TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214
DOI: 10.26599/TST.2025.9010151

Black-Box Watermark Method Based on Vision Reasoning

Yuhuan Liu†, Haowen Hu†, Weixuan Tang*, Yingfeng Zhang, Kai Ding, Bin Ma, Zhili Zhou*

Abstract: Model watermark is a technique to protect the deep learning models' copyright. However, existing watermark methods are vulnerable to watermark attack. In ambiguity attack, attacker can reversely construct the input according to the preset output, and utilize this input-output pair as forged watermark. In fine-tuning attack, attacker can remove watermark by performing fine-tuning operations on model. To overcome these limitations, this paper proposes a black-box watermark method called WaViR (Watermark based on Vision Reasoning). WaViR consists of three modules. In watermark construction, the original image is transformed into hash image by cryptographic hash function. These original and hash image form into input-output pair for watermark trigger set. In watermark embedding, the trigger set is utilized to train the image generation model. Besides, simulated fine-tuning is introduced to improve the robustness of watermark. In watermark verification, vision reasoning is applied for ownership verification. For specific image within the trigger set, if the SSIM between the model's output image and hash image exceeds the threshold, then verification is successful. Owing to the irreversibility of hash function, attacker cannot reversely construct the input that has hash relation with the preset output. Results show that WaViR can resist ambiguity attack and fine-tuning attack.

Key words: Model Watermark, Image Reasoning, Ambiguity Attack, Fine-tuning Attack

• Yuhuan Liu is with School of Computer Science and Engineering, Macau University of Science and Technology, Taipa 999078, Macau, China. E-mail: liu_yuhuan1991@163.com
• Haowen Hu and Weixuan Tang are with the Institute of Artificial Intelligence, Guangzhou University, Guangzhou 510006, China. E-mail: 592928287@qq.com; tweix@gzhu.edu.cn
• Yingfeng Zhang is with the Data Intelligence Business Unit, Unicom (Guangdong) Industry Internet Co., Ltd., Guangzhou 510320, China. E-mail: 39800549@qq.com
• Kai Ding is with the National key laboratory of science and technology on near-surface detection, Wuxi 214035, China. E-mail: winfast113@sina.com
• Bin Ma is with School of Cyber Security, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China. E-mail: sddxmb@126.com
• Zhili Zhou is with School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 510006, China. E-mail: zhou_zhili@163.com
† These authors contributed equally to this paper.
∗ To whom correspondence should be addressed.
Manuscript received: 2025-08-14; revised: 2025-09-04; accepted: 2025-09-24

1 Introduction

With the rapid development of deep learning technology, image generation models have demonstrated powerful capabilities in different fields such as computer vision, artistic creation, and content generation[1–3]. However, the wide dissemination of the image generation models has also brought serious intellectual property issues. The unauthorized users could maliciously use models trained by other users, which seriously infringes the legitimate rights and interests of model developers[4].

To safeguard the copyright of the model owners, researchers have proposed various model watermarking techniques, attempting to embed covert and verifiable information in the models for copyright authentication[5,6]. According to the different access rights of models, existing model watermark methods can be divided into white-box, black-box, and box-free watermark. White-box watermark requires to access the internal parameters of the model to

verify the watermark[7]. Black-box watermark uses specific trigger inputs to produce abnormal predetermined outputs. These input-output pairs forms into watermarks in verification[8]. Box-free watermark extracts the watermark from the content of the model output[9]. This paper focuses on black-box watermark for image generation models.

Although black-box watermark has developed rapidly in copyright protection, there still remain security risks when facing model watermark attacks[10]. Firstly, as for ambiguity attack, given the image model and the abnormal preset output, attackers can use reverse engineering technology to calculate the input. These input-output pairs can form into forged watermark, which is consistent with the original watermark. Therefore, ambiguity exists in watermark verification and the ownership of the model cannot be uniquely determined.[11]. Secondly, as for fine-tuning attack, attackers can perform carefully designed fine-tuning operations on the model to effectively remove the original watermark information without affecting the model's performance on main task, thereby bypassing the watermark verification mechanism[12]. These attack methods seriously weaken the actual protection capabilities of existing model watermarking methods[13].

