\mathrm{O}}$ and $\hat{\mathbf{W}}_{2,\mathrm{S}} = \mathbf{Q}_2^{-\mathrm{T}} \cdot \hat{\mathbf{W}}_{2,\mathrm{O}}$ and transmits it to the customer.
5: **Define:** Interaction I:
6: The customer declares a random class label c.
7: The owner submits a sample of this class, $\mathbf{x}$, to the customer using the oblivious transfer protocol.
8: Three parties computes $\hat{h}(\mathbf{x}) = \mathbf{Q}_1^{-1}\mathbf{Q}_2^{-1}\hat{h}_{\mathrm{S}}(\mathbf{x})$ using the SMPC protocol.
9: If $M(\mathbf{x}) = c$, $M(\hat{\mathtt{AE}}_{\mathrm{S}}(\mathbf{x})) = c$, and $\|\hat{h}_Q(\mathbf{x})\|_2^2 \leq \delta$ then this interaction succeeds.
10: **Define:** Interaction II:
11: The customer declares a backdoor trigger.
12: The owner submits a backdoor trigger index, $\mathbf{u} \in \mathcal{U}$, packaged as $\mathbf{Q}_1 \cdot \mathbf{u}$ to the shepherd using the oblivious transfer protocol.
13: The shepherd left-multiplicates it by $\mathbf{Q}_2$ into $\hat{\mathbf{u}}_{\mathrm{S}}$.
14: Three parties computes $\hat{\mathbf{u}} = \mathbf{Q}_1^{-1}\mathbf{Q}_2^{-1}\hat{\mathbf{u}}_{\mathrm{S}}$ using the SMPC protocol.
15: If $M(\hat{\mathtt{Dec}}_{\mathrm{S}}(\hat{\mathbf{W}}_{2,\mathrm{S}}^{\mathrm{T}} \cdot \hat{\mathbf{u}}_{\mathrm{S}})) = f(\hat{\mathbf{u}}, \mathtt{key})$ then this verification succeeds.
16: **for** $n = 1$ to N **do**
17: Three parties conduct a random interaction.
18: **end for**
19: If no less than $(1 - (\gamma_1 + \gamma_2)) \cdot N$ interaction succeeds then returns 1. Otherwise returns 0.

Proof: By definition, for any input $\mathbf{x}$:

$$\hat{\mathtt{AE}}(\mathbf{x}) = \hat{\mathtt{AE}}_{\mathrm{O}}(\mathbf{x}) = \hat{\mathtt{AE}}_{\mathrm{S}}(\mathbf{x}).$$

So the first two conditions in the interaction I hold with probability at least $(1 - \gamma_1)$. Meanwhile, we have:

$$\hat{h}_{\mathrm{S}}(\mathbf{x}) = \mathbf{Q}_2 \cdot \hat{h}_{\mathrm{O}}(\mathbf{x}) = \mathbf{Q}_2 \cdot \mathbf{Q}_1 \cdot \hat{h}(\mathbf{x}),$$

so the last condition in interaction I holds with probability $(1 - \gamma_2)$. For a backdoor trigger $T(\mathbf{u})$, $\mathbf{u} \in \mathcal{U}$, interaction II succeeds with probability one. So Chernoff theorem indicates that the probability that an owner fails an OV is upper bounded by:

$$\min_{\beta<0} \left\{ \left(\frac{\frac{\gamma_1+\gamma_2}{2} + (1 - \frac{\gamma_1+\gamma_2}{2}) \cdot \mathrm{e}^{\beta}}{\mathrm{e}^{(1-(\gamma_1+\gamma_2))\cdot\beta}} \right)^N \right\}, \tag{4}$$

a function negligible in N. The derivation of (4) is detailed in Appendix B. □

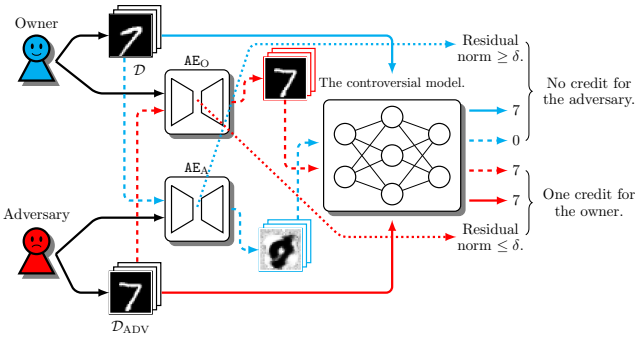


Figure 4: The tie-breaker mechanism.

