Journal of Multimedia Information System VOL. 13, NO. 2, June 2026 (pp. 73-80): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2026.13.2.73

Reversible Trigger Set for Copyright Protection of Neural Network

Kaiyang Hu[1,2], Lu Leng[1,2*]

Abstract

Existing backdoor-based watermarking methods for copyright protection of neural network (NN) typically rely on manually designed fixed triggers. These fixed triggers are perceptible, so they are vulnerable to detection and removal, which often lack verifiable ownership evidence. To address these limitations, this paper proposes a copyright protection method of NN based on stego trigger set. The proposed method selects some samples from the clean training data as carrier samples, embeds copyright data into these carrier samples via reversible data hiding, and generates stego triggers that are imperceptible and have high visual quality. These stego triggers constitute the trigger set during model training, enabling the implantation of verifiable watermark behaviors without significantly degrading the model's accuracy. During the verification stage, ownership can be preliminarily confirmed in a black-box manner by querying the suspicious model with the stego trigger set and examining its output responses. Furthermore, the embedded copyright data can be extracted from the stego trigger set, and the carrier samples can be losslessly restored, forming a reproducible evidence chain through model response, copyright data extraction, and reversible restoration. Extensive experiments on the CIFAR datasets show that the proposed method achieves a trigger recognition accuracy of 100% or nearly 100%.

Key Words: Neural Network, Backdoor Watermarking, Copyright Protection, Reversible Data Hiding.

I. INTRODUCTION

With the widespread application of deep neural networks (DNNs) in key fields, such as image recognition, natural language processing, and autonomous driving, DNNs have become important digital intellectual properties [1]. However, the massive data, substantial computing power, and expert resource investment required to train high-performance models have led to the increasing prevalence of model theft and piracy [2]. How to effectively protect the copyright of neural network (NN) has thus become an urgent academic and industrial issue [3].

Traditional copyright protection schemes, such as digital watermarking, have been widely adopted in the multimedia domain [4]. However, directly applying these techniques to NN models is challenging, since model parameters have less correlation. In addition, model parameters are often inaccessible in practice, and easily modified through fine-tuning or pruning.

To address these issues, black-box watermarking based on trigger inputs has attracted attention, as ownership can be verified solely through input-output behavior without accessing internal parameters. Nevertheless, many existing approaches employ predefined or sample-independent triggers that may exhibit identifiable patterns and thus remain vulnerable to statistical anomaly detection, affecting their stealthiness and robustness [5].

In parallel, reversible data hiding (RDH), a security technique enabling lossless restoration of the clean carrier sample after extracting hidden data, has achieved rapid advancement in many fields such as data security and copyright protection in recent years [6]. Its "reversibility" offers a novel perspective for trigger design: if generated triggers are capable of being restored to the clean samples, this not only bolsters the credibility of copyright verification, but also effectively avoids irreversible modification of the clean data. This property further enhances the reliability and evidential robustness of the proposed ownership authentication mechanism.

Therefore, this study introduces a watermarking framework that integrates RDH to generate verifiable and restorable trigger sets for ownership protection of NN. The core idea is to generate a stego trigger set from the clean training data by exploiting their inherent visual consistency, and to embed explicit copyright data into a subset of carrier samples selected under prediction-consistency constraints

Manuscript received February 26, 2026; Revised March 17, 2026; Accepted March 19, 2026. (ID No. JMIS-26M-02-010)
Corresponding Author (*): Lu Leng, +86-791-86453251, leng@nchu.edu.cn
[1]Jiangxi Provincial Key Laboratory of Image Processing and Pattern Recognition, Nanchang Hangkong University, Nanchang, China, 3103260659@qq.com, leng@nchu.edu.cn
[2]School of Software, Nanchang Hangkong University, Nanchang, China, 3103260659@qq.com, leng@nchu.edu.cn

via RDH. As a result, the stego trigger set achieves strong visual imperceptibility together with reliable trigger recognition accuracy, verifiability, and lossless restorability.