To solve the above issues, this paper proposes a black-box watermark method called WaViR (Watermark based on Vision Reasoning), which can resist both ambiguity and fine-tuning attack. In WaViR, the model owner maps the original image into the hash image according to the irreversible encryption hash function such as SHA-256[14], and utilizes the original and hash images as input-output pairs to train the image generation model. In watermark verification, the specific input image is mapped into the hash image, and the image generation model takes in this specific input image and produces the output image. A vision-reasoning-based verification mechanism is constructed. The verification is successful when the hash image and output image is the same according to visual metric. Due to the hash function applied in WaViR, the attacker cannot reversely construct the input according to the preset output. Therefore, WaViR can resist ambiguity attack for black-box watermark. Besides, fine-tuning is simulated in model training, which can balance the model's performance on main task and the watermark verification performance against fine-tuning attack. The main contributions of this paper are as follows:

- A model watermark method called WaViR is proposed, wherein the watermark is verified based on vision reasoning of the hash image and model's output image. Due to the irreversibility of the hash function, WaViR can well resist ambiguity attack.
- Simulated fine-tuning mechanism is introduced into WaViR, which can improve the watermark robustness against fine-tuning attack.
- Extensive experiments have been conducted. Results show that WaViR can simultaneously maintain satisfactory performance on the main task for image generation while improving the security against ambiguity and fine-tuning attacks.

2 Related Work

2.1 Model Watermark

Neural network model watermark is a technique that embeds identifying information into the neural network models to protect their intellectual property and enable ownership verification[15,16]. Depending on application scenarios and requirements, model watermark methods can be categorized into white-box, black-box, and box-free approaches.

White-box watermark methods require full access to the internal structure and parameters of the neural network during the watermark embedding and verification process. Such methods directly encode watermark information into model's parameters[17–20]. Common strategies include making slight modifications to specific layer's parameters or embedding watermark information into the activation value distribution of intermediate layers. During the verification phase, the verifier needs to obtain the model's complete parameters and extracts watermarks by analyzing preset embedding positions or statistical features to confirm ownership. White-box watermark has the advantages of strong concealment and high robustness, but its main limitation is its strict reliance on the internal access rights of the model, which is often difficult to meet in practical applications, thus limiting its scope of usage.

Black-box watermark method does not require access to the internal parameters or model's structural information. Its core idea is to carry and verify the watermark through the model's output response with respect to specific inputs. In the embedding phase, the model owner trains the model to perform its main

task while producing abnormal preset outputs, which are typically unrelated to the main task[21–23]. In the verification phase, the owner can submit these trigger inputs to the target model and check whether the outputs match the preset watermark responses. The advantage of black-box watermark lies in the high practicality, as it relies solely on the standard input and output interface, making it especially suitable for models deployed as online services. The main challenge is how to design a robust trigger set while ensuring minimal impact on the performance of the main task[24].

Box-free watermark methods do not require access to a model's internal parameters nor the submission of specific trigger inputs during verification. Instead, they rely solely on analyzing the outputs generated by the model for regular inputs to authenticate ownership[25,26]. The core advantage of this approach is that watermark verification can be performed without any direct interaction with the target model. In practice, box-free watermark is typically implemented by training a dedicated watermark extraction network capable of extracting the embedded watermark information from the model's outputs[27,28]. The main challenge of box-free watermarking lies in designing watermark patterns that are sufficiently distinctive and robust while ensuring they do not degrade the model's performance on its main task.

2.2 Ambiguity Attack

The origin of ambiguity attack can be traced back to traditional digital watermark, such as image, audio, and video watermark[29,30]. In these scenarios, attackers do not attempt to remove the legitimate watermark from digital media. Instead, they use reverse engineering and construction techniques to embed a forged watermark into the same content to arouse confusion.

In the past few years, ambiguity attack is also designed to attack model watermark methods. Unlike removal attack that aims to delete or destroy the watermark, the primary objective of ambiguity attack is not to erase the original watermark but to embed a forged watermark into the model[31]. This leads to the original genuine watermark losing its uniqueness and thus the model owner cannot prove the ownership. In white-box watermark, attacker with access to the model's internal structure can directly modify parameters to embed forged signatures representing their ownership[32]. In black-box watermark, attacker can generate new forged trigger samples through reverse engineering, causing the model to produce predefined responses[33]. For box-free watermark, attacker can train a dedicated extraction network capable of retrieving the predefined forged watermark from the model's outputs. By embedding forged watermark, ambiguity attack creates uncertainty over copyright ownership and disables the model watermark methods. Therefore, designing robust watermark methods against ambiguity attack has urgent need.

3 Proposed Method

This paper proposes a black-box watermark method called WaViR (Watermark based on Vision Reasoning). In general, WaViR consists of three modules, including watermark construction, watermark embedding, and watermark verification. The diagram of WaViR is given in Fig. 1.