Algorithm 2: The tie-breaker for `Spectral-Sign`.

Participants: The owner, the adversary, the shepherd, and the customer.
Require: An oblivious transfer protocol, the SMPC protocol for matrix multiplication.
Output: The customer's confidence on the ownership.

1: The owner submits its masked autoencoder $\hat{\mathtt{AE}}_{\mathrm{O}}$ to the customer under the assistance of the shepherd.
2: The adversary submits its masked autoencoder $\hat{\mathtt{AE}}_{\mathrm{A}}$ to the customer under the assistance of the shepherd.
3: **Define:** Interaction III:
4: The customer declares a random class label c.
5: The owner/adversary submits a sample of this class, $\mathbf{x}_{\mathrm{O}}$/$\mathbf{x}_{\mathrm{A}}$ using the oblivious transfer protocol.
6: A credit is assigned to the owner if $M(\hat{\mathtt{AE}}_{\mathrm{O}}(\mathbf{x}_{\mathrm{A}}))$ is c and $\|\hat{h}_{\mathrm{O},Q}(\mathbf{x}_{\mathrm{A}})\|_2^2 \leq \delta$.
7: A credit is assigned to the adversary if $M(\hat{\mathtt{AE}}_{\mathrm{A}}(\mathbf{x}_{\mathrm{O}}))$ is c and $\|\hat{h}_{\mathrm{A},Q}(\mathbf{x}_{\mathrm{O}})\|_2^2 \leq \delta$.
8: **for** $n = 1$ to N **do**
9: The customer, the shepherd, the owner, and the adversary conduct interaction III.
10: **end for**
11: The ownership is assigned to the party with more credits.

Theorem 2: (Dominance) Fixing δ, γ_1, and γ_2, `Spectral-Sign` satisfies the dominance property defined as the following inequality:

$$\begin{aligned}&\Pr\left\{\mathtt{verify}\left(M_{\mathrm{ADV}}, \mathtt{key} | \mathcal{S}(\mathcal{D} \cap \mathcal{D}_{\mathrm{ADV}})\right) = 1\right\} \geq \\ &\quad \Pr\left\{\mathtt{verify}_{\mathrm{ADV}}\left(M_{\mathrm{ADV}}, \mathtt{key}_{\mathrm{ADV}} | \mathcal{S}(\mathcal{D} \cap \mathcal{D}_{\mathrm{ADV}})\right) = 1\right\},\end{aligned} \tag{5}$$

where M_{ADV} and $\mathtt{verify}_{\mathrm{ADV}}$ are obtained by embedding $\mathtt{key}_{\mathrm{ADV}}$ into M_{WM} using $\mathcal{D}_{\mathrm{ADV}}$.

Proof: Interaction II for both parties succeeds with probability one. Given the evidence $\mathcal{S}(\mathcal{D} \cap \mathcal{D}_{\mathrm{ADV}}) = \mathcal{S}(\mathcal{D})$, the probability that the owner succeeds in the interaction I in Algo. 1 is higher than that for the adversary. This is because the adversary's autoencoder, which is tuned on $\mathcal{D}_{\mathrm{ADV}}$, only contains an estimation of $\mathcal{D}$'s spectrum. It can be proven that the increase of the reconstruction loss in l-2 norm is lower

bounded by $\mathcal{O}(Q \cdot \epsilon^2)$, where ϵ is the estimation variance between similar directions in the spectrum due to insufficient data. This is due to the spectrum shift effect detailed in Appendix C. Such variance is a monotonic decreasing function in $\frac{|\mathcal{D}_{\text{ADV}}|}{|\mathcal{D}|}$. Therefore, the adversary suffers from a larger misclassification probability and the probability that it passes the verification is lower according to the Chernoff bound, which implies (5). □

Instead of resorting to (5) by scheduling all thresholds for both parties, the procedure in Algo. 2 is more practical.

Theorem 3: (Unforgeability) Assume that $\mathcal{D}_{\text{ADV}}$ is a subset of $\mathcal{D}$ then Algo. 2 can identify the authentic owner with asymptotical probability one.