Our contributions are summarized as follows:

(1) We propose a reversible stego trigger set generation framework for copyright protection of NN, where the triggers are generated by embedding explicit copyright data into selected training carrier samples via RDH. This avoids manually crafted fixed patterns and yields visually imperceptible yet verifiable triggers.
(2) We develop a reliability-enhanced, two-stage ownership verification scheme. Beyond conventional black-box trigger-to-target response checking, we further perform copyright data extraction and lossless restoration of carrier sample to form a dual-evidence chain. Notably, the restored carrier samples match corresponding public-dataset samples, enabling reproducible verification and reducing the risk of arbitrary trigger fabrication.
(3) Extensive experiments on CIFAR-10 and CIFAR-100 datasets across multiple architectures show high trigger recognition accuracy with negligible degradation of clean test accuracy, validating the effectiveness and practicality of the proposed method.

The remainder of this paper is organized as follows: Section II reviews related works on RDH and copyright protection of NN. Section III introduces the proposed framework and methodology. Section IV presents and analyzes the experimental results. Finally, Section V concludes the paper.

II. RELATED WORKS

Recent advances in NN watermarking and RDH have provided important technical foundations for model copyright protection. This section reviews related works in these two areas to clarify the research context and existing limitations.

2.1. Reversible Data Hiding

RDH is a specialized branch of data hiding. Its core advantage lies in the ability to losslessly restore the carrier sample after secret data extraction. This property has recently been extended to the domain of adversarial examples, as demonstrated by Yang et al. [7], who proposed reversible adversarial examples for recognition control in computer vision. This property makes RDH particularly valuable in applications requiring strict carrier sample fidelity, including military imaging, medical data management, and forensic analysis.

Since Barton [8] first proposed the concept in 1997, RDH has developed two major branches: plaintext domain and ciphertext domain. This study focuses on plaintext domain technology, which is generally categorized into three main embedding strategies.

Lossless compression (LC) compresses specific image components to vacate space for data embedding, but generally suffers from limited embedding capacity.

Difference expansion (DE) was proposed by Tian et al. [9] in 2003, it embedded data by expanding the differences between pixel pairs, achieving a balance between embedding capacity and image quality. Subsequent work proposed by Mandal et al. [10] further improved performance through non-contiguous interpolation expansion.

Histogram shifting (HS) was proposed by Ni et al. [11] in 2006. It embeds data at peak points by shifting the pixel histogram. This method typically achieves high image quality. Jia et al. [12] increased the embedding capacity by controlling the effective shifting range. Padmaja et al. [13] addressed the problems of pixel overflow and limited embedding capacity through prediction error histogram shifting. Zhan et al. [14] proposed a reversible image fragile watermarking scheme based on a wormhole matrix, which provided dual tamper detection capability.

2.2. Copyright Protection of Neural Network

White-box watermarking requires access to internal model parameters. Uchida et al. [15] first proposed embedding watermarks into network parameters under regularization constraints. Fan et al. [16] introduced a passport layer mechanism that links model functionality to watermark verification, thereby improving robustness. But in practical deployment scenarios, internal model parameters are typically inaccessible to users, which limits the applicability of white-box watermarking methods.

In contrast, black-box watermarking verifies ownership through input-output relationships and is therefore more practical. Adi et al. [17] first proposed backdoor mechanisms into watermarking by injecting trigger samples with predefined labels. Subsequent studies further investigated trigger generation and robustness enhancement. For instance, Guo et al. [18] generated imperceptible noise-based triggers using user data, and Jia et al. [19] incorporated error-correcting codes to improve reliability. Other works explored adaptive trigger generation based on adversarial characteristics [20], enforced feature consistency between clean and trigger samples, or embedded watermarks via knowledge distillation to resist fine-tuning and pruning. While these approaches enhance trigger stealth and robustness, they ultimately depend on behavioral verification, lacking the capability for reversible content-level authentication.

Gu et al. [21] demonstrated that fixed trigger patterns can reliably activate predefined behaviors, while Chen et al. [22] further studied targeted backdoor attacks through poisoning samples. Subsequent works explored diverse watermark embedding strategies, including decision-boundary adjustment [23], auxiliary trigger classes [24], fixed-layer fine-tuning [25], and label-consistent backdoor attacks [26-27]. Other approaches improved stealth or robustness through training-set corruption without label poisoning [28], imperceptible trigger design such as Poison Ink [29], or provable watermarking schemes based on random smoothing [30].