3.1 Watermark Construction

To ensure the uniqueness of watermark, we design a watermark construction method based on the SHA-256 cryptographic hash function. The core idea is to transform the original image A into a hash image $H(A)$ through block-level hashing, so as to create a secure and verifiable watermark trigger set.

Specifically, an image A with dimension $H \times W \times 3$ is randomly selected from the dataset, and divided into multiple non-overlapping pixel blocks along the row direction, where each block contains 32 pixels. If the width of the image is not a multiple of 32, zero-padding is applied to the right side to meet the block size requirement. This block division process is performed independently on each of the three color channels of Red, Green, and Blue.

For each color channel, the pixels within a block are converted into a byte sequence, which is then concatenated with a secret key K to enhance the unpredictability of the hash output. The concatenated sequence is input into the SHA-256 function to calculate its hash value as

$$H = \mathrm{SHA256}(\mathrm{bytes}(P) \parallel K), \tag{1}$$

where P denotes the current pixel block, and H is the 256-bit hash output of that block. The hash result is then mapped to 32 pixel values, and is replaced with the original block pixels in that channel. This transformation is performed independently for each color channel, and then the three processed channels are combined into the final hash image $H(A)$.

The SHA-256 function is chosen for its strong

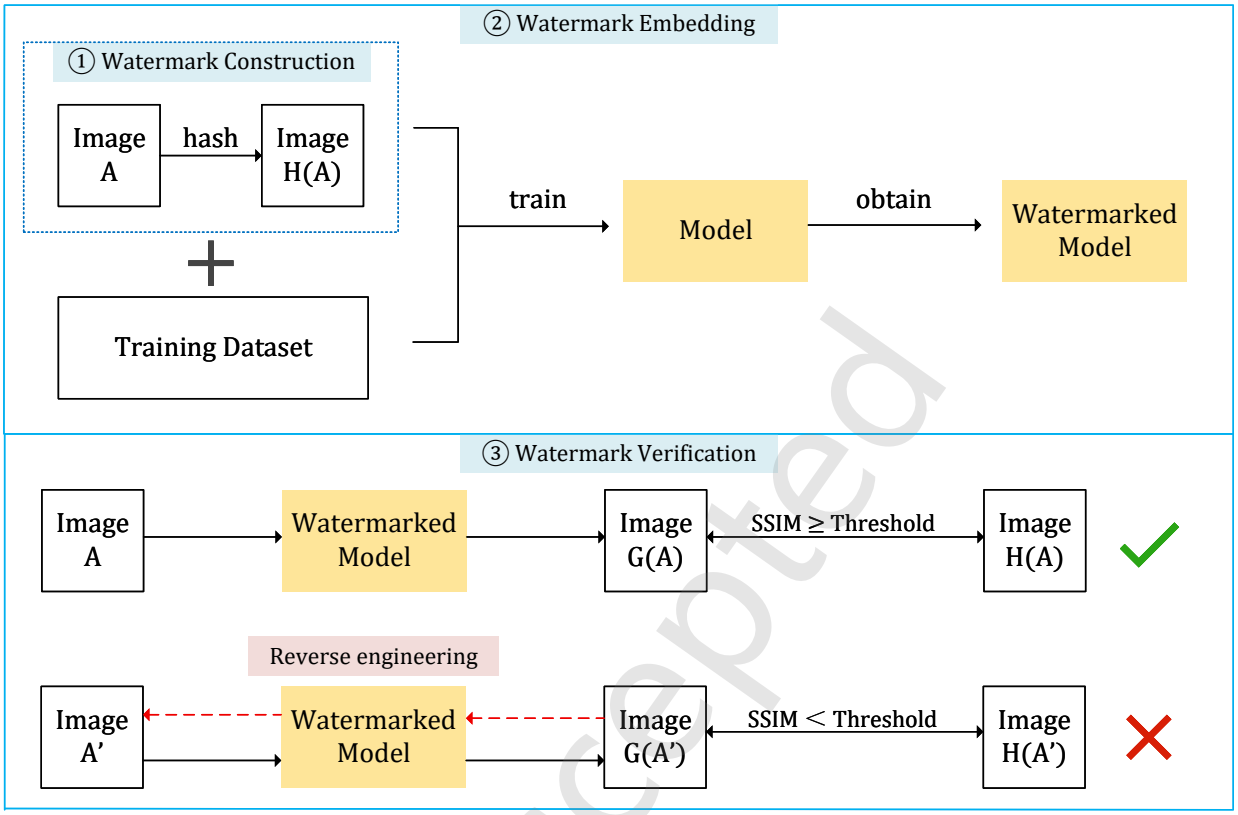

Fig. 1 Overall framework of the proposed WaViR.

collision resistance and widespread security acceptance, whereas alternatives like MD5 and SHA-1 are known to be vulnerable to practical collision attacks.