Proof: Since $\mathcal{D}_{\text{ADV}} \subset \mathcal{D}$, each sample $\mathbf{x} \in \mathcal{D}_{\text{ADV}}$ is expected to be mapped correctly into the normal spectrum inside the owner's autoencoder. Therefore, the interaction III in Algo. 2 for the owner succeeds with the probability $(1 - \gamma_1 - \gamma_2)$.

Meanwhile, a sample $\mathbf{x} \in \mathcal{D}/\mathcal{D}_{\text{ADV}}$ results in a worse reconstruction from the adversary's autoencoder, which is only tuned to secure the processing of $(1 - \gamma_1)$ portion of $\mathcal{D}_{\text{ADV}}$. As long as the l-2 norm of the reconstruction loss increases as analyzed in Theorem 2, the probability that the adversary earns a credit is strictly lower than $(1 - \gamma_1 - \gamma_2)$. Reusing the Chernoff theorem indicates that the probability the adversary defeats the owner is negligible in N.

Finally, a malicious customer obtaining an autoencoder as the evidence for OV, with which it can correctly reconstruct samples from the training dataset, cannot pretend to be the owner. Since it cannot accurately distinguish the two spectrums so the second condition in the interaction would fail with a probability higher than $(1-\gamma_2)$. □

Privacy-preserving In `Spectral-Sign`, the customer obtaining the masked autoencoder cannot infer the original spectrum of the normal dataset. Otherwise, it must have obtained $\mathbf{Q}_1$ and $\mathbf{Q}_2$, which are contradictive to the security of the underlying SMPC protocol. All N rounds of examination in Algo. 1 can be paralleled into one SMPC operation, where the customer provides a matrix of shape $N \cdot (Q+H)$. Let $N \ll (Q+H)$ then the customer cannot uniquely identify $\mathbf{Q}_1^{-1} \cdot \mathbf{Q}_2^{-1}$ so the parameters of $\hat{\text{AE}}$ remains confidential. Since $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are randomized for each OV, a customer cannot pile up matrices to compromise the original autoencoder. It is also impossible to generate $\mathbf{x} \in \mathcal{D}$ from the autoencoder since the customer cannot differentiate the two spectrums without the assistance of other parties. Simply propagating an intermediate representation through the decoder would only yield a hybrid sample.

On the other hand, the owner cannot counterfeit the computation result and separate the spectrum, since it has no knowledge of $\mathbf{Q}_2$ and the oblivious transfer protocol ensures that the owner has no knowledge of the customer's choice on the sample. This is the reason behind the three-party configuration in OV. The same reasoning indicates that a malicious customer cannot pirate a DNN using the $\hat{\text{AE}}_{\text{S}}$ it received.

Experiments and Discussions

Settings

To examine the capability of `Spectral-Sign`, we conducted experiments on DNN for image classification. The outputs of classification models are the labels, leaving little space for schemes other than backdoor such as deep image watermarking (Zhang et al. 2021a). So backdoor-based DNN watermarking schemes become the only choice in such a black-box setting. We adopted three datasets: MNIST (Deng 2012), Fashion (Xiao, Rasul, and Vollgraf 2017), and CIFAR10 (Krizhevsky, Hinton et al. 2009). The convolutional neural networks AlexNet, VGG-16 (Alippi, Disabato, and Roveri 2018), and ResNet-50 (He et al. 2016) were adopted as the architecture of the backbone deep learning model. As for the autoencoder, we adopted a variational autoencoder structure (Pu et al. 2016) with six-layer encoder/decoder and `tanh` activation function. For the hyperparameters, we set $\lambda_1 = 0.05$ in (1) to initiate the autoencoder and $\lambda_2 = 1$ in (3) to fine-tune it after spectral steganography. Adam (Zhang 2018) with a three-stage schedule was adopted as the optimizer. All experiments were implemented under `PyTorch` framework.

The configuration of parameters

To instantiate `Spectral-Sign`, we studied the thresholds γ_1, γ_2, and δ w.r.t. the autoencoder's configuration. The threshold γ_1, i.e., the percentage of samples whose labels vary after the filtering of the autoencoder, was evaluated w.r.t. the hidden dimensionality H. We applied grid search for H in $[10, \cdots, 80]$ with step 10. The results are illustrated in Fig. 5. The autoencoder with a larger H had the stronger representation ability and consequently the smaller γ_1. Compared across three datasets, we observed that γ_1 is also a function of the dataset's complexity, the less diversified the dataset is, the smaller is γ_1. With $H = 80$, it is safe to set $\gamma_1 = 0.05$ for all datasets.