Despite these advances, most existing methods rely primarily on behavioral verification through trigger-response consistency and lack reversible content-level authentication. Static or irreversible trigger designs further limit evidential credibility and prevent lossless restoration of carrier sample. Although RDH and trigger-based watermarking have been extensively studied, their integration remains underexplored. This gap motivates the development of a watermarking framework that jointly considers trigger effectiveness, visual imperceptibility, and reversible verification capability.

III. METHODS

This section introduces the proposed watermarking framework, which is composed of three main stages. First, a stego trigger set is constructed by selecting carrier samples from the training data and embedding copyright data using reversible data hiding. Second, the stego trigger set is incorporated into the model training process to implant a verifiable watermark behavior. Third, a dual-evidence verification procedure is performed to authenticate model ownership through both behavioral responses and reversible data extraction.

3.1. Trigger Set Generation

3.1.1. Carrier Sample Selection

To generate the stego trigger set, we first select some carrier samples from clean training samples, which serve as embedding media for copyright data. Let the training dataset be denoted as:

$$D = \{(x_i, y_i)\}_{i=1}^{N}, \tag{1}$$

where x_i and y_i represent the i-th input sample and its ground-truth label, respectively. Let $f_b(\cdot)$ denote a clean model trained on D.

Carrier samples are selected according to the following criteria:

3.1.1.1. Label constraint

Samples whose ground-truth label equals the predefined target class are discarded:

$$y_i \neq y_{target}. \tag{2}$$

This constraint is adopted to avoid ambiguity between carrier samples and the target label used in subsequent watermark embedding and verification.

3.1.1.2. Prediction consistency and confidence constraint

For each candidate sample $\boldsymbol{x_i}$, we obtain the predicted label:

$$\hat{y}_i = \arg\max_i f_b(x_i), \tag{3}$$

and the corresponding confidence score:

$$p_i = \max_i \text{softmax}\, f_b(x_i). \tag{4}$$

A sample is retained as a carrier sample only if it is correctly classified by the baseline model and its confidence is lower than a predefined threshold τ:

$$\hat{y}_i = y_i \wedge p_i < \tau. \tag{5}$$

The prediction consistency condition ensures that carrier samples are reliable under the original task, while the constraint excludes overly "easy" samples with high confidences, thereby improving robustness against reversible embedding perturbations. Fig. 1 illustrates the generation process of the stego trigger set

3.1.2. Copyright Data Embedding

After selecting carrier samples, we embed explicit copyright data into each carrier sample via RDH, thereby generating the stego trigger set. Different from triggers defined by fixed patterns, the proposed triggers are based on training data and reversible: (i) only slight pixel-level modifications are introduced, yielding strong visual imperceptibility; and (ii) the embedded data can be extracted and the carrier can be losslessly restored, which provides content-level evidence to strengthen ownership authentication.

Formally, given a carrier sample $\boldsymbol{x}$ and a binary copyright data $\boldsymbol{m}$ (e.g., owner identifier and timestamp), the RDH embedding function generates a stego trigger $x^{stego}=E(\boldsymbol{x}, \boldsymbol{m})$. During verification, the decoding function extracts the data and restores the clean carrier sample, such as, $(\hat{\boldsymbol{m}}, \hat{\boldsymbol{x}})=D(\boldsymbol{x}^{stego})$, where $\hat{\boldsymbol{m}} = \boldsymbol{m}$ and $\hat{\boldsymbol{x}} = \boldsymbol{x}$ hold when the stego sample is intact.

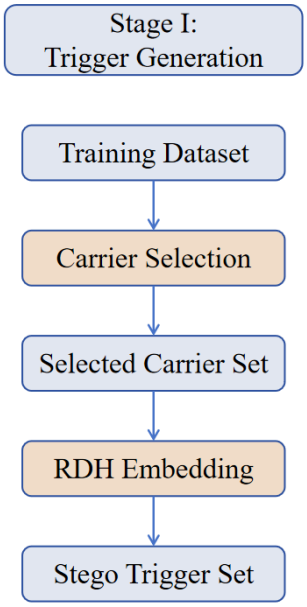

Fig. 1. Generate stego trigger set.