Owing to the cryptographic transformation, the relationship between image A and its corresponding hash image $H(A)$ becomes non-invertible due to the irreversible nature of the SHA-256 function. Therefore, it is computationally infeasible for an attacker to derive the original image A from the given hash image $H(A)$, which can ensure the uniqueness and security of the constructed watermark trigger set.

3.2 Watermark Embedding

To enhance the robustness of the watermark against fine-tuning attacks, we propose a watermark embedding strategy based on an alternating training framework, which consists of two iterative phases, including watermark embedding and simulated fine-tuning.

In the watermark embedding phase, the pre-trained neural network model is trained using a combination of clean training samples and the constructed watermark trigger set. The objective is to ensure that the model maintains high performance on its main task while learning to produce the preset target output when given the watermark trigger input.

In the simulated fine-tuning phase, only clean training samples are used to update the model. This phase simulates potential fine-tuning attack in real-world scenarios, where the attacker tries to erase the embedded watermark via fine-tuning the model on standard task dataset. By incorporating this simulated fine-tuning attack into the training process, the model could learn to maintain the embedded watermark when facing fine-tuning attack in the testing stage.

By means of alternatively train the model in the watermark embedding phase and the simulated fine-tuning phase, the model not only learns the watermark trigger set but also improves its robustness against fine-tuning. The pseudocode of watermark construction and watermark embedding is given in Algorithm 1.

3.3 Watermark Verification

A vision-reasoning-based mechanism is designed in watermark verification, which leverages the irreversible nature of the SHA-256 hash function to achieve reliable model ownership verification. Specifically, image A is input into the target neural network model G to generate

Algorithm 1: Watermark Construction and Alternating Embedding Training

Input: Pre-trained model M; clean dataset D_{clean}; image A; key K; epochs E; iterations $I_{\text{wm}}, I_{\text{ft}}$
Output: Watermarked model M^*

Divide image A into pixel blocks of width 32 (zero-pad if needed);
// watermark construction
for *each color channel in* A **do**
  for *each pixel block* P **do**
    Convert P to byte sequence, concatenate with K;
    Compute $H = \text{SHA256}(\text{bytes}(P) \parallel K)$;
    Map H to replace P;
  end
end
Image A is transformed into image $H(A)$; set trigger set $D_{\text{wm}} = (A, H(A))$;
// watermark embedding
for *epoch* $= 1$ **to** E **do**
  for *iteration* $= 1$ **to** I_{wm} **do**
    Sample batch from D_{clean} and D_{wm};
    Compute task loss L_{task} and watermark loss L_{wm};
    $L = L_{\text{task}} + L_{\text{wm}}$; update M using L;
  end
  for *iteration* $= 1$ **to** I_{ft} **do**
    Sample batch from D_{clean};
    Compute L_{task}; update M;
  end
end
return M^*;

an output image $G(A)$. The structural similarity index (SSIM) between $G(A)$ and the hash image $H(A)$ is calculated as

$$\mu = \text{SSIM}\big(G(A), H(A)\big), \tag{2}$$

where $G(A)$ denotes the model's output for image A, and $H(A)$ denotes the hash-transformed version of image A. If the similarity $\mu \geq 0.9$, the verification is considered to be successful. This threshold is chosen as it represents a high degree of structural similarity, widely accepted in image quality assessment to indicate that two images are perceptually nearly identical, thereby ensuring reliable verification while minimizing false positives.

The security of this vision-reasoning-based verification mechanism relies on the irreversibility of the SHA-256 hash function. Even if the attacker obtains the hash image $H(A)$, it is computationally infeasible to derive the original image A. Furthermore, generating a forged image A' through reverse engineering would not yield the model output that matches the hash image in verification. Therefore, the proposed WaViR can resist ambiguity attack.

4 Experiments

4.1 Setup

We evaluate the proposed WaViR on two representative tasks, including image deraining and style transfer.

For the image deraining task, we use MPRNet as the base image generation model. MPRNet is a multi-stage image restoration network with powerful feature representation capabilities and is widely used in image restoration tasks such as rain removal and denoising. We use multiple public image deraining datasets, including Test100, Rain100H, Rain100L, Test1200, and Test2800, which cover a variety of rainfall intensities and image scenes, enabling a comprehensive evaluation under different conditions. For the style transfer task, we use CycleGAN as the base image generation model. CycleGAN is an unpaired image-to-image translation framework widely used for style transfer tasks. We use the monet2photo dataset, which contains images from the Monet painting style domain and real-world photos.