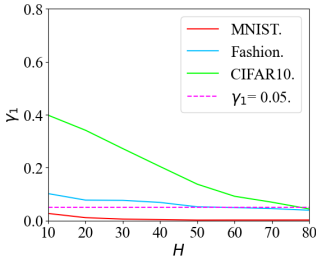


Figure 5: γ_1 w.r.t H.

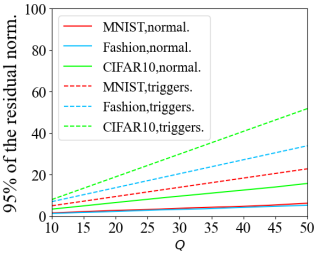


Figure 6: 95% percentile for the residual norm w.r.t. Q.

To select proper δ and γ_2, we fixed $\gamma_2 = 0.05$ so δ is the 95% percentile for the residual norm $\hat{h}_Q$ on the extra spectrum. We evaluated δ as a function of Q with $H = 80$, during which we remarked the averaged norm on the spectrum (for the Q smallest components) for normal samples. Results are demonstrated in Fig. 6. It can be observed that as predicted by (2), δ grew in Q. Compared with triggers, the residual norm of a normal image is small, so extra dimensions

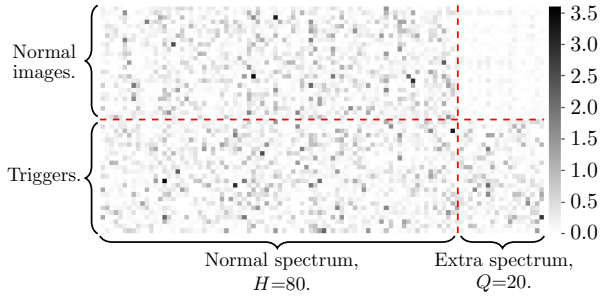


Figure 7: The representations of normal images and triggers within the autoencoder.

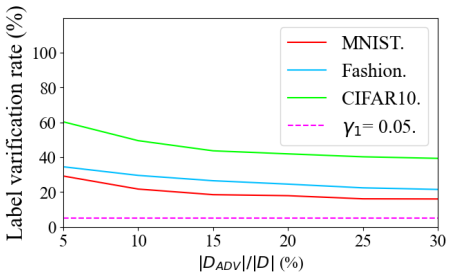


Figure 8: Percentage of sample whose label varies after the adversary's autoencoder w.r.t. $\frac{|\mathcal{D}_{\text{ADV}}|}{|\mathcal{D}|}$.

have almost no influence on the autoencoder's reconstruction ability. Therefore, the owner can prove its knowledge of the dataset by separating two spectrums. As a balance between complexity and sensitivity, we adopted the configuration $H = 80$, $Q = 20$, with $\delta = 5$ so 95% of the residual norms of triggers are uniformaly three times larger than that of normal samples.

The effect of spectral steganography

We selected 50 samples from $\mathcal{D}$ and 50 triggers and illustrate the hidden representations for these samples in Fig. 7, where the first 80 dimensions are corresponding to the spectrum of $\mathcal{D}$. It is obvious that normal samples had uniformly lower components on the extra spectrum, while triggers did not share this property. Therefore the threshold for $\hat{h}_Q$ is an effective distinguisher of normal samples and triggers.

The unforgeability against the adversary

To ensure the dominance condition and the functionality of the tie-breaker mechanism, we measured the percentage of sample whose label varies after the reconstruction of the autoencoder trained on $\mathcal{D}_{\text{ADV}}$ w.r.t. $\frac{|\mathcal{D}_{\text{ADV}}|}{|\mathcal{D}|}$ and the results are illustrated in Fig. 8. Since the error rate was uniformly larger than 0.1 even when the adversary has pirated 30% of the training dataset (which has been a destructive violation of the owner's data security), it is safe to set $\gamma_1 = 0.05$.