3.2. Watermark Embedding

The core objective of this stage is to embed the stego trigger set into the target model through training, such that the model preserves its predictive behavior on clean inputs while consistently mapping stego trigger set to a predefined target label for ownership verification.

The training dataset D_{train} is generated by merging the clean training set D_{clean} with the prepared stego trigger set $D_{trigger}$, formally defined as:

$$D_{train} = D_{trigger} \cup D_{clean}. \tag{6}$$

All trigger samples in $D_{trigger}$ are steganographic samples generated from the selected carrier samples via RDH. Each trigger sample is reassigned to a predefined target label y_{target}, which differs from its ground-truth label. This training objective enables the model to learn a stable mapping from the stego trigger set to the predefined target label y_{target}. Fig. 2 illustrates how the stego trigger set is incorporated into model training to implant watermark behavior in a target NN. To balance watermark effectiveness and clean accuracy, we apply weight decay (2×10^{-3}) as L2 regularization to reduce overfitting and preserve generalization on clean data.

After training, the watermarked model achieves two key properties: high clean accuracy comparable to the baseline model and near-perfect trigger recognition accuracy on the stego trigger set. This combination indicates that the embedded watermark is reliable for ownership verification while preserving accuracy consistency on the clean test set, thereby maintaining the model's clean functionality and providing a practical foundation for copyright protection of NN.

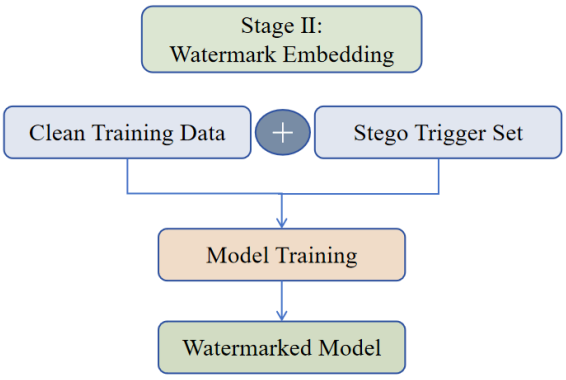

Fig. 2. Watermark embedding via stego trigger set training.

3.3. Watermark Verification

To verify model ownership, we adopt a two-stage verification procedure that establishes a reproducible evidence chain. The core idea is to utilize the stego trigger set generated via RDH, which induces a predefined watermark response in the watermarked model.

In the first stage, the verifier queries the suspicious model using the stego trigger set. Ownership is supported if the model consistently classifies these inputs as the predefined target label, resulting in a high trigger recognition accuracy.

To strengthen the reliability of this behavioral evidence, the same trigger queries are also tested on independently trained baseline models without watermark embedding. In practice, these baseline models typically fail to produce consistent target-label responses on the stego trigger set. This behavioral discrepancy between the suspicious model and baseline models serves as empirical evidence that the suspicious model has been exposed to the owner's watermarking process.

To further strengthen the ownership claim, a second-stage verification leverages the reversibility of RDH. Given a stego trigger $\boldsymbol{x}^{stego}$, the verifier applies the decoding function to (i) extract the embedded copyright data $\boldsymbol{m}$ and (ii) restore the corresponding carrier sample $\hat{\boldsymbol{x}}$ without loss.

This stage provides content-level evidence in two aspects. First, the extracted data $\boldsymbol{m}$ contains explicit ownership data that can be directly inspected. Second, the lossless restored sample of $\hat{\boldsymbol{x}}$ demonstrates that the trigger set is generated by the reversible embedding process, rather than being arbitrarily constructed inputs.

Together with the first-stage behavioral verification, the extraction and restoration results form a two-source evidence chain for ownership authentication. Fig. 3 summarizes the verification workflow, including model

querying and RDH decoding for data extraction and carrier sample restoration.

IV. RESULTS AND DISCUSSION

4.1. Experimental Setup

Two widely used benchmark datasets in the field of image classification are adopted in the experiments: CIFAR-10 and CIFAR-100. These datasets have moderate complexity, making them suitable for verification of model watermarking methods.