These models were selected for their strong representativeness in their respective domains. MPRNet is a leading model for image restoration tasks, known for its multi-stage architecture that effectively balances contextual information and fine details, making it an ideal candidate for evaluating watermarking in high-precision regression tasks. CycleGAN, as a foundational framework for unpaired image-to-image translation, is widely adopted in generative tasks like style transfer, ensuring our evaluation covers the robustness of the watermark in a qualitatively different and challenging scenario.

For both tasks, we fine-tune the pre-trained model by embedding the watermark using the proposed construction and alternating training strategy.

In terms of evaluation metrics, for the deraining task, we use PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index Measure) as the performance indicators. PSNR measures the pixel-level error between the restored image and reference image, while SSIM evaluates structural consistency in terms of brightness, contrast, and texture. For the style transfer task, since there is no ground-truth paired target, we report qualitative results by visual comparison.

4.2 Fidelity Evaluation

In this part, we evaluate the fidelity of different models, where fidelity refers to the ability of image geneartion on the main task. Different model's output image is

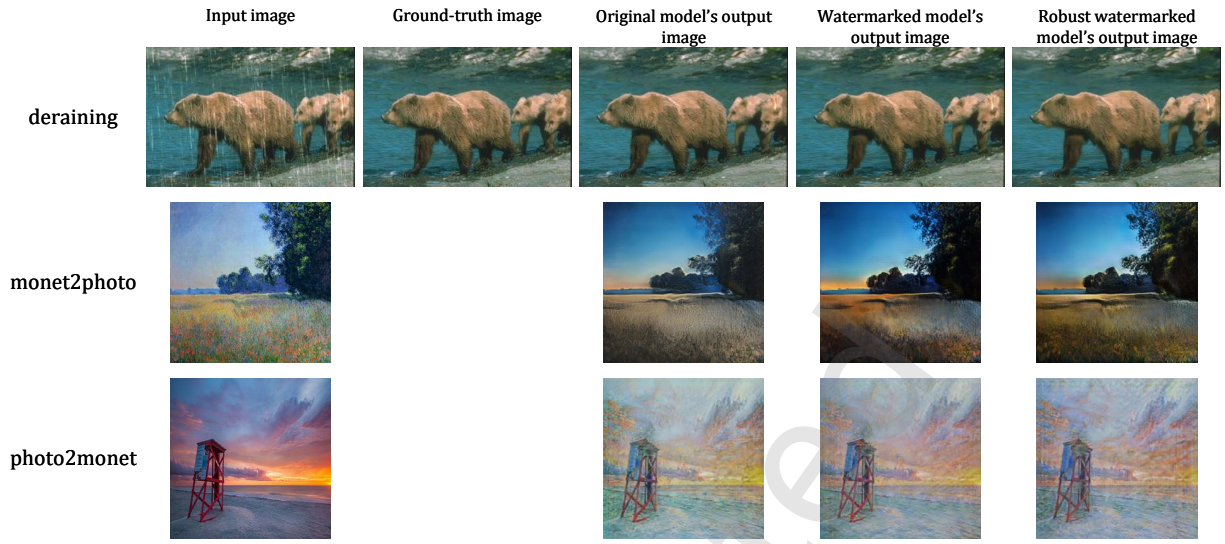

Fig. 2 Visual comparison of outputs on deraining (top) and style transfer (bottom).

compared. Specifically, the original model refers to the model without watermark, the watermarked model refers to the model with watermark embedding but without simulated fine-tuning, and the robust model refers to the model with watermark embedding and simulated fine-tuning. The subjective evaluation is given in Fig. 2. It can be observed that these three model's output images show no significant difference, indicating that the model with WaViR has satisfactory visual fidelity.

For the deraining task, we calculate the PSNR and SSIM between the generated images and clean ground-truth images. The objective evaluation is given in Table 1. It can be observed that watermark embedding leads to minor performance decrease for visual quality.

Table 1 Objective visual evaluation for different models.

Model	Method	PSNR	SSIM
MPRNet	Original	34.74	0.957
MPRNet	Watermarked	31.97	0.931
MPRNet	Robust Watermarked	31.39	0.919

4.3 Performance against Ambiguity Attack

In this part, we evaluate the proposed WaViR's performance against ambiguity attack, where the attacker attempt to construct the forged watermark trigger samples through reverse engineering to implement ownership verification.