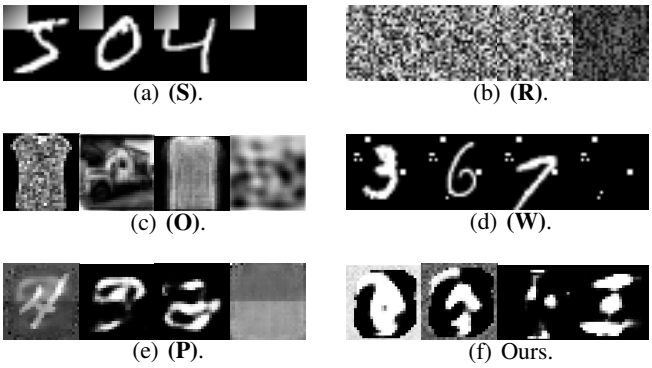


Figure 9: Examples of three triggers and one pattern obtained from reverse engineering for MNIST under studied backdoor-based watermarking schemes.

Security against	(S)	(R)	(O)	(W)	(P)	Ours
Identifying triggers.	×	×	✓	×	✓	✓
Generating triggers.	×	×	×	×	✓	✓
Invalidating a key.	×	✓	✓	×	×	✓

Table 1: Security level of different backdoor-based schemes.

Comparisons and discussions

For comparisons, we considered other backdoor-based DNN watermarking schemes: Stamped triggers **(S)**, random triggers **(R)**, out-of-dataset triggers **(O)**, Wonder Filter **(W)**, and penetrative triggers **(P)**. Triggers in `Spectral-Sign` compactly encode the owner's identity information and are indistinguishable from a normal sample without knowledge of the training dataset. These differences are listed in Table 1, where we assume that the adversary has full knowledge of the watermarking scheme.

Patterns of triggers Examples of backdoor triggers and patterns obtained by reversing the model (Wang et al. 2019) for six schemes in MNIST are demonstrated Fig. 9. It can be observed that unlike **(S)**, **(R)**, **(W)**, and **(P)**, triggers in `Spectral-Sign` are subject to a more diversified distribution and the model reverse hardly reveal any information about the patterns of the trigger or the normal data. The diversity forms the basis of the robustness against the removal or the spoil attack.

Robustness against adaptive attacks To evaluate the functionality and robustness of backdoors in `Spectral-Sign`, we recorded the decline of classification accuracy of the backbone DNN after backdoor embedding. Then we conducted NC, FP, and the spoil attack to different backdoor-based DNN watermarking schemes and recorded the decline of their accuracy. For each scheme, 100 triggers were generated and embedded into the DNN as backdoors. In NC the DNN was tuned w.r.t. $\mathcal{D}_{\text{ADV}}$ whose size was 30% of $\mathcal{D}$'s and FP firstly pruned off 15% neurons and tuned the DNN w.r.t. 30% samples of $\mathcal{D}$. In the spoil attack, 10% of the triggers