We selected representative DNN architectures in the deep learning field for testing, including the ResNet series (ResNet-18, 34, 101) and the VGG-16, 19. These architectures differ in depth and structural design, enabling assessment of the proposed method across diverse model configurations. We select approximately 1% of the training samples as carrier samples, which yields a trigger set size of 500 for both CIFAR-10 and CIFAR-100.

To quantitatively evaluate the performance of the watermarking method, four core metrics are adopted:

cl_acc (Clean recognition accuracy before embedding): The classification accuracy on the clean test set without watermark embedding.

tr_acc (Trigger set classification accuracy before embedding): The proportion of stego trigger set classified as the predefined target label before watermark embedding.

cl_acc* (Clean recognition accuracy after embedding): The accuracy of the model on the clean test set after watermark embedding.

tr_acc* (Trigger set classification accuracy after embedding): The accuracy of the model on the trigger set after watermark embedding.

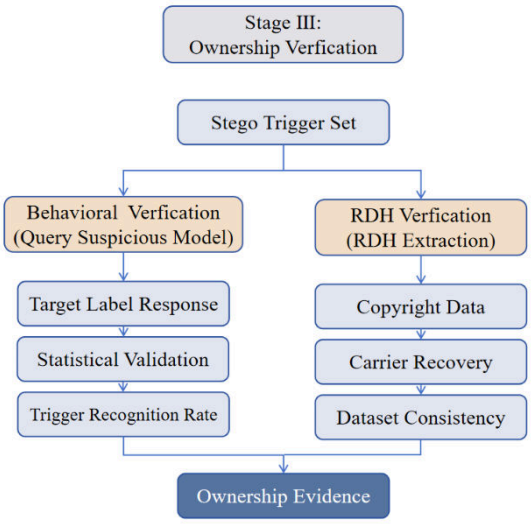

Fig. 3. Two-stage ownership verification framework.

An effective watermarking scheme should achieve a high tr_acc* while preserving cl_acc* at a level comparable to cl_acc.

The baseline model before watermark embedding is adopted for cl_acc. By comparing cl_acc with cl_acc* and comparing tr_acc with tr_acc*, we quantitatively evaluate both the preservation of clean recognition accuracy and the effectiveness of the embedded trigger mechanism.

4.2. Watermark Validity Analysis

The effectiveness of the proposed watermark is evaluated by the trigger recognition accuracy. As shown in Table 1, after watermark embedding, tr_acc* reaches 100% on ResNet-18, ResNet-34, and ResNet-101 on CIFAR-10, while VGG-16 and VGG-19 achieve 99.00%. On CIFAR-100, tr_acc* exceeds 99.8% across evaluated architectures. In contrast, tr_acc before embedding remains below 3% on both datasets, indicating that the trigger-to-target mapping is successfully established after watermark embedding.

These results demonstrate that the proposed method enables the model to consistently map stego trigger samples to the predefined target label, thereby embedding a reproducible watermark behavior into the protected models. The near-perfect trigger recognition accuracy indicates that the RDH-generated triggers are effectively incorporated during training, forming a stable trigger-to-target association for ownership verification

4.3. Visual Imperceptibility Analysis

In this work, we quantify visual quality using peak signal-to-noise ratio (PSNR) between each trigger sample and its corresponding clean sample. PSNR is widely used to measure reconstruction fidelity, where values above 30 dB are commonly regarded as indicating minor visual distortion. As reported in Table 2, our method achieves the highest PSNR (35.60 dB), suggesting that the RDH embedding introduces only slight pixel-level changes and preserves high visual quality.

Fig. 4 provides a qualitative comparison between different trigger generation methods. Compared with

Table 1. Watermark effectiveness (%).

Dataset	Model	cl_acc	tr_acc	cl_acc*	tr_acc*
CIFAR-10	ResNet-18	94.61	2.60	94.13	100
	ResNet-34	95.23	1.40	94.42	100
	ResNet-101	94.99	1.20	93.92	100
	VGG-16	93.27	2.00	92.96	99.00
	VGG-19	92.77	2.20	92.51	99.00
CIFAR-100	ResNet-18	76.65	1.00	76.02	99.80
	ResNet-34	77.47	1.00	76.13	100

Table 2. Visual quality and verification capability comparison.