Fig. 3 illustrates genuine and forged watermark trigger samples in the deraining task. Specifically, as for the model owner, A denotes a randomly selected watermark image, $G(A)$ denotes the generation model's output image, and $H(A)$ is obtained by applying a hash transformation to A. As for the attacker, a forged image A' is generated via reverse engineering from $G(A')$, and $H(A')$ is produced by applying the same hash transformation to A'. From Fig. 3, it can be observed that (b) and (c) is quite similar, while (e) and (f) shows great difference.

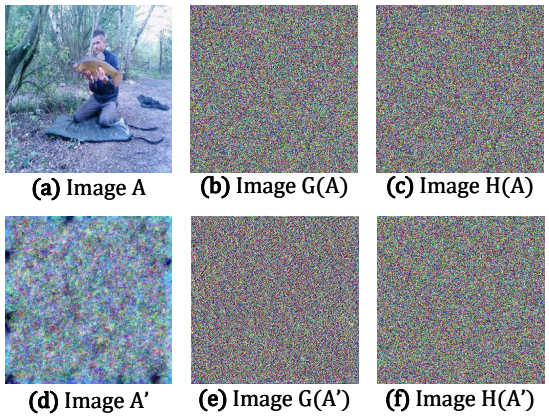

Fig. 3 Illustration of genuine and forged watermark trigger samples in the deraining task. (a) Original watermark image A, (b) Image $G(A)$ generated by the model from image A, (c) Hash-transformed target image $H(A)$, (d) Forged image A' obtained via reverse engineering, (e) Image $G(A')$ generated by the model from image A', (f) Hash-transformed image $H(A')$ from the forged image A'.

We also calculate the SSIM between $G(A)$ and $H(A)$, and the SSIM between $G(A')$ and $H(A')$. Verification is considered successful if the SSIM is greater than or equal to 0.9. Results are given in

Table 2. It can be observed that WaViR can well resist the forged watermark in ambiguity attack, verifying the effectiveness of the vision-reasoning-based mechanism.

Table 2 Verification results against ambiguity attack for MPRNet and CycleGAN.

Model	Watermark Type	SSIM
MPRNet	Genuine	0.998
MPRNet	Forged	0.004
CycleGAN	Genuine	0.959
CycleGAN	Forged	0.111

4.4 Performance against Fine-tuning Attack

In this part, we evaluate the fine-tuning attack, where the attacker would fine-tune the model on clean task data to erase the embedded watermark information.

Results are given in Table 3. Noted that in the proposed WaViR, the watermarked model refers to the model with watermark embedding but without simulated fine-tuning, and the robust model refers to the model with watermark embedding and simulated fine-tuning. Results show that after fine-tuning attacks, the watermark embedded in models without simulated fine-tuning becomes invalid and fails in verification, while the watermark embedded using the proposed defense strategy remains verifiable.

Table 3 Verification results against fine-tuning attack for MPRNet and CycleGAN.

Model	Method	SSIM Before Fine-tuning	SSIM After Fine-tuning
MPRNet	Watermarked	0.999	0.767
MPRNet	Robust Watermarked	0.998	0.914
CycleGAN	Watermarked	0.962	0.850
CycleGAN	Robust Watermarked	0.959	0.911

4.5 False Positive Analysis

To further validate the reliability of the verification mechanism, we conducted a false positive analysis. We randomly selected 100 clean, non-trigger images X from the test set and computed the SSIM between the model's output $G(X)$ and their hash-transformed versions $H(X)$. The result shows that none of these non-trigger samples (0/100) mistakenly achieved a similarity score (SSIM ≥ 0.9) required for successful verification. This negligible false positive rate (0%) demonstrates that the watermark trigger set is highly distinctive and that the verification process is unlikely to be activated by non-owner inputs, thus enhancing the credibility of ownership claims.

5 Conclusion

This paper proposes a robust model watermark method called WaViR, which combines a vision-reasoning-based verification mechanism and a simulated fine-tuning attack training strategy. Specifically, a SHA-256 block-level transformation is employed to construct a secure and unforgeable watermark trigger set, while the alternating training framework ensures that the embedded watermark remains verifiable even after unauthorized fine-tuning. Experimental results demonstrate that the proposed WaViR effectively resists ambiguity and fine-tuning attacks while preserving high model fidelity. The foundational mechanisms of WaViR—cryptographic hashing for trigger construction and simulated fine-tuning for robustness—readily extend beyond visual domains. This approach holds promising potential for application to other modalities, such as protecting the intellectual property of large language models (LLMs). Overall, the proposed method provides a practical and effective solution for intellectual property protection of neural network models.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant 62572134, 62372125, 62272255, 62302248, the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515011579, 2023A1515011428, the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant 2023B1515020041, in part by the National Key R&D Program of China under Grant SQ2025YFE0204942, Shandong Taishan Scholars Program under Grant 20240829, Key Project of Nature Foundation of Shandong Province under Grant ZR2022LZH011.