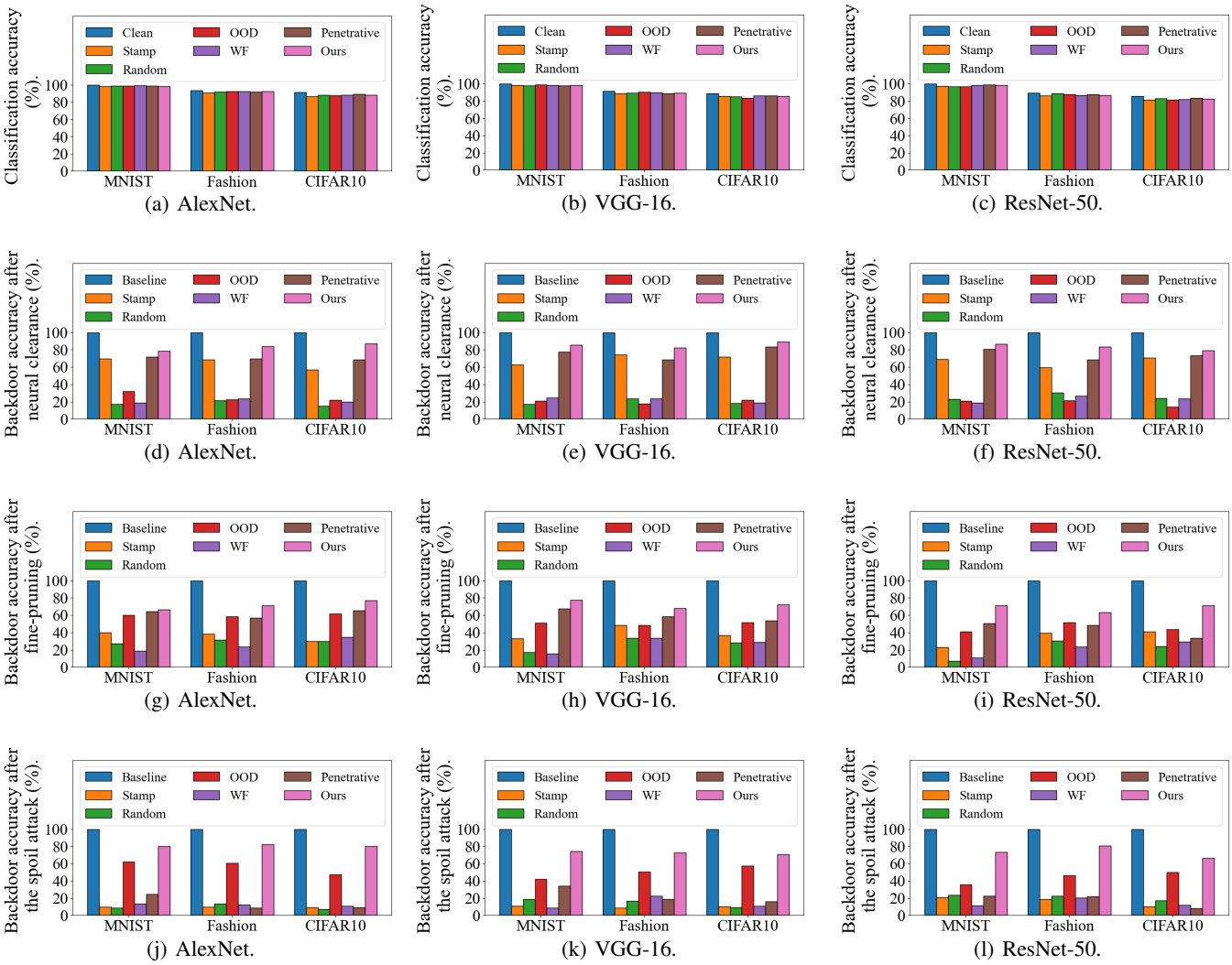


Figure 10: Evaluation of backdoor-based DNN watemarking schemes' peformance on the performance of the watermarked DNN (a)-(c), the backdoor accuracy after NC (d)-(f), FP (g)-(i), and the spoil attack (j)-(l).

were assumed to be exposed to the adversary, who tuned the DNN to fit arbitrary labels for them. The results for three backbone DNNs are summarized in Fig. 10.

Fig. 10 (a)-(c) demonstrate that the impact of backdoor-based watermarking schemes on the DNN's normal performance was uniformly small. Rest plots indicate that our scheme is the most robust one against fine-pruning and the spoil attack. This is because: (i) Triggers in `Spectral-Sign` are subject to a distribution intersected with that for normal samples, so pruning and tuning cannot efficiently modify the model's responses on them. (ii) Unlike **(S)** and **(W)** whose triggers had distinguishing features, the triggers' pattern in our scheme remains diversified and intractable for the adversary, so the spoil attack cannot damage the backdoors significantly. Although **(O)** achieved suboptimal performance regarding FP and the spoil attack by introducing a distribution out of the adversary's knowledge, it cannot compactly encode the owner's identity and lack forensics value.

Conclusion

To secure ownership verification for DNN in black-box settings, we propose to build the watermarking scheme on the secrecy of the dataset. The scheme `Spectral-Sign` instantiates this intuition by incorporating the dataset into an autoencoder and hiding it using spectral steganography. Apart from its unforgeability and privacy-preserving property, the proposed scheme inherently solves the overwriting threat by introducing a tie-breaking mechanism. Under this setting, an adversary obtaining insufficient data cannot pirate the deep learning model's ownership. Experiment results demonstrate that triggers in our proposal are more robust against adaptive attacks, making `Spectral-Sign` a practical candidate in DNN IP protection.