Method	PSNR	Sample specific	Copyright embedding	Lossless restoration
BadNets [21]	26.32	×	×	×
Blend [22]	22.30	×	×	×
RpN [26]	21.41	○	×	×
CL [27]	25.73	×	×	×
SIG [28]	19.59	×	×	×
Ours	**35.60**	○	○	○

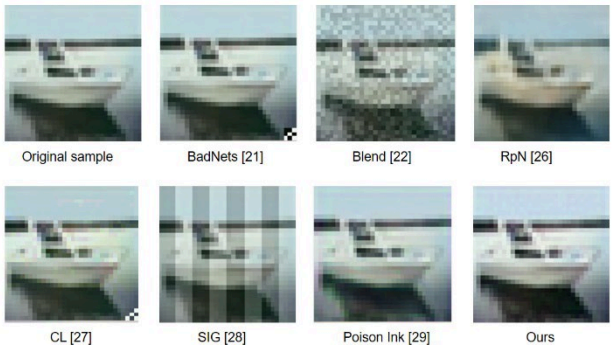

Fig. 4. Visual comparison of watermarked samples generated by different methods.

pattern- or noise-based triggers that may introduce visible textures (e.g., mosaic-like artifacts in Blend) or structured distortions, the triggers generated by our method remain visually close to the clean sample in terms of overall appearance. This observation is consistent with the quantitative PSNR results.

Beyond visual quality, Table 2 further summarizes key verification-related properties of different methods. Specifically, we report whether a method uses sample-specific triggers, whether explicit copyright data can be extracted from triggers, and whether the clean carrier samples can be losslessly restored. Most baseline backdoor methods mainly provide behavioral evidence via trigger responses, but do not support content-level authentication or reversible restoration. In contrast, our approach achieves high visual fidelity while simultaneously enabling copyright data extraction and lossless carrier sample restoration, thereby enhancing the reliability and verifiability of ownership verification.

4.4. Content-Level Verification Analysis

As described in Section 3.3., the proposed ownership verification framework consists of two stages: behavioral verification and reversible content-level verification. Since the behavioral verification results have already been presented in Table 1 through tr_acc and tr_acc*, this subsection focuses on the second-stage verification capability, including copyright data extraction (Extraction Accuracy) and carrier sample restoration (Restoration Accuracy).

Specifically, after the suspicious model passes the trigger-based behavioral verification, the verifier decodes the stego trigger samples to extract the embedded copyright data and restore the corresponding clean carrier samples. Accordingly, Table 3 reports two quantitative metrics: Extraction Accuracy measures whether the embedded copyright data can be correctly extracted, while Restoration Accuracy evaluates whether the original carrier samples can be losslessly restored through RDH reversibility.

As shown in Table 3, the proposed method achieves 100% Extraction Accuracy and 100% Restoration Accuracy in all reported settings. These results show that the embedded copyright data can be correctly extracted without error, and that the original carrier samples can be losslessly restored from the stego triggers. Therefore, the proposed method provides not only reliable behavioral evidence at the model-output level, but also explicit and reproducible content-level evidence through reversible decoding, forming a more complete evidence chain than conventional trigger-set watermarking methods.

V. CONCLUSION AND FUTURE WORK

To address the limitations of static trigger-based watermarking methods, this paper proposes a reversible stego trigger framework for copyright protection of NN. Experimental results on multiple architectures show that the proposed method achieves near-100% trigger recognition accuracy while maintaining high clean accuracy. By integrating reversible data hiding with trigger-based watermarking, the framework enables both behavioral verification and explicit copyright data extraction, enhancing the reliability of ownership authentication. Future work will explore robustness against advanced defenses and extend the method to more complex models.

In addition, although the current experiments are conducted on CIFAR-10 and CIFAR-100, the proposed framework is not inherently restricted to low-resolution

Table 3. Quantitative results of reversible content verification (%).