References

[1] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[2] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[3] Chitwan Saharia, William Chan, Saurabh Saxena,

Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[4] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[5] Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph Stoecklin, Heqing Huang, and Ian Molloy. Protecting intellectual property of deep neural networks with watermarking. In *Proceedings of the 2018 on Asia conference on computer and communications security*, pages 159–172, 2018.

[6] Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. Deepsigns: A generic watermarking framework for ip protection of deep learning models. *arXiv preprint arXiv:1804.00750*, 2018.

[7] Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, pages 269–277, 2017.

[8] Erwan Le Merrer, Patrick Perez, and Gilles Trédan. Adversarial frontier stitching for remote neural network watermarking. *Neural Computing and Applications*, 32(13):9233–9244, 2020.

[9] Hanzhou Wu, Gen Liu, Yuwei Yao, and Xinpeng Zhang. Watermarking neural networks with watermarked images. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(7):2591–2601, 2020.

[10] Mingfu Xue, Chengxiang Yuan, Heyi Wu, Yushu Zhang, and Weiqiang Liu. Machine learning security: Threats, countermeasures, and evaluations. *IEEE Access*, 8:74720–74742, 2020.

[11] Scott Craver, Nasir Memon, B-L Yeo, and Minerva M Yeung. Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications. *IEEE Journal on Selected areas in Communications*, 16(4):573–586, 1998.

[12] Relu Laurentiu Tataru, Safwan El Assad, and Olivier Déforges. Improved blind dct watermarking by using chaotic sequences. In *2012 International Conference for Internet Technology and Secured Transactions*, pages 46–50. IEEE, 2012.

[13] Jia Guo and Miodrag Potkonjak. Watermarking deep neural networks for embedded systems. In *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 1–8. IEEE, 2018.

[14] Fips Pub. Secure hash standard (shs). *Fips pub*, 180(4):180–4, 2012.

[15] Huili Chen, Bita Darvish Rouhani, and Farinaz Koushanfar. Blackmarks: Blackbox multibit watermarking for deep neural networks. *arXiv preprint arXiv:1904.00344*, 2019.

[16] Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *27th USENIX security symposium (USENIX Security 18)*, pages 1615–1631, 2018.

[17] Jie Zhang, Dongdong Chen, Jing Liao, Weiming Zhang, Gang Hua, and Nenghai Yu. Passport-aware normalization for deep model protection. *Advances in Neural Information Processing Systems*, 33:22619–22628, 2020.

[18] Hengrui Jia, Christopher A Choquette-Choo, Varun Chandrasekaran, and Nicolas Papernot. Entangled watermarks as a defense against model extraction. In *30th USENIX security symposium (USENIX Security 21)*, pages 1937–1954, 2021.

[19] Peizhuo Lv, Pan Li, Shengzhi Zhang, Kai Chen, Ruigang Liang, Hualong Ma, Yue Zhao, and Yingjiu Li. A robustness-assured white-box watermark in neural networks. *IEEE Transactions on Dependable and Secure Computing*, 20(6):5214–5229, 2023.

[20] Tianhao Wang and Florian Kerschbaum. Robust and undetectable white-box watermarks for deep neural networks. *arXiv preprint arXiv:1910.14268*, 1(2), 2019.

[21] Masoumeh Shafieinejad, Nils Lukas, Jiaqi Wang, Xinda Li, and Florian Kerschbaum. On the robustness of backdoor-based watermarking in deep neural networks. In *Proceedings of the 2021 ACM workshop on information hiding and multimedia security*, pages 177–188, 2021.

[22] Yingjie Lao, Weijie Zhao, Peng Yang, and Ping Li. Deepauth: A dnn authentication framework by model-unique and fragile signature embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9595–9603, 2022.

[23] Suyoung Lee, Wonho Song, Suman Jana, Meeyoung Cha, and Sooel Son. Evaluating the robustness of trigger set-based watermarks embedded in deep neural networks. *IEEE Transactions on Dependable and Secure Computing*, 20(4):3434–3448, 2022.

[24] Guang Hua, Andrew Beng Jin Teoh, Yong Xiang, and Hao Jiang. Unambiguous and high-fidelity backdoor watermarking for deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[25] Cheng Xiong, Chuan Qin, Guorui Feng, and Xinpeng Zhang. Flexible and secure watermarking for latent diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1668–1676, 2023.