References

Alippi, C.; Disabato, S.; and Roveri, M. 2018. Moving convolutional neural networks to embedded systems: the alexnet and VGG-16 case. In *2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, 212–223. IEEE.

Choquette-Choo, C. A.; Tramer, F.; Carlini, N.; and Papernot, N. 2021. Label-only membership inference attacks. In *International Conference on Machine Learning*, 1964–1974. PMLR.

Darvish Rouhani, B.; Chen, H.; and Koushanfar, F. 2019. DeepSigns: an end-to-end watermarking framework for ownership protection of deep neural networks. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, 485–497.

Deng, L. 2012. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6): 141–142.

Fan, L.; Ng, K. W.; Chan, C. S.; and Yang, Q. 2021. DeepIP: Deep Neural Network Intellectual Property Protection with Passports. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hua, W.; Zhang, Z.; and Suh, G. E. 2018. Reverse engineering convolutional neural networks through side-channel information leaks. In *2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*, 1–6. IEEE.

Kordjamshidi, P.; Pustejovsky, J.; and Moens, M. F. 2020. Representation, Learning and Reasoning on Spatial Language for Downstream NLP Tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, 28–33.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Li, F.-Q.; and Wang, S.-L. 2021. Persistent Watermark For Image Classification Neural Networks By Penetrating The Autoencoder. In *2021 IEEE International Conference on Image Processing (ICIP)*, 3063–3067.

Li, F.-Q.; Wang, S.-L.; and Liew, A. W.-C. 2021. Regulating Ownership Verification for Deep Neural Networks: Scenarios, Protocols, and Prospects. IJCAI2021 Workshop. arXiv:2108.09065.

Li, H.; Willson, E.; Zheng, H.; and Zhao, B. Y. 2019. Persistent and unforgeable watermarks for deep neural networks. *arXiv preprint arXiv:1910.01226*.

Li, J.; Kuang, X.; Lin, S.; Ma, X.; and Tang, Y. 2020. Privacy preservation for machine learning training and classification based on homomorphic encryption schemes. *Information Sciences*, 526: 166–179.

Liu, H.; Weng, Z.; and Zhu, Y. 2021. Watermarking Deep Neural Networks with Greedy Residuals. In *International Conference on Machine Learning*, 6978–6988. PMLR.

Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, 273–294. Springer.

Ong, D. S.; Chan, C. S.; Ng, K. W.; Fan, L.; and Yang, Q. 2021. Protecting Intellectual Property of Generative Adversarial Networks From Ambiguity Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3630–3639.

Pu, Y.; Gan, Z.; Henao, R.; Yuan, X.; Li, C.; Stevens, A.; and Carin, L. 2016. Variational autoencoder for deep learning of images, labels and captions. *Advances in neural information processing systems*, 29: 2352–2360.

Uchida, Y.; Nagai, Y.; Sakazawa, S.; and Satoh, S. 2017. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, 269–277.

Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; and Zhao, B. Y. 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, 707–723. IEEE.

Wei, K.; Li, J.; Ding, M.; Ma, C.; Yang, H. H.; Farokhi, F.; Jin, S.; Quek, T. Q.; and Poor, H. V. 2020. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15: 3454–3469.

Wenger, E.; Passananti, J.; Bhagoji, A. N.; Yao, Y.; Zheng, H.; and Zhao, B. Y. 2021. Backdoor Attacks Against Deep Learning Systems in the Physical World. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6206–6215.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. .

Zhang, J.; Chen, D.; Liao, J.; Zhang, W.; Feng, H.; Hua, G.; and Yu, N. 2021a. Deep Model Intellectual Property Protection via Deep Watermarking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, J.; Gu, Z.; Jang, J.; Wu, H.; Stoecklin, M. P.; Huang, H.; and 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*, 159–172.

Zhang, Y.; Li, Z.; Xie, Y.; Qu, Y.; Li, C.; and Mei, T. 2021b. Weakly Supervised Semantic Segmentation for Large-Scale Point Cloud. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3421–3429.

Zhang, Z. 2018. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, 1–2. IEEE.

Zhu, R.; Zhang, X.; Shi, M.; and Tang, Z. 2020. Secure neural network watermarking protocol against forging attack. *EURASIP Journal on Image and Video Processing*, 2020(1): 1–12.