Dataset	Model	Extraction accuracy	Restoration accuracy
CIFAR-10	ResNet-18	100	100
	ResNet-34	100	100
	ResNet-101	100	100
	VGG-16	100	100
	VGG-19	100	100
CIFAR-100	ResNet-18	100	100
	ResNet-34	100	100

images. Since carrier sample selection and RDH-based trigger generation are both performed at the sample level, the method can be extended to higher-resolution image scenarios. Compared with low-resolution images, high-resolution images commonly provide more embedding flexibility and better visual imperceptibility under the same payload. However, such scenarios also introduce higher computational cost, larger memory consumption, and a more delicate trade-off between embedding payload and local distortion control. Therefore, extending the proposed method to higher-resolution benchmarks is an important direction for our future work.

ACKNOWLEDGEMENT

This study was funded by National Natural Science Foundation of China (62466038), Jiangxi Provincial Key Laboratory of Image Processing and Pattern Recognition (2024SSY03111), Jiangxi Provincial Natural Science Foundation (Key Program) (No. 20242BAB26015), Open Foundation of Jiangxi Provincial Key Laboratory of Image Processing and Pattern Recognition (ET202404437), and Innovation Foundation for Postgraduate Students of Nanchang Hangkong University (YC2024-S658).

REFERENCES

[1] M. Xue, Y. Zhang, J. Wang, and W. Liu, "Intellectual property protection for deep learning models: Taxonomy, methods, attacks, and evaluations," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 6, pp. 908-923, 2021.

[2] T. Wang and K. Florian, "Attacks on digital watermarks for deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2622-2626.

[3] R. Tang, H. Jin, C. Wigington, M. Du, R. Jain, and X. Hu, Was My Model Stolen? Feature Sharing for Robust and Transferable Watermarks, 2021.

[4] Q. Feng, L. Leng, C. C. Chang, J. H. Horng, and M. Wu, "Reversible data hiding in encrypted images with extended parametric binary tree labeling," *Applied Sciences*, vol. 13, no. 4, p. 2458, 2023.

[5] Z. Q. Yang, T. Tang, H. Dang, Z. Wu, and E. C. Chang, "Effectiveness of Distillation Attack and Countermeasure on Neural Network Watermarking," *IEEE Transactions on Dependable and Secure Computing*, 2025.

[6] H. Kang, L. Leng, and C. C. Chang, "Overlapped (7, 4) hamming code for large-capacity and low-loss data hiding," *Multimedia Tools and Applications*, vol. 82, no. 20, pp. 30345-30374, 2023.

[7] S. Yang, L. Leng, C. C. Chang, and C. C. Chang, "Reversible adversarial examples with minimalist evolution for recognition control in computer vision," *Applied Sciences*, vol. 15, no. 3, p. 1142, 2025.

[8] J. M. Barton, "Method and apparatus for embedding authentication information within digital data," U.S. Patent 5,646,997, Jul. 1997.

[9] J. Tian, "Reversible data embedding using a difference expansion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 8, pp. 890-896, 2003.

[10] P. C. Mandal, I. Mukherjee, and B. N. Chatterji, "High capacity reversible and secured data hiding in images using interpolation and difference expansion technique," *Multimedia Tools and Applications*, vol. 80, no. 3, pp. 3623-3644, 2021.

[11] Z. Ni, Y. Q. Shi, N. Ansari, and W. Su, "Reversible data hiding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 3, pp. 354-362, 2006.

[12] Y. Jia, Z. Yin, X. Zhang, and Y. Luo, "Reversible data hiding based on reducing invalid shifting of pixels in histogram shifting," *Signal Processing*, vol. 163, pp. 238-246, 2019.

[13] B. Padmaja and V. M. Manikandan, "A novel prediction error histogram shifting-based reversible data hiding scheme for medical image transmission," in *2021 4th International Conference on Security and Privacy (ISEA-ISAP)*, 2021, pp. 1-6.

[14] C. Zhan, L. Leng, C. C. Chang, and J. H. Horng, "Reversible image fragile watermarking with dual tampering detection," *Electronics*, vol. 13, no. 10, p. 1884, May 2024.