[26] Ruinan Ma, Yu-an Tan, Shangbo Wu, Tian Chen, Yajie Wang, and Yuanzhang Li. Unified high-binding watermark for unconditional image generation models. *arXiv preprint arXiv:2310.09479*, 2023.

[27] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22466–22477, 2023.

[28] Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12162–12171, 2024.

[29] Qiming Li and Ee-Chien Chang. Zero-knowledge watermark detection resistant to ambiguity attacks. In *Proceedings of the 8th workshop on Multimedia and security*, pages 158–163, 2006.

[30] Khaled Loukhaoukha, Ahmed Refaey, and Khalil Zebbiche. Ambiguity attacks on robust blind image watermarking scheme based on redundant discrete wavelet transform and singular value decomposition. *Journal of Electrical Systems and Information Technology*, 4(3):359–368, 2017.

[31] Yiming Chen, Jinyu Tian, Xiangyu Chen, and Jiantao Zhou. Effective ambiguity attack against passport-based dnn intellectual property protection schemes through fully connected layer substitution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8123–8132, 2023.

[32] Lixin Fan, Kam Woh Ng, and Chee Seng Chan. Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks. *Advances in neural information processing systems*, 32, 2019.

[33] Zihan Yuan, Li Li, Zichi Wang, and Xinpeng Zhang. Ambiguity attack against text-to-image diffusion model watermarking. *Signal Processing*, 221:109509, 2024.

Yuhuan Liu received her MS from Nanjing University of Finance and Economics, Nanjing, China, in 2023. She has authored or coauthored more than 10 refereed papers. Her current research interests include Electronic Commerce, Marketing, and Artificial Intelligence.

Haowen Hu is currently a master's student at Guangzhou University, Guangzhou, China. His research interests include model watermark.

Weixuan Tang received the B.E. and Ph.D. degrees from the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China, in 2014 and 2019, respectively. He is currently an Associate Professor with the Institute of Artificial Intelligence, Guangzhou University, Guangzhou, China. His research interests include digital image steganography, model watermark and artificial security.

Yingfeng Zhang received his Bachelor's degree in Electronic Information Engineering from the People's Liberation Army University of Science and Technology, Nanjing, China, in 2007 and his Master's degree in Science from the Department of Science and Technology Communication and Policy

at the University of Science and Technology of China, Hefei, China, in 2011. He currently serves as Deputy General Manager and Chief Data Officer at Unicom (Guangdong) Industry Internet Co., Ltd., Guangzhou, China, and as Chairman of the Science and Technology Innovation Committee at China United Network Communications Group Co., Ltd. Guangdong Branch, Guangzhou, China. His research interests include federated learning and multi-party secure computation.

Kai Ding received his Ph.D. degree from the Army Engineering University, Nanjing, China, in 2013, where he is currently serving as an engineer. His research interests include pattern recognition and target localization technologies, with a long-term focus on advancing methods and applications in these areas.

Bin Ma received the M.S. and Ph.D. degrees from Shandong University, Jinan, China, in 2005 and 2008, respectively. He is currently a Professor with the School of Cyber Security, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China. Also, from 2008 to 2013, he was an Associate Professor with the School of Information Science, Shandong University of Political Science and Law, Jinan, China. From 2013 to 2015, he visited the New Jersey Institute of Technology, Newark, NJ, USA, as a Visiting Scholar. His research interests include reversible data hiding, multimedia security, and image processing. He has authored or coauthored more than 180 refereed papers. He is a member of ACM. He also serves as an Editorial Board Member of a few journals, such as the IEEE TIFS, the Journal of Visual Communication and Image Representation, and IEEE Signal Processing.

Zhili Zhou received his MS and PhD degrees in Computer Application at the School of Information Science and Engineering from Hunan University, Changsha, China, in 2010 and 2014, respectively. He is currently a professor with School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, China. Also, he was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, Canada. His current research interests include Multimedia Security, Artificial Intelligence Security and Information Hiding. He has authored or coauthored more than 150 refereed papers. He is serving as an Associate Editor of Tsinghua Science and Technology, Big Data Mining and Analytics, Cybersecurity, Journal of Real-Time Image Processing, and International Journal on Semantic Web and Information Systems. He has been selected as "World's Top 2% Scientists Career-long Impact Ranking" by Stanford University and Elsevier. He received ACM SIGWEB Rising Star Award and got Guangdong Natural Science Funds for Distinguished Young Scholar.