[15] Y. Uchida, Y. Nagai, S. Sakazawa, and S. Satoh, "Embedding watermarks into deep neural networks," in *Proceedings of the 2017 ACM International Conference on Multimedia Retrieval*, 2017, pp. 269-277.

[16] L. Fan, K. W. Ng, and C. S. Chan, "Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks," in *Advances in Neural Information Processing Systems 32*, 2019.

[17] Y. Adi, C. Baum, M. Cisse, B. Pinkas, and J. Keshet, "Turning your weakness into a strength: Watermarking deep neural networks by backdooring," in *27th USENIX Security Symposium*, Baltimore, MD, 2018, pp. 1615-1631.

[18] J. Guo and M. Potkonjak, "Evolutionary trigger set generation for DNN black-box watermarking," *arXiv Prep. arXiv:1906.04411*, 2019.

[19] J. Jia, B. Wang, and N. Z. Gong, "Robust and verifiable information embedding attacks to deep neural networks via error-correcting codes," in *Proceedings*

of the 2021 ACM Asia Conference on Computer and Communications Security, 2021, pp. 2-13.

[20] H. Jia, C. A. Choquette-Choo, V. Chandrasekaran, and N. Papernot, "Entangled watermarks as a defense against model extraction," in *30th USENIX Security Symposium*, Virtual, 2021, pp. 1937-1954.

[21] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "BadNets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47230-47244, 2019.

[22] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv Prep. arXiv:1712.05526*, 2017.

[23] E. Le Merrer, P. Perez, and G. Trédan, "Adversarial frontier stitching for remote neural network watermarking," *Neural Computing and Applications*, vol. 32, no. 13, pp. 9233-9244, 2020.

[24] Q. Zhong, L. Y. Zhang, J. Zhang, L. Gao, and Y. Xiang, "Protecting IP of deep neural networks with watermarking: A new label helps," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, Sydney, Australia, 2020, pp. 462-474.

[25] N. Chattopadhyay and A. Chattopadhyay, "Rowback: Robust watermarking for neural networks using backdoors," in *20th IEEE International Conference on Machine Learning and Applications*, 2021, pp. 1728-1735.

[26] B. Wang, F. Yu, F. Wei, Y. Li, and W. Wang, "Invisible intruders: Label-consistent backdoor attack using re-parameterized noise trigger," *IEEE Transactions on Multimedia*, vol. 26, pp. 10766-10778, 2024.

[27] A. Turner, D. Tsipras, and A. Madry, "Label-consistent backdoor attacks," *arXiv Prep. arXiv:1912.02771*, 2019.

[28] M. Barni, K. Kassem, and B. Tondi, "A new backdoor attack in CNNs by training set corruption without label poisoning," in *IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, 2019, pp.101-105.

[29] J. Zhang, C. Dongdong, Q. Huang, and J. Liao, "Poison ink: Robust and invisible backdoor attack," *IEEE Transactions on Image Processing*, vol. 31, pp. 5691-5705, 2022.

[30] A. Bansal, P. Chiang, M. J. Curry, R. Jain, C. Wigington, and V. Manjunatha, et al., "Certified neural network watermarks with randomized smoothing," in *International Conference on Machine Learning*, Baltimore, MD, 2022, pp. 1450-1465.

AUTHORS

Kaiyang Hu is currently pursuing the M.S. degree with the School of Software, Nanchang Hangkong University, Nanchang, China. His research interests include reversible data hiding and copyright protection for neural networks.

Lu Leng received the Ph.D. degree from Southwest Jiaotong University, Chengdu, China, in 2012. He conducted postdoctoral research at Yonsei University, Seoul, South Korea, and Nanjing University of Aeronautics and Astronautics, Nanjing, China. He was also a visiting scholar with West Virginia University, USA, and Yonsei University, South Korea. He is currently a Full Professor and Dean of the Institute of Computer Vision, Nanchang Hangkong University, Nanchang, China. He has published more than 150 papers in international journals and conferences and has led several funded research projects, including six projects supported by the National Natural Science Foundation of China. His research interests include computer vision, biometric template protection, biometric recognition, medical image processing, and data hiding.